Zbornik 28. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2025

Zvezek B

Proceedings of the 28th International Multiconference

INFORMATION SOCIETY – IS 2025

Volume B

Kognitivna znanost
Cognitive Science

Urednika / Editors

Anka Slavka Ozimič, Borut Trpin, Toma Strle

http://is.ijs.si

9. oktober 2025 / 9 October 2025
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

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

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

Ljubljana, oktober 2025

Informacijska družba
ISSN 2630-371X

PREDGOVOR MULTIKONFERENCI INFORMACIJSKA DRUŽBA 2025

28. mednarodna multikonferenca *Informacijska družba* se odvija v času izjemne rasti umetne inteligence, njenih aplikacij in vplivov na človeštvo. Vsako leto vstopamo v novo dobo, v kateri generativna umetna inteligenca ter drugi inovativni pristopi oblikujejo poti k superinteligenci in singularnosti, ki bosta krojili prihodnost človeške civilizacije. Naša konferenca je tako hkrati tradicionalna znanstvena in akademsko odprta, pa tudi inkubator novih, pogumnih idej in pogledov.

Letošnja konferenca poleg umetne inteligence vključuje tudi razprave o perečih temah današnjega časa: ohranjanje okolja, demografski izzivi, zdravstvo in preobrazba družbenih struktur. Razvoj UI ponuja rešitve za številne sodobne izzive, kar poudarja pomen sodelovanja med raziskovalci, strokovnjaki in odločevalci pri oblikovanju trajnostnih strategij. Zavedamo se, da živimo v obdobju velikih sprememb, kjer je ključno, da z inovativnimi pristopi in poglobljenim znanjem ustvarimo informacijsko družbo, ki bo varna, vključujoča in trajnostna.

V okviru multikonference smo letos združili dvanajst vsebinsko raznolikih srečanj, ki odražajo širino in globino informacijskih ved: od umetne inteligence v zdravstvu, demografskih in družinskih analiz, digitalne preobrazbe zdravstvene nege ter digitalne vključenosti v informacijski družbi, do raziskav na področju kognitivne znanosti, zdrave dolgoživosti ter vzgoje in izobraževanja v informacijski družbi. Pridružujejo se konference o legendah računalništva in informatike, prenosu tehnologij, mitih in resnicah o varovanju okolja, odkrivanju znanja in podatkovnih skladiščih ter seveda Slovenska konferenca o umetni inteligenci.

Poleg referatov bodo okrogle mize in delavnice omogočile poglobljeno izmenjavo mnenj, ki bo pomembno prispevala k oblikovanju prihodnje informacijske družbe. »Legende računalništva in informatike« predstavljajo domači »Hall of Fame« za izjemne posameznike s tega področja. Še naprej bomo spodbujali raziskovanje in razvoj, odličnost in sodelovanje; razširjeni referati bodo objavljeni v reviji *Informatica*, s podporo dolgoletne tradicije in v sodelovanju z akademskimi institucijami ter strokovnimi združenji, kot so ACM Slovenija, SLAIS, Slovensko društvo Informatika in Inženirska akademija Slovenije.

Vsako leto izberemo najbolj izstopajoče dosežke. Letos je nagrado *Michie-Turing* za izjemen življenjski prispevek k razvoju in promociji informacijske družbe prejel **Niko Schlamberger**, priznanje za raziskovalni dosežek leta pa **Tome Eftimov**. »Informacijsko limono« za najmanj primerno informacijsko tematiko je prejela odsotnost obveznega pouka računalništva v osnovnih šolah. »Informacijsko jagodo« za najboljši sistem ali storitev v letih 2024/2025 pa so prejeli Marko Robnik Šikonja, Damir Vreš in Simon Krek s skupino za slovenski veliki jezikovni model GAMS. Iskrene čestitke vsem nagrajencem!

Naša vizija ostaja jasna: prepoznati, izkoristiti in oblikovati priložnosti, ki jih prinaša digitalna preobrazba, ter ustvariti informacijsko družbo, ki koristi vsem njenim članom. Vsem sodelujočim se zahvaljujemo za njihov prispevek — veseli nas, da bomo skupaj oblikovali prihodnje dosežke, ki jih bo soustvarjala ta konferenca.

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

FOREWORD TO THE MULTICONFERENCE INFORMATION SOCIETY 2025

The 28th International Multiconference on the Information Society takes place at a time of remarkable growth in artificial intelligence, its applications, and its impact on humanity. Each year we enter a new era in which generative AI and other innovative approaches shape the path toward superintelligence and singularity — phenomena that will shape the future of human civilization. The conference is both a traditional scientific forum and an academically open incubator for new, bold ideas and perspectives.

In addition to artificial intelligence, this year's conference addresses other pressing issues of our time: environmental preservation, demographic challenges, healthcare, and the transformation of social structures. The rapid development of AI offers potential solutions to many of today's challenges and highlights the importance of collaboration among researchers, experts, and policymakers in designing sustainable strategies. We are acutely aware that we live in an era of profound change, where innovative approaches and deep knowledge are essential to creating an information society that is safe, inclusive, and sustainable.

This year's multiconference brings together twelve thematically diverse meetings reflecting the breadth and depth of the information sciences: from artificial intelligence in healthcare, demographic and family studies, and the digital transformation of nursing and digital inclusion, to research in cognitive science, healthy longevity, and education in the information society. Additional conferences include Legends of Computing and Informatics, Technology Transfer, Myths and Truths of Environmental Protection, Knowledge Discovery and Data Warehouses, and, of course, the Slovenian Conference on Artificial Intelligence.

Alongside scientific papers, round tables and workshops will provide opportunities for in-depth exchanges of views, making an important contribution to shaping the future information society. *Legends of Computing and Informatics* serves as a national »Hall of Fame« honoring outstanding individuals in the field. We will continue to promote research and development, excellence, and collaboration. Extended papers will be published in the journal *Informatica*, supported by a long-standing tradition and in cooperation with academic institutions and professional associations such as ACM Slovenia, SLAIS, the Slovenian Society Informatika, and the Slovenian Academy of Engineering.

Each year we recognize the most distinguished achievements. In 2025, the Michie-Turing Award for lifetime contribution to the development and promotion of the information society was awarded to **Niko Schlamberger**, while the Award for Research Achievement of the Year went to **Tome Eftimov**. The »Information Lemon« for the least appropriate information-related topic was awarded to the absence of compulsory computer science education in primary schools. The »Information Strawberry« for the best system or service in 2024/2025 was awarded to Marko Robnik Šikonja, Damir Vreš and Simon Krek together with their team, for developing the Slovenian large language model GAMS. We extend our warmest congratulations to all awardees.

Our vision remains clear: to identify, seize, and shape the opportunities offered by digital transformation, and to create an information society that benefits all its members. We sincerely thank all participants for their contributions and look forward to jointly shaping the future achievements that this conference will help bring about.

Mojca Ciglarič, Chair of the Program Committee
Matjaž Gams, Chair of the Organizing Committee

KONFERENČNI ODBORI
CONFERENCE COMMITTEES

International Programme Committee

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

Organizing Committee

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

Programme Committee

Mojca Ciglarič, chair
Bojan Orel
Franc Solina
Viljan Mahnič
Cene Bavec
Tomaž Kalin
Jozsef Györkös
Tadej Bajd
Jaroslav Berce
Mojca Bernik
Marko Bohanec
Ivan Bratko
Andrej Brodnik
Dušan Caf
Saša Divjak
Tomaž Erjavec
Bogdan Filipič
Andrej Gams
Matjaž Gams
Mitja Luštrek
Marko Grobelnik
Nikola Guid
Marjan Heričko
Borka Jerman Blažič Džonova
Gorazd Kandus
Urban Kordeš
Marjan Krisper
Andrej Kuščer
Jadran Lenarčič
Borut Likar
Janez Malačič
Olga Markič
Dunja Mladenič
Franc Novak
Vladislav Rajkovič
Grega Repovš
Ivan Rozman
Niko Schlamberger
Gašper Slapničar
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 28. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2025

Zvezek B

Proceedings of the 28th International Multiconference

INFORMATION SOCIETY – IS 2025

Volume B

Kognitivna znanost
Cognitive Science

Urednika / Editors

Anka Slavka Ozimič, Borut Trpin, Toma Strle

http://is.ijs.si

9. oktober 2025 / 9 October 2025
Ljubljana, Slovenia

PREDGOVOR

Dobrodošli na letošnji konferenci Kognitivna znanost v okviru multikonference Informacijska družba. Konferenca tudi letos združuje raziskovalce in raziskovalke, ki jih povezuje zanimanje za kognitivne procese in njihovo umeščenost v širši naravni in družbeni kontekst. Kognitivna znanost je interdisciplinarno raziskovalno polje, ki povezuje filozofijo, psihologijo, nevroznanost, lingvistiko, računalništvo, umetno inteligenco in sorodne discipline. Prav na presečišču različnih pristopov nastajajo nova vprašanja, metode in rešitve, ki bogatijo razumevanje kognicije in odpirajo pot k inovativnim aplikacijam.

Tudi letošnji program odraža to raznolikost. Filozofski prispevki se lotevajo temeljnih vprašanj zavesti, življenja in t. i. težkega problema; drugi se posvečajo socialni epistemologiji. Več raziskav je namenjenih velikim jezikovnim modelom: njihovi kognitivni zmogljivosti, vlogi v analizi psihiatričnih intervjujev ter razmerju med uspešnim jezikovnim vedenjem in odsotnostjo zavesti. Empirični in aplikativni prispevki obravnavajo rehabilitacijo z robotsko podporo, spremljanje faznih prehodov v psihoterapiji, sinhronijo med terapevtom in klientom, gradnjo ontologij sebstva ter raziskovanje strategij v delovnem spominu. Tak nabor tem potrjuje, da kognitivna znanost v Sloveniji in širše ostaja živahno raziskovalno polje, ki se nenehno odpira novim izzivom.

Posebno mesto ima plenarno predavanje red. prof. dr. Olge Markič, ene osrednjih osebnosti pri razvoju kognitivne znanosti v Sloveniji. Njeno delo je pomembno prispevalo k uveljavitvi interdisciplinarnega pristopa in k oblikovanju raziskovalne skupnosti, ki jo danes soustvarjamo.

Del programa je tudi okrogla miza o zaupanju. Gre za temo, ki presega meje posameznih disciplin in se dotika tako epistemologije in etike kot psihologije, sociologije ter raziskav umetne inteligence. Zaupanje je ključen pogoj za znanstveno sodelovanje, za delovanje družbenih institucij in za odgovorno uporabo novih tehnologij.

Konferenca Kognitivna znanost 2025 ostaja prostor srečevanja in dialoga med raziskovalkami in raziskovalci različnih disciplin in generacij. Upamo, da bo tudi tokrat spodbudila plodno izmenjavo idej, oblikovanje novih sodelovanj ter skupno refleksijo o prihodnjih poteh raziskovanja kognicije.

Dobrodošli!

Anka Slana Ozimič

Borut Trpin

Toma Strle

FOREWORD

Welcome to this year's Cognitive Science conference, held within the multiconference Information Society. Once again, the conference brings together researchers who share an interest in cognitive processes and their place within the broader natural and social context. Cognitive science is an interdisciplinary research field that integrates philosophy, psychology, neuroscience, linguistics, computer science, artificial intelligence, and related disciplines. It is precisely at the intersection of these diverse approaches that new questions, methods, and solutions emerge, enriching our understanding of cognition and opening the way to innovative applications.

This year's program reflects this diversity. Philosophical contributions address fundamental questions concerning consciousness, life, and the so-called "hard problem"; others focus on issues in social epistemology. Several papers investigate large language models: their cognitive capacities, their role in the analysis of psychiatric interviews, and the relation between successful linguistic performance and the absence of consciousness. Empirical and applied contributions deal with robot-assisted rehabilitation, the monitoring of phase transitions in psychotherapy, therapist–client synchrony, the construction of ontologies of the self, and the study of strategies in working memory. Taken together, these contributions demonstrate that cognitive science in Slovenia and beyond remains a dynamic field of research, continuously opening itself to new challenges.

A special place is reserved for the keynote lecture by Prof. Olga Markič, one of the central figures in the development of cognitive science in Slovenia. Her work has significantly contributed to the establishment of the interdisciplinary approach and to the formation of the research community of which we are part today.

The program also includes a round table on trust. This theme transcends disciplinary boundaries and touches upon epistemology and ethics as well as psychology, sociology, and research on artificial intelligence. Trust is a crucial condition for scientific collaboration, for the functioning of social institutions, and for the responsible use of new technologies.

The Cognitive Science 2025 conference continues to serve as a venue for encounters and dialogue among researchers from different disciplines and generations. We hope that this year's meeting will once again stimulate fruitful exchanges of ideas, foster new collaborations, and inspire collective reflection on the future directions of cognitive science.

Welcome!

Anka Slana Ozimič

Borut Trpin

Toma Strle

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;

Urška Martinc, Center za kognitivno znanost, Pedagoška fakulteta, Univerza v Ljubljani

Rehabilitacija roke z robotsko podprto obravnavo pri otrocih in mladostnikih več let po akutno nastali možganski okvari

Tina Bregant†
CIRIUS Kamnik
Slovenija
tina.bregant@cirius-kamnik.si

Renata Pavlinič
CIRIUS Kamnik
Slovenija
renata.pavlinic@cirius-kamnik.si

Patricija Šinkovec
CIRIUS Kamnik
Slovenija
patricija.sinkovec@cirius-kamnik.si

Povzetek

Izhodišča: Pri povrnitvi funkcije rok sta pomembni reorganizacija in plastičnost možganske skorje ter kortikospinalne proge. Za spodbujanje in modulacije nevronske plastičnosti uporabljamo rehabilitacijske strategije: zgodnje intervencije s ponavljajočo se v cilj usmerjeno intenzivno terapijo (motorični trening, trening z omejevanjem, robotski trening), kar pripomore k boljšemu okrevanju in povrnitvi funkcije roke.

Cilji: Ugotoviti vpliv delovne terapije z robotsko podprto obravnavo v primerjavi z vplivi klasičnih pristopov delovne terapije na funkcijo roke.

Metode: V 4-tedensko raziskavo je bilo vključenih 32 otrok in mladostnikov (od tega 15 žensk) z okvarjeno funkcijo zgornjega uda zaradi akutno nastale možganske okvare pred nekaj leti. V eksperimentalni skupini z robotskim treningom je bilo 9 žensk in 7 moških, s povprečno starostjo 17,9 let; v kontrolni skupini (standardna delovna terapija) je bilo 6 žensk in 10 moških, s povprečno starostjo 16,85 let. Okvare so nastale večinoma perinatalno – ob rojstvu oziroma v prvih tednih po rojstvu. Za ocenjevanje funkcije roke smo uporabili standardne instrumente ocenjevanja funkcije roke (ARAT, Box&Blocks, mišična moč). V analizi smo zaradi majhnega vzorca in boljše povednosti glede izboljšanja funkcije rok uporabili deskriptivno statistiko.

Rezultati: Po zaključku terapij so rezultati ocenjevanja testa ARAT in merjenja mišične moči v zgornjih udih pokazali večji napredek pri eksperimentalni skupini. Pri testu Box&Blocks pa je boljše rezultate dosegla kontrolna skupina. Po zaključenih terapijah so otroci in mladostniki podali subjektivno mnenje o zadovoljstvu glede terapij.

Zaključki: Tako klasični pristopi delovne terapije kot robotsko podprte obravnave pomembno vplivajo na izboljšanje funkcije zgornjih udov tudi nekaj let po nastali okvari. V skupini z robotsko napravo smo dosegli večji napredek na področju fine motorike. V skupini s klasičnimi pristopi je bil večji napredek na področju grobe motorike. Pri obeh terapijah je prevladovalo zadovoljstvo z njimi. Ugotavljamo, da je smiselna uporaba kombinacije obeh pristopov, saj s tem pridobimo izboljšanje tako grobe motorike kot fine motorike.

Ključne besede

Roka, zgornji ud, motorična skorja, kortikospinalna proga, plastičnost, delovna terapija, robotska rokavica

Information Society 2025, 6–10 October 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.cogni.6

1 Uvod

Funkcija roke je ključnega pomena za ohranjanje samostojnosti in skrbi zase pri dnevnih aktivnostih. Zato je obnovljena oz. povrnjena funkcija roke pogosto eden najpomembnejših ciljev za bolnike z možgansko okvaro [1]. Funkcija rok se običajno po možganski poškodbi izboljšuje počasi, najbolj pogosto šele za izboljšanjem funkcije trupa in spodnjih udov; najkasneje se povrnejo finomotorične spretnosti, kjer sodelujejo drobne mišice rok. Do 80 % preživelih odraslih po možganski kapi ima okvare v področju zgornjih udov, le redki dosežejo popolno funkcionalno okrevanje po 6 mesecih po možganski kapi [2]. Zato je izguba funkcije zgornjih udov (rok) eden od dejavnikov, ki prispevajo k zmanjšanju splošne kakovosti življenja, kar pomembno vpliva na dnevne aktivnosti, družabne aktivnosti ter pri odraslih vrnitev k poklicu. Pri otrocih in mladostnikih pa so za uspešnost v šoli pomembne finomotorične spretnosti, koordinacija oko-oko in oko-roka ter grafomotorične spretnosti, ki so po okvarah možganov lahko pomembno okrnjene.

Obseg motorične okvare med akutno ishemično možgansko kapjo je odvisen predvsem od obsega in integritete kortikospinalnega trakta, ki je bil poškodovan. Edino za območja, kjer je kortikospinalni trakt zelo zgoščen (komprimiran), kot je npr. v področju ponsa, je korelacija med motorično okvaro in velikostjo ishemične lezije majhna. Pri bolnikih z bolj ohranjeno integriteto kortikospinalnega trakta je izboljšanje po akutni ishemiji boljše, rehabilitacija pa uspešnejša. Kortikospinalna proga (imenovana tudi piramidna proga) je snop vlaken, ki povezuje možgansko skorjo s hrbtenjačo in omogoča hoteno gibanje udov. Večina vlaken (75–90 %) prestopi na nasprotno stran v podaljšani hrbtenjači (t. i. križanje piramidne proge) ter se končuje v mišičnih skupinah udov [3]. Pri otrocih živčevje in živčne povezave še zorijo; spretnosti s področja grobe in fine motorike zorijo in šele nekaj let po rojstvu se razvijejo odrasli vzorci gibanja. Zato je posledice perinatalne poškodbe nemogoče v celoti presojati takoj po dogodku; s terapevtskimi postopki nadaljujemo tudi nekaj let po nastali okvari.

2 Ocena funkcije roke

Pri funkciji zgornjih udov sta pomembna mišična moč in gibljivost v sklepih. Moč prijema in stiska roke (pesti) lahko merimo z dinamometrom; obsege gibljivosti pa z goniometrom (kotomerom). Obstaja več lestvic za oceno delovanja roke. Nobeden od testov ni tako univerzalen oz. ne pokriva vseh področij, da bi ga lahko enoznačno uporabljali, ne glede na patologijo roke. Roka je namreč tako kompleksna, da v svoji polni funkciji zahteva anatomsko integriteto, gibljivost, mišično moč, občutljivost, natančnost, spretnost in koordinacijo, specifične grobe in finomotorične veščine (prijeme), soročnost, a hkrati tudi dober kognitivni nadzor. Pri oceni funkcije roke smo z meritvami nekoliko omejeni in zato še dodatno ocenjujemo uspešnost in hitrost izvajanja praktičnih nalog [4].

Mišično moč v zgornjih udih merimo z dinamometrom. Praviloma z meritvijo poskušamo objektivizirati napredek oz. spremembe v mišični moči v zgornjih udih. Z vsako roko praviloma opravimo tri zaporedne meritve, kot končni rezultat pa upoštevamo povprečje. Rezultati se beležijo v merski enoti kilogram (kg). Testi, ki so standardizirani in jih lahko uporabljamo za oceno funkcije roke, so npr. test devetih zatičev (angl. nine hole peg test, NHPT, ki je bil razvit za vrednotenje spretnosti prstov – fino ročno spretnost), test škatle in kock (angl. box and blocks test, BBT), ki določa grobo motoriko in lateralizacijo, Jebsenov test (angl. Jebsen-Taylor hand function test, JTHFT) za oceno lateralizacije, grobe in fine motorike, test ARAT (angl. action research arm test) in sistema razvrščanja, kot sta: lestvica BFMF (angl. bimanual fine motor function) ali lestvica MACS (angl. manual ability classification system), s katerima lahko razvrstimo otroke s cerebralno paralizo glede na njihove sposobnosti rokovanja s predmeti pri dnevnih aktivnostih. Za oceno celotnega stanja je uporabna lestvica funkcijske neodvisnosti (angl. functional independence measure, FIM) [4].

3 Povrnitev funkcije roke

Spastičnost (v 20–40 %) in šibkost (spastična pareza) sta najbolj pogosti težavi po možganski poškodbi [5, 6]. Okrevanje v smislu povrnitve moči in normaliziranja tonusa ter s tem tudi motorične funkcije, pripisujemo zlasti hitri reorganizaciji (plastičnosti) korteksa in kortikospinalne proge, medtem ko neugodna plastičnost in pretirana vzdražnost retikulospinalne proge najverjetneje povzročata največ težav. Spodbude in modulacija nevronske plastičnosti z rehabilitacijskimi strategijami, kot so zgodnje intervencije s ponavljajočo se ciljno usmerjeno intenzivno terapijo (npr. motorični trening), ustrezna neinvazivna možganska stimulacija (npr. nevromodulacija s transkranialno stimulacijo) in farmakološka sredstva (vključno z apliciranjem toksina botulinum lokalno), so ključ do funkcionalnega motoričnega okrevanja [7]. Sinaptične povezave v osrednjem živčevju so plastične, kar pomeni, da jih je mogoče spremeniti na podlagi učenja [8].

4 (Re)habilitacija

V rehabilitaciji se pogosto osredotočamo na boljšo sklepno gibljivost, večjo moč in boljšo funkcijo. Z robotsko pomočjo lahko podpremo plastično reorganizacijo nitja kortikospinalne proge [9]. S pomočjo tehnologije okrepimo mehanizme biomehanske povratne zanke (t. i. biofeedback) [10]. S tem povečamo dotok informacij glede gibanja, kar presega informacije, ki so sicer na voljo, in so lahko v nasprotju s senzoričnimi (ali notranjimi) povratnimi informacijami [11], saj z robotskim gibanjem pravilnejše gibe “vsiljujemo”. Povečanje povratnih informacij o gibanju ima večje klinične učinke kot senzorične povratne informacije. Spodbudi tudi nevronsko plastičnost po poškodbi možganov [12]. Robotska naprava omogoča usposabljanje bolnikov na intenziven, k nalogam usmerjen način terapije od zgoraj navzdol, kar povečuje skladnost in motivacijo bolnikov. Kognitivna stimulacija od zgoraj navzdol se omogoča z uvedbo vizualnih povratnih informacij, izvedenih z igranjem posebnih iger, česar se tudi poslužujemo [13]. Z robotom lahko dodatno preko strojnega učenja tudi optimiziramo zahtevani vzorec gibanja. Zato je kompleksnost motorične naloge mogoče z robotiko natančneje nadzorovati kot s konvencionalnimi pristopi zdravljenja.

Raziskave kažejo, da z novimi rehabilitacijskimi protokoli lahko vseeno dosežemo motorično izboljšanje tudi kasneje, celo še leto dni po dogodku, ki pa ni tako izrazito kot na začetku [14, 15]. Med takšne programe se uvrščata: terapija z omejevanjem oz. z omejevanjem spodbujajoča terapija (CIMT) [16] in robotski trening [15, 16].

Z omejevanjem spodbujajoča terapija – CIMT, pri kateri omejimo funkcijo neprizadete roke, se izkazuje v intenzivni rehabilitacijski obravnavi kot zelo koristna, čeprav je lahko tudi frustrirajoča. Ta terapevtski postopek spodbuja funkcijo okvarjenega zgornjega uda med izvajanjem različnih aktivnosti. S tem spodbuja procese plastičnosti in reorganizacije možganov ter tako prispeva k izboljšanju funkcije okvarjenega zgornjega uda [4].

5 Raziskava

Metode

V raziskavo je bilo vključenih 32 otrok in mladostnikov s prisotno okvarjeno funkcijo zgornjega uda zaradi možganske okvare, ki je nastala perinatalno (v prvih tednih po rojstvu oz. ob rojstvu). V eksperimentalno skupino je bilo vključenih 9 žensk ter 7 moških, povprečna starost je bila 17,9 let. V kontrolno skupino je bilo vključenih 6 žensk in 10 moških, povprečna starost je bila 16,85. Raziskava je potekala strnjeno 4 tedne. Zanimal nas je vpliv intenzivnih terapij podprtih z robotsko napravo v primerjavi s terapijami, ki vključujejo klasične delovno-terapevtske pristope na funkcijo zgornjih udov ter primerjava rezultatov.

Pred pričetkom intenzivnih terapij so bila izvedena ocenjevanja (ARAT, Box&Blocks ter merjenje mišične moči z dinamometrom). Vse oblike terapij so potekale 3 krat tedensko. Terapije z robotsko napravo Syrebo so se izvajale 20 minut, klasične delovno-terapevtske obravnave pa 30 minut. Terapije z robotsko napravo so vključevale pasivno razgibavanje, vaje proti uporu, aktivne vaje ter funkcionalne vaje. Na okvarjen zgornji ud

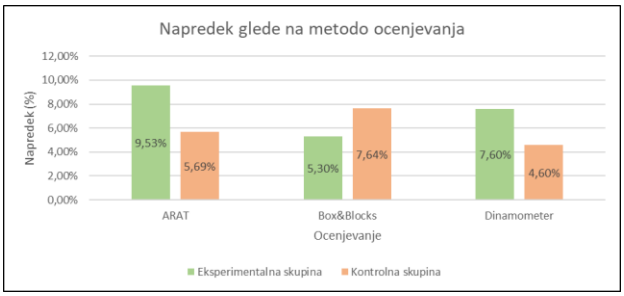
smo namestili robotsko rokavico. Naprava nam je omogočala izbiro med različnimi programi in funkcijami. Na osnovi različnih programov se je zahtevnost terapij tedensko stopnjevala.

Klasične delovno-terapevtke obravnave so bile razdeljene v tri dele: 10 minut pasivnega sproščanja ramenskega obroča, predel nadlahti, podlahti in zapestja, 10 minut terapevtsko kolo in 10 minut primerne aktivnosti usmerjene na funkcijo rok (HomeClinico, aktivnosti na Movi mizi, aktivnosti za izboljšanje mišične moči in obseg gibanja…). Vrste aktivnosti ter intenzivnost se je individualno prilagajalo vsakemu posamezniku. Pri obeh skupinah so bila po končanih intenzivnih terapijah ponovno opravljena testiranja ARAT, Box&Blocks ter merjenje mišične moči z dinamometrom. Udeleženci so izpolnili tudi nestandardiziran vprašalnik o zadovoljstvu.

V analizi smo zaradi majhnega vzorca in boljše povednosti glede izboljšanja funkcije rok uporabili deskriptivno statistiko.

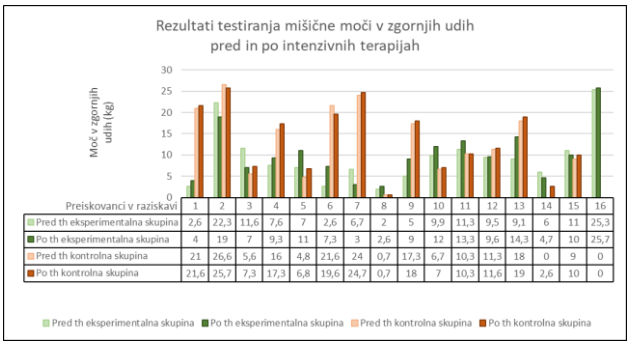
Rezultati

Rezultati so pokazali izboljšanje funkcije rok tako pri eksperimentalni skupini kot pri kontrolni skupini. Kot je prikazano na Sliki 1 je bil v eksperimentalni skupini dosežen večji napredek pri ocenjevanju ARAT in merjenju mišične moči z dinamometrom. Pri ocenjevanju Box&Blocks pa je večji napredek dosega kontrolna skupina. Pri ocenjevanju ARAT je bil zabeleženo izboljšanje v eksperimentalni skupini za 9,53 % in v kontrolni skupini za 5,69 %, večji napredek je bil dosežen v eksperimentalni skupini. Rezultati ocenjevanja Box&Blocks so pokazali večji napredek pri kontrolni skupini in sicer za 2,34 %.



Slika 1: Napredek pri ponovno izvedenih ocenjevanjih po končanih intenzivnih terapijah in primerjava med eksperimentalno in kontrolno skupino glede na metodo ocenjevanja.

Mišična moč v zgornjih udih se je po končanih terapijah v eksperimentalni skupini povečala za 7,60 %, v kontrolni za 4,60 %. V primerjavi med eksperimentalno in kontrolno skupino je bil pri končnih ocenjevanjih mišične moči večji napredek za 3,00 % v eksperimentalni skupini. Rezultati merjenja mišične moči so prikazani na Sliki 2. Ugotavljamo, da so bili pri posameznih O/M končni rezultati ocenjevanj slabši kot pri prvem ocenjevanju, zaradi odstopanj na področju procesnih spretnosti (pozornost, koncentracija, dnevno razpoloženje), kar je povezano z njihovo poškodbo možganov, ki vpliva na utrujanje ter procesne sposobnosti.



Preiskovanci v raziskavi	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Pred th eksperimentalna skupina	2,6	22,3	11,6	7,6	7	2,6	6,7	2	5	9,9	11,3	9,5	9,1	6	11	25,3
Po th eksperimentalna skupina	4	19	7	9,3	11	7,3	3	2,6	9	12	13,3	9,6	14,3	4,7	10	25,7
Pred th kontrolna skupina	21	26,6	5,6	16	4,8	21,6	24	0,7	17,3	6,7	10,3	11,3	18	0	9	0
Po th kontrolna skupina	21,6	25,7	7,3	17,3	6,8	19,6	24,7	0,7	18	7	10,3	11,6	19	2,6	10	0

Slika 2: Rezultati testiranja mišične moči zgornjih udov z dinamometrom v eksperimentalni in kontrolni skupini pred in po izvedbi intenzivnih obravnav.

Po zaključenih terapijah smo izvedli nestandardiziran vprašalnik o zadovoljstvu. Večina pozitivnih odzivov nakazuje, da ima eksperimentalni pristop z uporabo robotske rokavice pozitiven vpliv na izboljšanje grobe motorike in senzoričnega zaznavanja, zmanjšanje mišične napetosti v zgornjih okončinah ter povečano motivacijo za izvajanje terapij. Pri enem udeležencu pa so se pojavili negativni stranski učinki, in sicer povečanje mišičnega tonusa celotnega telesa. Vsa mnenja so bila podana subjektivno.

6 Zaključek

V prispevku smo osvetlili, kaj se dogaja s funkcijo roke po okvari možganov in kako lahko na funkcijo roke vplivamo z usmerjenimi terapevtskimi metodami. Ob spontanem okrevanju pri rehabilitaciji se zanašamo na plastičnost možganov, ki jo spodbujamo na pravi način, in sicer z motoričnim treningom in delovno-terapevtskimi obravnavami ali pa s sodobno tehnologijo (uporaba nevromodulacijskih tehnik kortikalnega draženja, robotika, navidezna resničnost, principi igric). Z razvojem tehnologije in razumevanjem mehanizmov delovanja živčevja upamo, da bomo tudi po možganskih okvarah dosegli čim boljšo funkcijo, zlasti funkcijo zgornjega uda, dokler ne bo medicina toliko napredovala, da bomo znali nadomestiti tudi izgubljene nevrone in njihove povezave oz. učinkovito preprečiti neugodne dogodke v našem živčevju.

Pomembno dejstvo pri izvajanju obravnav s pomočjo robotskih naprav je vidik bolnikove varnosti. Pri robotski napravi z možnostjo nastavitve moči in intenzivnosti izvajanja pasivnih vaj preprečimo možnost nastanka poškodb v primeru povišanega mišičnega tonusa ali mišičnega krča. Za varno izvajanje obravnav je potrebno poznavanje delovanje robotske naprave in ustrezno usposobljeni strokovni delavec, v našem primeru delovni terapevti. Za zagotavljanje varnosti je med izvajanjem obravnav potreben stalni nadzor delovnega terapevta.

Po zaključeni raziskavi smo ugotovili, da robotska naprava za rehabilitacijo zgornjega uda vpliva le na posamezne sklepe zgornjega uda, medtem ko pri klasičnih delovno-terapevtskih pristopih v obravnavi zajamemo celotno področje zgornjega uda ter posturalno kontrolo trupa. Po izvedenih terapijah z uporabo robotske naprave je bil zaznan večji napredek na področju fine

motorike zgornjih udov, po izvedenih terapijah s klasičnimi pristopi delovne terapije pa na področju grobe motorike.

Na podlagi te raziskave ugotavljamo smiselnost kombinacije obeh pristopov, kjer zajamemo celostno področje funkcije zgornjih udov.

Literatura

[1] Jorgensen HS, Nakayama H, Raaschou HO, Vive -Larsen J, Stoier M, Olsen TS. Outcome and time course of recovery in stroke. Part II: time course of recovery the Copenhagen stroke study. Arch Phys Med Rehabil 1995; 76: 406–12.

[2] Hayward KS, Kramer SF, Thijs V, Ratcliffe J, Ward NS, Churilov L et al. A systematic review protocol of timing, efficacy and cost effectiveness of upper limb therapy for motor recovery post-stroke. Syst Rev 2019; 8 (1): 187.

[3] Rong D, Zhang M, Ma Q, Lu J, Li K. Corticospinal tract change during motor recovery in patients with medulla infarct: a diffusion tensor imaging study. Bio-med Res Int 2014; 2014: 524096

[4] Bregant, T. in sod., (2024). Rehabilitacija roke pri otrocih in mladostnikih po možganski okvari. Slovenska pediatrija, 31, str. 180–187. doi.org/10.38031/slovpediatr-2024-4-02.

[5] Bregant T, Derganc M, Neubauer D. Uporaba magnetnoresonančnega slikanja z difuzijskimi tenzorji v pediatriji. Zdrav Vestn 2012; 81: 533–42

[6] Yoo YJ, Kim JW, Kim JS, Hong BY, Lee KB, Lim SH. Corticospinal tract integrity and long-term hand function prognosis in patients with stroke. Front Neurol 2019; 10: 374

[7] Dalamagkas K, Tsintou M, Rathi Y, O'Donnell LJ, Pasternak O, Gong X et al. Individual variations of the human corticospinal tract and its hand-related motor fibers using diffusion MRI tractography. Brain Imaging Behav 2020; 14(3): 696–714.

[8] Kamper DG, Fischer HC, Cruz EG, Rymer WZ. Weakness is the primary contributor to finger impairment in chronic stroke. Arch Phys Med Rehabil 2006; 87: 1262

[9] Zorowitz RD, Gillard PJ, Brainin M. Poststroke spasticity: sequelae and burden on stroke survivors and caregivers. Neurology 2013; 80: S45–52.

[10] Li S. Spasticity, motor recovery, and neural plasticity after stroke. Front Neurol 2017; 8: 120.

[11] Classen J, Liepert J, Wise SP, Hallett M, Cohen LG. Rapid plasticity of human cortical movement representation induced by practice. J Neurophysiol 1998; 79: 1117–23.

[12] Cinnera AM, Bonnì S, D'Acunto A. Cortico-cortical stimulation and robot-assisted therapy (CCS and RAT) for upper limb recovery after stroke: study protocol for a randomised controlled trial. Trials 2023; 24: 823

[13] Giggins OM, Persson UM, Caulfield B. Biofeedback in rehabilitation. J Neuroeng Rehabil 2013; 10: 60

[14] Morone G, Spitoni GF, De Bartolo D, Ghanbari Ghooshchy S, Di Iulio F Paolucci S et al. Rehabilitative devices for a top-down approach. Expert Rev Med Devices 2019; 16 (3): 187–95

[15] Poli P, Morone G, Rosati G, Masiero S. Robotic technologies and rehabilitation: new tools for stroke patients' therapy. Biomed Res Int 2013; 2013: 153872.

[16] Hatem SM, Saussez G, Della Faille M, Prist V, Zhang X, Dispa D et al. Rehabilitation of motor function after stroke: a multiple systematic review focused on techniques to stimulate upper extremity recovery. Front Hum Neurosci 2016; 10: 442.

Staring, Guessing, and Imagining: Strategies in Visual Working Memory

Benjamin Bušelič
University of Ljubljana, Faculty of Arts, Department of Psychology, Slovenia
benjamin.buselic@ff.uni-lj.si

Nina Purg Suljič
University of Ljubljana, Faculty of Arts, Department of Psychology, Slovenia
nina.purg@ff.uni-lj.si

Andrej Jablanovec
University of Ljubljana, Faculty of Arts, Department of Psychology, Slovenia
andrej.jablanovec@gmail.com

Grega Repovš
University of Ljubljana, Faculty of Arts, Department of Psychology, Slovenia
grega.repovs@ff.uni-lj.si

Anka Slana Ozimič
University of Ljubljana, Faculty of Arts, Department of Psychology, Slovenia
Anka.SlanaOzimic@ff.uni-lj.si

Abstract

Although working memory (WM) capacity is often treated as a stable limit, performance in WM tasks is not determined by capacity alone. Emerging evidence suggests that it is also influenced by the strategies individuals adopt to meet task demands – a factor that remains insufficiently explored. This study investigated how strategy use in visual working memory varies depending on the specific requirements of the task, namely the features and combinations of features of visual stimuli to be remembered. Forty-eight students completed a visual WM span task in which they had to remember colors, shapes or both properties of visual stimuli. When both features had to be remembered, colors and shapes were either presented in separate objects (*both separate* condition) or combined within the same objects (*both integrated* condition). Following each task condition, participants reported how often they had used specific strategies by completing a strategy questionnaire. Results showed that visually oriented strategies (e.g., focusing on visual features and imagery) were most common across all conditions. Significant task condition effects emerged for the *staring* and *guessing* strategies, which were reported most often in the *both separate* condition. Furthermore, *active pattern search* was positively correlated with WM span in the *colors* condition, while *passive waiting* was negatively correlated with WM span in the *both separate* condition. These findings highlight that performance in WM tasks reflects not only capacity limits but also the strategies individuals adopt.

Keywords

Visual working memory, working memory strategies, task condition, working memory span

Information Society 2025, 6–10 October 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.cogni.11

1 Introduction

The current understanding of working memory (WM) is often based on the multicomponent model of WM proposed by Baddeley and Hitch [1] who conceptualized WM as a system used for the short-term maintenance and manipulation of information [2]. Previous research [3, 4] has highlighted the importance of WM in everyday tasks, including language and reading comprehension, problem solving, and learning.

Given this central role in cognition, researchers have been particularly interested in the capacity limits of WM. Initial efforts to estimate this capacity suggested that it is highly limited. Miller [5] introduced the concept of the magical number 7 ± 2, describing individuals' WM capacity as the ability to retain approximately 7 ± 2 units of information. However, later studies, particularly in the domain of visual WM, have proposed even lower estimates. Cowan [6] estimated the capacity of visual WM to be closer to 3-4 items.

While these estimates help define the capacity limits of WM, task performance is not determined by capacity alone. Individuals often employ strategies that allow them to optimize how information is encoded and maintained. Such strategies do not increase capacity per se but can improve task performance by making more efficient use of available capacity.

Miller [5] described the phenomenon of chunking, a strategy in which individuals combine separate units of information into larger, meaningful ones (e.g., instead of remembering numbers 2 and 3 separately, they are stored together as 23). Subsequent research identified other strategic approaches, as a means to enhance performance [7].

More recent studies have taken a more open-ended approach to investigating WM strategies, allowing participants to report strategies they had spontaneously used, rather than limiting them to narrow predefined categories. For example, Oblak et. al. [8] used qualitative methods to explore individuals' experiences during a WM task, identifying a variety of strategies employed. Building on this work, Slana Ozimič et. al. [9] reported that the strategy use depends on specific task conditions.

However, previous research has either examined strategy use in a very open-ended manner – typically through interviews or free-

response formats – or has focused on a narrow set of predefined strategies. What has been lacking is a structured yet comprehensive approach to quantitatively assess a broad range of strategies across task conditions. To address this gap – and building on previous literature (e.g., [8–10]) - we developed a structured questionnaire that included a broad set of strategies relevant to visual WM tasks. Using this questionnaire, we examined whether different task conditions encourage the use of different strategies, and whether the spontaneous use of such strategies is related to individuals' WM performance.

2 Methods

2.1 Participants

The study included 48 students (38 female, 8 male, 2 other), aged between 18 and 27 years (M = 19.98 years, SD = 2.23 years). None of the participants reported neurological diseases or conditions and all participants had normal or corrected-to-normal vision.

2.2 Behavioral task and strategy questionnaire

A behavioral task was used to assess visual WM span. The task was presented on a Windows 11 computer using PsychoPy (v2023.1.1), and each participant completed two sessions lasting about 60 minutes. The task included four conditions (two conditions per session), the order of which was pseudo-randomized across participants. In the *colors* condition, participants had to memorize the colors of circles; in the *shapes* condition, they had to memorize the shapes of black outlines. In the *both separate* condition, they were presented with an equal number of colored circles and shape outlines and were asked to remember both features. In the *both integrated* condition, each presented object combined both features—a unique shape filled with a unique color—and participants were instructed to remember both the shape and the color of each object. In each trial, participants were presented with objects defined by color and/or shape for 500 ms. After a 2 s delay interval, they selected from the array of all possible colors and/or shapes those they remembered being shown, by clicking on them (up to the number originally presented). The number of stimuli increased until a stable WM span was obtained in each task condition. Throughout the trial, participants continually repeated the syllables »ta-ma« to suppress verbal rehearsal.

After each task condition, participants completed strategy questionnaire, consisting of 37 items, each formulated as a statement describing a possible strategy (e.g., *"While viewing the stimuli, I actively searched for a pattern in the presented items"*). The items were grouped into three phases of working memory (encoding, maintenance, and recall) and included visual, spatial, verbal, motor, auditory, long-term memory, and transmodal strategies. Participants reported, for each statement, the estimated frequency of its use during the preceding condition, expressed as a percentage.

2.3 Data analysis„

Data were analyzed using R [11]. To assess the effect of task condition on the frequency of strategy use, one-way ANOVAs were conducted separately for each strategy, with task condition (*color, shape, both integrated, both separate*) as the independent variable, and the frequency of strategy use as the dependent variable. In addition, Pearson's correlation analyses were performed to examine relationships between strategy use and WM span within each task condition. All statistical tests were performed separately for each strategy, while FDR corrections were applied within task phases – encoding, maintenance and recall – to reflect the grouping of strategies by phase.

3 Results

First we examined the internal consistency of the questionnaire, which was excellent (Cronbach's $\alpha = .88$), with an average inter-item correlation of .16, indicating that items were related but not redundant. We then examined the mean self-reported frequency of strategy use across all task conditions. The three most frequently reported strategies during encoding were *identifying distinctive features*, *inspecting visual features*, and *representing*. During the maintenance phase, participants predominantly relied on *afterimage*, *rehearsing a visual image*, and *impression*, whereas in the recall phase, the most frequently endorsed strategies were *comparing with a visual image*, *hunch*, and *applying verbal descriptions*. Strategies that were, on average, used in less than 20% of trials were excluded from further analyses, as their low overall frequency suggested limited relevance for interpreting task performance (Figure 1).

On the remaining strategies, we conducted one-way ANOVAs to test for differences in strategy frequency across task conditions. After applying FDR correction, two strategies (*staring* and *guessing*) showed significant task condition effects. For *staring*, a significant effect of task condition was found, $F(3, 183) = 5.99$, $p = .018$, $\eta^2 = .09$, indicating small-to-medium effect size. *Staring* was reported most frequently in the *both separate* condition, followed by the *shapes* and *both integrated* conditions, and least frequently in the *colors* condition. Post hoc comparisons using Tukey's tests revealed that the significant effect of condition was primarily driven by higher reported use of the *staring* strategy in *both separate* condition compared to the *colors* condition (mean diff. = 22.82 %, SE = 7.14 %, $p < .001$).

For *guessing*, there was also a significant effect of task condition, $F(3, 183) = 5.28$, $p = .022$, $\eta^2 = .08$, indicating small-to-medium effect size. Tukey's post hoc comparisons indicated that *guessing* was reported significantly more often in *both separate* condition compared to the *colors* condition (mean diff. = 19.29 %, SE = 6.71 %, $p = .001$), and s*hape* condition (mean diff. = –14.51 %, SE = 6.71 %, $p = .024$).

Lastly, correlation analyses were conducted to examine relationship between the self-reported frequency of each strategy use and WM span within each task condition. After FDR correction only two correlations remained statistically significant. Use of *establishing a pattern* strategy positively correlated with WM span in the *colors* condition ($r = .45$, $p = .045$), while use of *waiting* strategy negatively correlated with WM span in the *both separate* condition ($r = -.44$, $p = .016$).

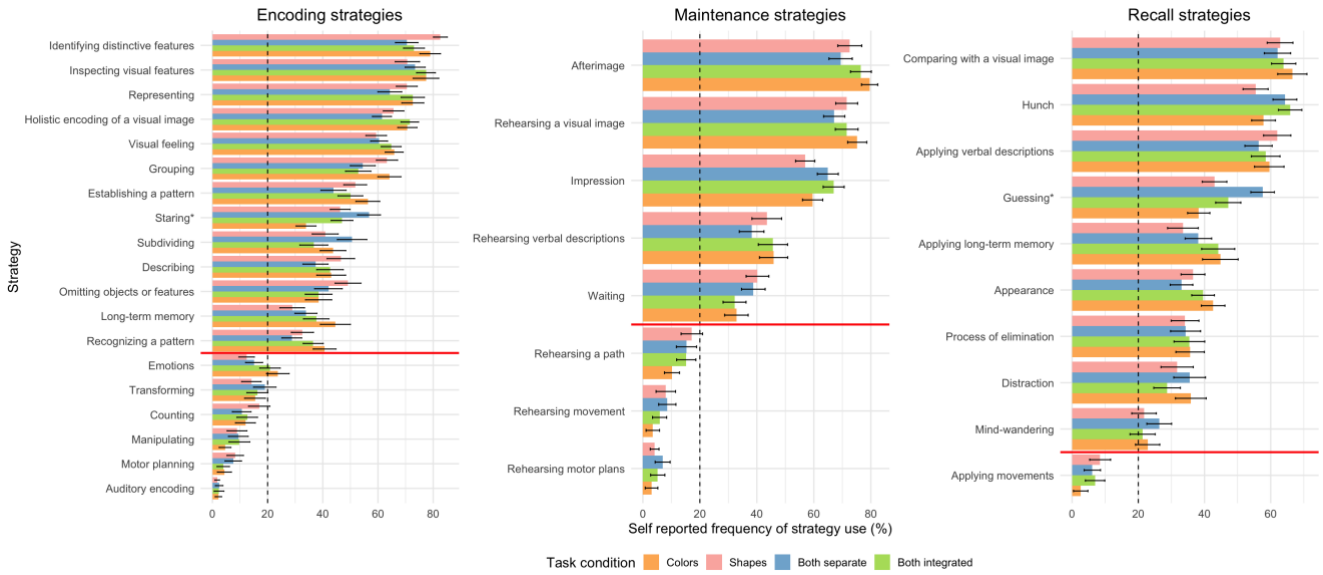


Figure 1: Average self-reported frequency (%) of strategy use across four task conditions

Note. Error bars represent ±1 *SE*. Black vertical dotted lines represent 20% cut-off. Strategies below red horizontal lines were excluded from further analyses; * Strategies with statistically significant one-way ANOVAs after FDR correction.

4 Discussion

The aim of this study was to examine strategies individuals use under different visual WM task conditions, and whether the use of these strategies is related to WM performance. Using a newly developed, literature-based strategy questionnaire, the findings show that the use of strategies during WM tasks differs across task conditions, consistent with previous findings [9].

The most commonly reported strategies across all task conditions were visually based, such as focusing on the visual features of the stimuli (*identifying distinctive features*), relying on afterimage, or mentally comparing the current image with the one stored in WM (*comparing with a visual image*). The predominance of visual strategies is consistent with the nature of the task, which required remembering visual properties – colors and shapes – and thus naturally engages visual encoding and maintenance mechanisms [12]. In contrast participants rarely used motor strategies (e.g., *motor planning*, *rehearsing motor plans*), likely because such strategies are more effective in tasks involving spatial or movement-related information [13].

Significant differences between task conditions emerged for the *staring* strategy, which reflects a passive approach where participants simply looked at the screen with the hope of remembering the stimuli, and the *guessing* strategy, characterized by providing a response without confidence or clear memory of the stimuli. Both strategies showed a similar pattern of use across task conditions: they were reported most frequently in the *both separate* condition, followed by the *shapes* and *both integrated* conditions, and least frequently in the *colors* condition. Post hoc analyses indicated that the difference in *staring* was driven by higher reported use in the *both separate* compared to the *colors* condition, while for *guessing*, significant differences were found between the *both separate* condition and the *shapes* condition and the *colors* conditions. This pattern suggests that participants were more likely to rely on passive or less effortful strategies when task demands increased – particularly when multiple visual features had to be encoded simultaneously in separate objects. Similar findings have been reported in research showing that individuals tend to adopt less demanding strategies as task complexity and cognitive load increases [14].

Finally, our analysis showed that the use of two specific strategies was significantly related to WM span. In the *colors* condition, participants who more frequently *established a pattern*, showed larger WM span, suggesting that combining colors into meaningful patterns supported memory performance. In contrast, in the *both separate* condition, greater reliance on *waiting* – waiting for the prompt to provide the answer – was associated with lower WM span. This indicates that disengaging from active retrieval processes hindered performance in more complex tasks.

The present findings demonstrate that using a structured questionnaire allowed us to identify specific links between WM strategy use and performance - something that was not captured in our previous study [9], which relied on an open-ended interview approach.

5 Conclusion

Taken together these findings suggest that WM strategies play an important role in the dynamic processes underlying WM. The complexity of these processes cannot be captured by WM span alone. While WM span provides useful estimate of capacity, it does not account for the individual differences in strategy use that may influence task performance. Beyond the laboratory, such strategies are likely engaged in everyday contexts, for example, when navigating environments, remembering instructions, or interpreting visual information in educational or occupational settings. Future research in this field should further

examine variability in the deployment of strategies, including how such strategies manifest on a neural level.

Acknowledgements

This work was supported by the Slovenian Research and Innovation Agency (Z5-50177 to N.P.S., J7-5553, J3-9264 and P3-0338 to G.R.).

References

[1] Baddeley, A.D. and Hitch, G. 1974. Working Memory. *Psychology of Learning and Motivation*. Elsevier. 47–89.

[2] Baddeley, A. 2012. Working Memory: Theories, Models, and Controversies. *Annual Review of Psychology*. 63, 1 (Jan. 2012), 1–29. https://doi.org/10.1146/annurev-psych-120710-100422.

[3] Takeuchi, H., Taki, Y. and Kawashima, R. 2010. Effects of Working Memory Training on Cognitive Functions and Neural Systems. *Reviews in the Neurosciences*. 21, 6 (Jan. 2010). https://doi.org/10.1515/REVNEURO.2010.21.6.427.

[4] Unsworth, N., Redick, T.S., Heitz, R.P., Broadway, J.M. and Engle, R.W. 2009. Complex working memory span tasks and higher-order cognition: A latent-variable analysis of the relationship between processing and storage. *Memory*. 17, 6 (Aug. 2009), 635–654. https://doi.org/10.1080/09658210902998047.

[5] Miller, G.A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*. 63, 2 (Mar. 1956), 81–97. https://doi.org/10.1037/h0043158.

[6] Cowan, N. 2010. The Magical Mystery Four: How Is Working Memory Capacity Limited, and Why? *Current Directions in Psychological Science*. 19, 1 (Feb. 2010), 51–57. https://doi.org/10.1177/0963721409359277.

[7] Morrison, A.B. and Chein, J.M. 2011. Does working memory training work? The promise and challenges of enhancing cognition by training working memory. *Psychonomic Bulletin & Review*. 18, 1 (Feb. 2011), 46–60. https://doi.org/10.3758/s13423-010-0034-0.

[8] Oblak, A., Slana Ozimič, A., Repovš, G. and Kordeš, U. 2022. What Individuals Experience During Visuo-Spatial Working Memory Task Performance: An Exploratory Phenomenological Study. *Frontiers in Psychology*. 13, (May 2022), 811712. https://doi.org/10.3389/fpsyg.2022.811712.

[9] Slana Ozimič, A., Oblak, A., Kordeš, U., Purg, N., Bon, J. and Repovš, G. 2023. The Diversity of Strategies Used in Working Memory for Colors, Orientations, and Positions: A Quantitative Approach to a First-Person Inquiry. *Cognitive Science*. 47, 8 (Aug. 2023), e13333. https://doi.org/10.1111/cogs.13333.

[10] Gonthier, C. 2021. Charting the Diversity of Strategic Processes in Visuospatial Short-Term Memory. *Perspectives on Psychological Science*. 16, 2 (Mar. 2021), 294–318. https://doi.org/10.1177/1745691620950697.

[11] R Core Team 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

[12] Van Ede, F. 2020. Visual working memory and action: Functional links and bi-directional influences. *Visual Cognition*. 28, 5–8 (Sept. 2020), 401–413. https://doi.org/10.1080/13506285.2020.1759744.

[13] Purg Suljič, N., Kraljič, A., Rahmati, M., Cho, Y.T., Slana Ozimič, A., Murray, J.D., Anticevic, A. and Repovš, G. 2024. Individual differences in spatial working memory strategies differentially reflected in the engagement of control and default brain networks. *Cerebral Cortex*. 34, 8 (Aug. 2024), bhae350. https://doi.org/10.1093/cercor/bhae350.

[14] Tavares, W., Ginsburg, S. and Eva, K.W. 2016. Selecting and Simplifying: Rater Performance and Behavior When Considering Multiple Competencies. *Teaching and Learning in Medicine*. 28, 1 (Jan. 2016), 41–51. https://doi.org/10.1080/10401334.2015.1107489.

Machine bias: new experiments with COMPAS data*

Ana Farič†
Faculty of Education
University of Ljubljana
Slovenia
af27987@student.uni-lj.si

Ivan Bratko
Faculty of Computer and Information Science
University of Ljubljana
Slovenia
bratko@fri.uni-lj.si

Abstract

This paper revisits the debate on machine bias through an analysis of the COMPAS recidivism prediction system. While some studies claim COMPAS is racially biased and others argue the opposite, our replication and extension of prior work show that across diverse methods accuracy consistently converges at around 66-67%. Moreover, error distributions follow a stable pattern: higher false positive rates for black defendants and higher false negative rates for white defendants. We argue that this convergence reflects inherent difficulty of this prediction problem and probably yet unexplained asymmetries in this domain. Our findings suggest that debates on fairness should move beyond model choice to address systemic disparities that shape observed outcomes.

Keywords

artificial intelligence, machine bias, fairness, COMPAS system

1 Introduction

Calls for unbiased AI systems are increasingly more common in regulation debates. For example, in September 2024, one of the GPAI (*Global Partnership on Artificial Intelligence*) working groups released a report [13] recommending that AI system providers be held liable for discriminatory impacts and required to compensate individuals harmed by algorithmic bias. Although the group attempted to clarify and better define the notion of bias in the revised report [14], released in November 2024, it ultimately offered no practical metrics or other criteria to determine with confidence whether a system is biased or not. This highlights a broader challenge: while the demand for unbiased AI systems is growing, even well-intentioned policymakers struggle to translate abstract concepts of fairness into actionable, measurable criteria. The COMPAS recidivism prediction system exemplifies these definitional difficulties. Some studies claim COMPAS is racially biased, while others disagree, depending on which fairness metric is applied. Beyond the technical debate, this inconsistency raises a deeper question: to what extent do observed disparities reflect the various models versus the underlying data and social context?

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

Information Society 2025, 6–10 October 2025, Ljubljana, Slovenia

http://doi.org/10.70314/is.2025.cogni.13

2 Understanding the concept of machine bias

In computer science, dozens of fairness metrics and bias definitions exist, often in contradiction with one another [16, 17, 18, 19]. Philosophers, legal scholars, and social scientists have long debated the meaning of bias, and computer scientists face the additional challenge of operationalizing these abstract concepts into measurable criteria [12]. Despite numerous attempts to resolve this ambiguity, no consensus has emerged. Even mathematical definitions, while precise, often lack concrete examples that would make them applicable in real-world decision-making contexts.

Bias in machine learning (ML) is a multifaceted concept, encompassing both technical and social dimensions. Researchers identify three broad types of bias:
1. Inductive/learning bias: in supervised learning, an algorithm seeks a function that predicts outcomes from data. Many functions may fit the training data, but most fail to generalize well. Preferential bias is needed to select certain functions over others, guiding learning toward useful generalizations. As such, bias is a necessary component that enables learning [16].
2. Historical bias: reflects real-world prejudices embedded in the data. As such, even perfectly measured data may produce biased outcomes if the underlying reality is discriminatory [1, 16, 19].
3. Biases that arise during data generation: specification, measurement/observation, sampling/population, annotator bias etc. [5, 16, 24].
In practice, bias is most often discussed in terms of its social consequences, such as when models classify individuals differently based on protected attributes like race or gender [16].

Computer scientists have formalized bias through various fairness metrics, including:
1. Demographic parity: equal positive prediction rates across groups [16];
2. Equalized odds: equal false positive (FPR) and false negative rates (FNR) across groups [15];

3. Predictive parity: equal prediction accuracy across groups [23].
These metrics are mutually incompatible with the exception of certain very trivial conditions [17].
For a more comprehensive survey of existing bias definitions and their limitations, we refer the reader to our previous work [9].

3 COMPAS

A good example that demonstrates the problem of a lack of a unified definition is COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*). It is a model still used by US courts to assess the likelihood that a defendant will reoffend within two years of the evaluation date if released [2, 20]. The assessment is based on 137 attributes about the defendant, including personal information and criminal history. Race is not considered in the evaluation [8]. The model assists judges in making decisions about bail and sentencing, particularly in determining whether defendants awaiting trial are too dangerous to be released [6].

The model assigns defendants a score between 1 and 10, indicating how likely they are to reoffend. These scores have a significant impact on the lives of the defendants. Those rated as medium- or high-risk (scores 5-10) are more often held in detention until trial, while low-risk defendants (scores 1-4) are more frequently released [2, 6].

3.1 Previous research on COMPAS

In 2016, ProPublica [2] sparked an intense debate by claiming that COMPAS was biased against black defendants. Over the following years, researchers reached contradictory conclusions, highlighting the difficulty of assessing bias in this system. Northpointe, the developer of COMPAS, rejected ProPublica's claims, arguing that ProPublica's analysis was methodologically flawed, and that ProPublica should have used standard fairness measures such as AUC-ROC, under which COMPAS showed no racial bias [7]. Similarly, Flores et al. [10] argued that there is no significant difference in predictive accuracy between white and black defendants. In AI literature, the negative assessments of COMPAS prevail.

Subsequent studies further complicated the debate. Dressel and Farid [8] showed that COMPAS performs no better than laypeople in predicting recidivism, and that a simple linear classifier with only two or seven attributes produces accuracy results comparable to COMPAS's 137-attribute system. Rudin [22] reached a similar conclusion with a three-rule interpretable model based on just two attributes. These findings questioned the added value of complex risk assessment tools in this domain.

Other research emphasizes inherent trade-offs in fairness metrics. Corbett-Davies et al. [6] and Zafar et al. [27] all highlight the impossibility of simultaneously satisfying competing fairness definitions, given differing base rates of recidivism across racial groups and different fairness metrics. Conceptual and formal tensions such as these, help explain why analyses of COMPAS produce conflicting assessments of the same system [9].

4 Our analysis

We noticed an intriguing pattern in the previously mentioned studies: models of varying complexity produce comparable accuracy, and FNR and FPR. While these results have been reported independently, they have not been systematically compared within the same analytical framework, nor have their broader implications been thoroughly examined. In this paper, we aim to replicate prior findings to confirm their robustness, and to extend the discussion by investigating why such convergence occurs across different methods.

4.1 Method

We used the publicly available COMPAS dataset released by ProPublica [2] on GitHub (https://github.com/propublica/compas-analysis). To ensure comparability with previous studies, we selected the same version of the dataset as used by [8]. The dataset contains 53 attributes, including demographic information, criminal history, and COMPAS risk scores, including protected attributes such as race and sex, for 7214 defendants from Broward County, Florida. Following the previous researchers, we filtered the dataset to include only black and white defendants, resulting in final 6150 individuals.

We trained the following models using the Orange data mining platform, applying an 80%/20% training/testing split that was repeated 10 times to compute true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), both overall and separately for black and white defendants:
1. Logistic regression: simple linear classifier, trained with either 6 (sex, age, prior crimes, crime degree, number of juvenile misdemeanors and felonies) or 2 attributes (age and priors). We excluded the crime description attribute (used by [8]) because it contains over 400 different values, which they reduced to 63 for human judgement purposes. As we could not reproduce this exact transformation, we omitted it to test whether the remaining attributes alone suffice.
2. Decision tree: was constructed to approximate Rudin's [22] rule-based model. Two attributes (age category and priors) were used and its depth was limited to 5.

Models were evaluated using:
1. Accuracy: proportion of correct predictions on test set.
2. FPR: proportion of non-recidivists incorrectly predicted to reoffend.
3. FNR: proportion of recidivists incorrectly predicted not to reoffend.
For each model, TP, TN, FP, and FN were first recorded for each of the 10 repetitions. These counts were then pooled

across repetitions, and accuracy, FPR and FNR were calculated from the pooled counts for each race group. Finally, metrics were averaged across all repetitions to produce values reported in tables 1 and 2.

The results from our models were directly compared to reported metrics from [2, 8, 22], allowing us to compare predictive performance across models.

4.2 Results

The results from previous researchers are summarized in table 1. Our results are summarized in table 2.

Across all methods (table 2), overall accuracy converged around 66-67%, consistent with the performance reported by ProPublica [2] and Dressel and Farid [8]. While our exact error rates differ somewhat from those reported previously, the same pattern in error distribution was observed; black defendants exhibited higher FPR, whereas white defendants exhibited higher FNR. Finally, compared to [8], who incorporated a reduced version of the *crime description* attribute, our results suggest that excluding this feature does not substantially change performance.

Table 1: Columns A-E summarize predictive performance across different models and conditions from previous researchers. Column A reports human judgements without access to information about race; B reports human judgements with race, C shows COMPAS predictions as reported by ProPublica, D and E show logistic regression (LR) models trained on 7 or 2 attributes respectively. Accuracy (CA), FPR and FNR are reported overall and separately for two races.

	A: human (no race)	B: human (race)	C: COMPAS	D: LR-7	E: LR-2
CA (overall)	67.0%	66.5%	65.2%	66.6%	66.8%
CA (black)	68.2%	66.2%	64.9%	66.7%	66.7%
CA (white)	67.6%	67.6%	65.7%	66.0%	66.4%
FPR (black)	37.1%	40.0%	40.4%	42.9%	45.6%
FPR (white)	27.2%	26.2%	25.4%	25.3%	25.3%
FNR (black)	29.2%	30.1%	30.9%	24.2%	21.6%
FNR (white)	40.3%	42.1%	47.9%	47.3%	46.1%

Table 2: Columns A-D summarize predictive performance of our models. A shows LR trained on 6 attributes, excluding race, B shows LR trained on the same 6 attributes with race included; C shows LR trained on 2 attributes, D shows a decision tree (DT) trained on 2 attributes. CA, FPR and FNR are reported overall and separately for black and white defendants.

	A: LR-6 (no race)	B: LR-6 (race)	C: LR-2	D: DT-2
CA (overall)	67.2%	67.1%	66.5%	66.8%
CA (black)	66.9%	67.2%	66.7%	66.8%
CA (white)	67.1%	66.4%	66.1%	67.7%
FPR (black)	29.0%	31.1%	31.1%	35.2%
FPR (white)	15.1%	15.2%	16.5%	20.1%
FNR (black)	36.7%	34.5%	35.4%	31.3%
FNR (white)	61.7%	61.4%	60.6%	51.0%

Figures 1-3 present an example of a decision tree trained on two attributes (age category: 1. < 25, 2. 25-45, 3. > 45) and number of prior offences, with the tree depth limited to 5. The tree splits defendants into subgroups, with the leaves representing predicted recidivism risk (0=predicted not to reoffend, 1=predicted to reoffend) and the proportion of majority class.

To improve readability, the tree is divided into three figures: figure 1 shows the root node and the initial split by the number of priors, figure 2 shows the left subtree (defendants with less or equal 2 priors), and figure 3 shows the right subtree (defendants with > 2 priors). Among these, defendants older than 45 are further divided by priors. The utmost right leaf in figure 3 predicts that defendants with more than 20 priors will reoffend (1), with probability 76.9%.

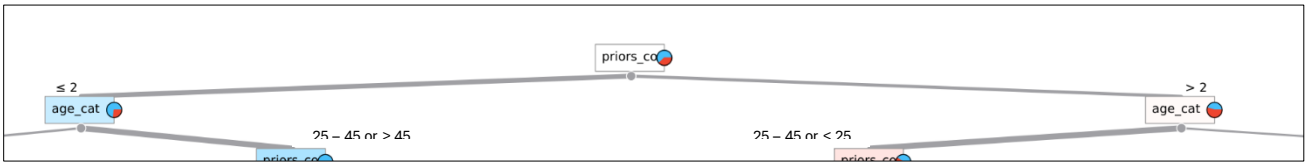


Figure 1: Root node of the decision tree and the initial split into left and right subtrees.

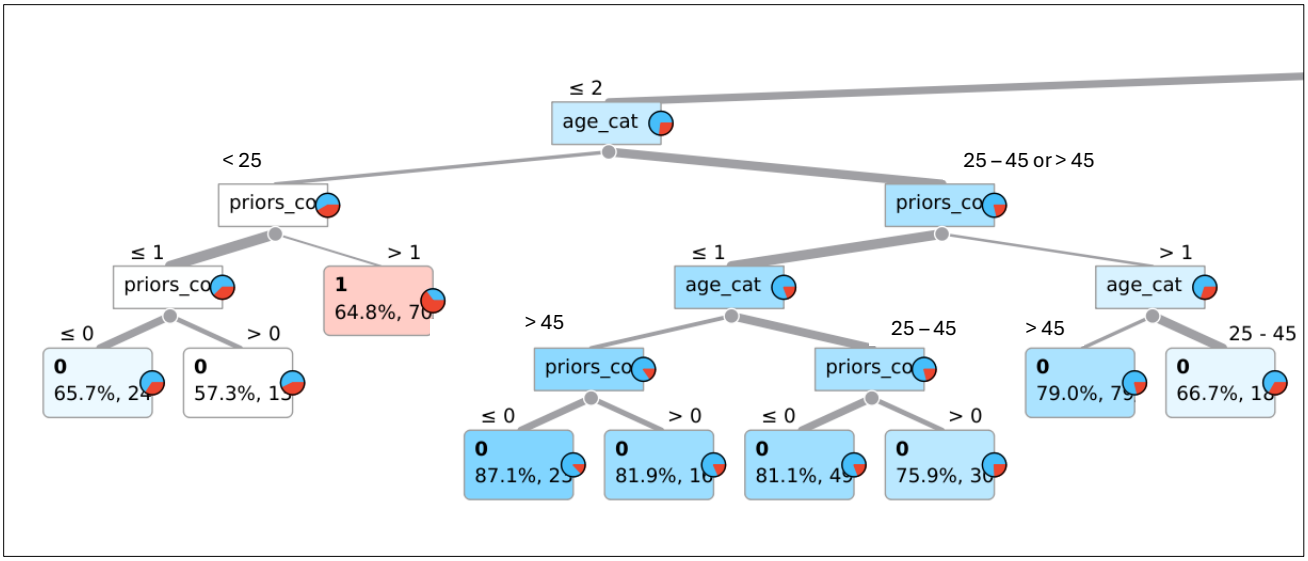


Figure 2: Left subtree of the decision tree

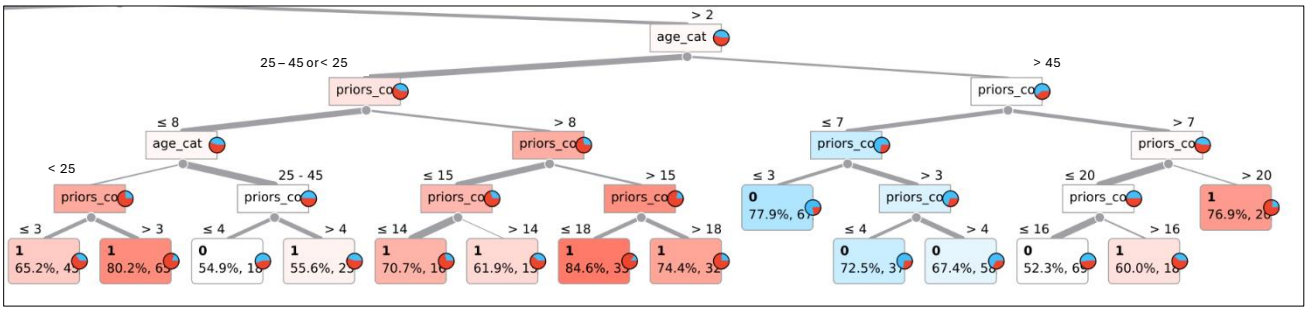


Figure 3: Right subtree of the decision tree

To further illustrate how the decision tree makes predictions, figures 4 and 5 show the distribution of defendants by race (the first two columns in both figures correspond to black defendants, the right ones to white defendants) within two example leaves. The bar charts make it clear that, although the prediction is the same within each leaf (1 = will reoffend, shown in red; the blue color indicates the number of defendants who did not reoffend), the underlying racial composition of these subgroups can vary substantially. Additionally, the figures illustrate how estimated prediction errors and class balance differ across leaves.

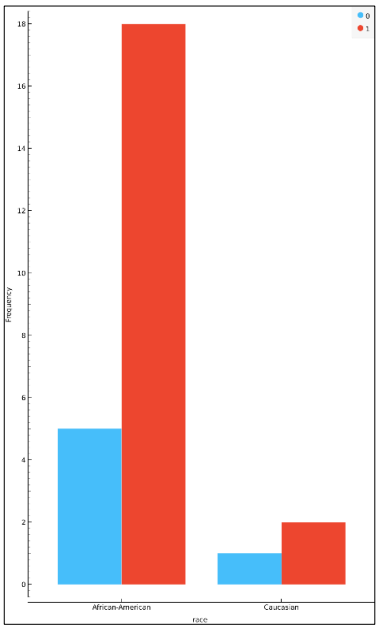


Figure 4: Distribution of defendants by race in the leaf (right-most leaf in figure 3) corresponding to the path: > 2 priors --> age > 45 --> > 20 priors. All defendants in this subgroup are predicted to reoffend, with 76.9% probability.

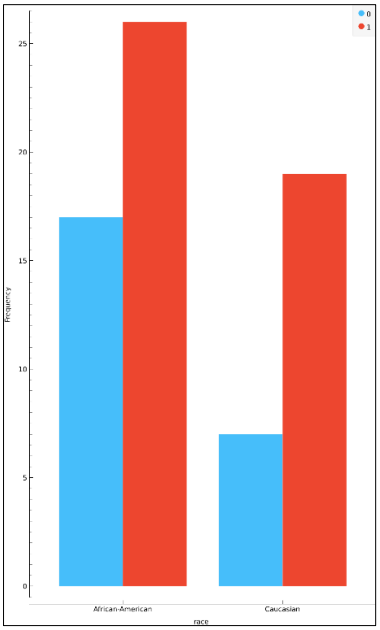


Figure 5: Distribution of defendants by race in the leaf (left-most leaf in figure 3) corresponding to the path: > 2 priors --> age < 25 --> 3 priors. All defendants in this subgroup are predicted to reoffend with probability 65.2%.

4.3 Discussion

Our findings confirm and extend those of [2, 8, 22]. Across models of varying nature and complexity (black box, logistic regression, interpretable rule-based, and even human judgement) predictive accuracy consistently hovers around 66-67%. Moreover, we replicated the characteristic error distribution pattern; higher FPR for black defendants, and higher FNR for white defendants. We extend the mentioned prior research by demonstrating this convergence using decision trees as an approximation of Rudin's [22] interpretable rules.

While [22] emphasizes the use of inherently interpretable models, and [8] question the overall utility of algorithmic recidivism prediction, our work shifts the discussion toward the underlying reasons why all these methods yield similar results, in particular similar error patterns.

The convergence of predictions across methods suggests that the limitations may lie less in model choice and more in the data and domain itself, which prior analyses often overlook.

Beyond dataset quality, structural factors such as racial disparities in arrests and sentencing likely drive the consistent error patterns observed across all models. As [11] reports, the lifetime likelihood of imprisonment for black men was one in three for those born in 1981, and one in five for those born in 2001. A report from 2018 [21] emphasizes that the imprisonment rate for black adults is 5.9 the rate for white adults – and even higher in some states. These disparities exist for both least and more serious offences; 56% of people imprisoned nationwide for a drug offence are black or Latino, and 48% of people serving life sentences are black. Another report [25] emphasizes that 56.4% of those serving life without parole sentences are black.

Additionally, Williamsons' framework [26] emphasizes that the higher crime rates observed among black individuals are not indicative of inherent crime tendencies, but rather reflect systemic economic disparities which are often the result of historical and ongoing policies that have marginalized black communities, limiting their access to resources and opportunities. Therefore, the convergence of predictive models like COMPAS with other simple ML models and even lay people judgements may not solely be a technical issue but also a reflection of deeper societal inequalities.

While our study confirms agreement across models and highlights the importance of structural factors, several avenues for further research remain. At a methodological level, further work could explore different versions of the ProPublica dataset, test additional feature combinations, and evaluate a wider range of ML models to assess the robustness of these patterns. At a broader level, additional research should examine the underlying systemic factors that drive disparities in recidivism predictions, thus contextualizing algorithmic predictions within real-world social dynamics and inform policy discussions on the responsible use of predictive models in the justice system.

5 Conclusion

Our study revisits the COMPAS recidivism prediction debate by replicating and extending previous findings and discussion why different methods (ranging from black-boxes to simple linear predictors, interpretable rule-based models, and human judgements) consistently converge on similar predictive performance and error patterns. Across all methods, accuracy hovered around 66-67%, with characteristic error distributions showing higher FPR for black defendants and higher FNR for white defendants.

While prior research has documented this convergence, its proper interpretation and broader implications have received less attention. We emphasize that COMPAS may be wrongfully vilified; while skepticism regarding algorithmic risk assessment is warranted, using ML systems to inform decisions should not be dismissed outright, as they hold potential to support informed decision-making if used responsibly.

More importantly, we argue that the debate should shift toward understanding why such convergence occurs. Our findings suggest that it reflects domain-specific and structural factors, including disparities in arrest, sentencing, and systemic socio-economic inequalities that induce observed recidivism rates. By examining these elements, alongside limitations in commonly used datasets, we can better contextualize predictive performance and the persistence of racial disparities.

References

[1] Alelyani, S. 2021. Detection and Evaluation of Machine Bias. *Applied Sciences* 11, 4. DOI:https://doi.org/10.3390/app11146271.

[2] Angwin, J., Kirchner, L., Larson, J. & Mattu, S. 2016. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica.*

[3] Barenstein, M. 2019. ProPublica's COMPAS Data Revisited. *ArXiv:1906.04711.* DOI: https://doi.org/10.48550/arXiv.1906.04711.

[4] Beck, A. J. 2021. *Race and Ethnicity of Violent Crime Offenders and Arrestees, 2018.* U.S. Department of Justice, Statistical Brief.

[5] Chakraborty, J., Majumder, S. & Menzies, T. 2021. Bias in Machine Learning Software: Why? How? What to do? *ESEC/SIGSOFT FSE.* DOI:https://doi.org/10.1145/3468264.3468537.

[6] Corbett-Davies, S., Pierson, E., Feller, A. & Goel, S. 2016. A computer program used for bail and sentencing decisions was labeled against blacks. It's actually not that clear. *The Washington Post.*

[7] Dieterich, W., Mendoza, S. & Brennan, T. 2016. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity.

[8] Dressel, J. & Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1. DOI:https://doi.org/10.1126/sciadv.aao5580.

[9] Farič, A. & Bratko, I. 2024. Machine Bias: A Survey of Issues. *Informatica* 48, 2. DOI: https://doi.org/10.31449/inf.v48i2.5971.

[10] Flores, A. W., Lowenkamp, C. T. & Bechtel, K. 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* 80, 2.

[11] Ghandnoosh, N. 7.12.2023. *One in Five: Racial Disparity in Imprisonment – Causes and Remedies.* The Sentencing Project. https://www.sentencingproject.org/reports/one-in-five-racial-disparity-in-imprisonment-causes-and-remedies/.

[12] Goel, N., Yaghini, M. & Faltings, B. 2018. Non-Discriminatory Machine Learning through Convex Fairness Criteria. *The 23nd AAAI Conference on Artificial Intelligence* 32, 1.

[13] GPAI 2024. *Towards Substantive Equality in Artificial Intelligence: Transformative AI Policy for Gender Equality and Diversity.* Report, September 2024, Global Partnership on AI.

[14] GPAI 2024. *Towards Substantive Equality in Artificial Intelligence: Transformative AI Policy for Gender Equality and Diversity.* Report, November 2024, Global Partnership on AI.

[15] Hardt, M., Price, E. & Srebro, N. 2016. Equality of Opportunity in Supervised Learning. *ArXiv:1610.02413.* DOI:https://doi.org/10.48550/arXiv.1610.02413.

[16] Hellstrom, T., Dignum, V. & Bensch, S. 2020. Bias in Machine Learning – What is it Good for? *ArXiv:2004.006686.* DOI:https://doi.org/10.48550/arXiv.2004.00686.

[17] Kleinberg, J., Mullainathan, S. & Raghavan, M. 2021. Inherent Trade-Offs in the Fair Determination of Risk Scores. *ArXiv:1609.05807.* DOI:https://doi.org/10.48550/arXiv.1609.05807.

[18] Mehrabi, A., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys.* DOI:https://doi.org/10.1145/3457607.

[19] 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. DOI:https://doi.org/10.1002/widm.1356.

[20] Porębski, A. 2023. Machine learning and law. *Research Handbook on Law and Technology.* Cheltenham, UK: Edward Elgar Publishing.

[21] *Report to the United Nations on Racial Disparities in the U.S. Criminal Justice System.* 19.4.2018. The Sentencing Project. https://www.sentencingproject.org/reports/report-to-the-united-nations-on-racial-disparities-in-the-u-s-criminal-justice-system/.

[22] Rudin, C. 2018. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *ArXiv:1811.10154.* DOI:https://doi.org/10.48550/arXiv.1811.10154.

[23] Saravanakumar, K. K. 2021. The Impossibility Theorem of Machine Fairness: A Causal machine learning algorithms interaction. *ArXiv:2007.06024.* DOI:https://doi.org/10.48550/arXiv.2007.06024.

[24] Sun, O., Nasraoui, O. & Shafto, P. 2020. Evolution and impact of bias in human and machine learning algorithms interaction. *PLOS ONE* 15, 8. DOI:https://doi.org/10.1371/journal.pone.0235502.

[25] Walsh, A. 15.8.2016. *The criminal justice system is riddled with racial disparities.* Prison Policy Initiative. https://www.prisonpolicy.org/blog/2016/08/15/cjrace/.

[26] Williamson Kramer, C. 13.2.2024. *Systemic Racism in Crime: Do Blacks Commit More Crimes Than Whites?* Liberty Matters. https://oll.libertyfund.org/publications/liberty-matters/2024-02-13-systemic-racism-in-crime-do-blacks-commit-more-crimes-than-whites.

[27] Zafar, M. B., Valera, I., Gomez Rogrigues, M. & Gummadi, K. P. 2017. Fairness constraints: Mechanisms for fair classification. *ArXiv:1507.05259.* DOI:https://doi.org/10.48550/arXiv.1507.05259.

Primerjava lastnosti človeške kognicije in umetne inteligence

Comparison of the characteristics of human cognition and artificial Intelligence

Monika Jamšek†
Faculty of Public Administration,
University of Ljubljana
Gosarjeva ulica 5
1000 Ljubljana, Slovenia
monika.jamsek@gmail.com

Rok Smodiš
Faculty of Education,
University of Ljubljana
Kardeljeva ploščad 16
1000 Ljubljana, Slovenia
rs68734@student.uni-lj.si

Marko Jordan
Odsek za inteligentne sisteme
Jozef Stefan Institute
1000 Ljubljana, Slovenija
marko.jordan@ijs.com

Matjaž Gams
Odsek za inteligentne sisteme
Jozef Stefan Institute
1000 Ljubljana, Slovenija
matjaz.gams@ijs.com

Povzetek

Referat ponuja poglobljen teoretični okvir, ki primerja ključne kognitivne funkcije človeka in sistemov umetne inteligence (UI). Na podlagi najnovejših raziskav iz nevroznanosti, kognitivne psihologije, umetne inteligence in filozofije uma so predstavljeni procesi pozornosti, spomina, učenja, jezika, kreativnosti in čustev ter njihova analogija v umetnih sistemih. V ospredje so postavljene prednosti in omejitve obeh perspektiv, izpostavljene so etične razsežnosti ter nevarnost kognitivne atrofije, ki jo lahko povzroči pretirana uporaba UI. Osrednji del referata predstavlja pregledna tabela, ki vizualno prikaže razlike, prednosti in slabosti človeka in UI. Posebej je obravnavan koncept hibridne inteligence, ki nakazuje prihodnost sodelovanja obeh oblik inteligence. Zaključek poudarja, da prihodnost ne bo temeljila na nadomeščanju človeka, temveč na komplementarnem partnerstvu, ki odpira novo kognitivno paradigmo.

Ključne besede

človeška kognicija, umetna inteligenca, čustva, etika UI, hibridna inteligenca, primerjalna analiza

Abstract

The paper offers an in-depth theoretical framework that compares key cognitive functions of humans and artificial intelligence (AI) systems. Drawing on the latest research from neuroscience, cognitive psychology, artificial intelligence, and the philosophy of mind, it presents processes such as attention, memory, learning, language, creativity, and emotions, along with their analogs in artificial systems. The focus is on the strengths and limitations of both perspectives, highlighting ethical dimensions and the risk of cognitive atrophy that may result from excessive use of AI. The core of the paper features a comparative table that visually displays the differences, strengths, and weaknesses of humans and AI. Special attention is given to the concept of hybrid intelligence, which points toward a future of collaboration between both forms of intelligence. The conclusion emphasizes that the future will not revolve around replacing humans, but rather around a complementary partnership that opens up a new cognitive paradigm.

Keywords

human cognition, artificial intelligence, emotions, AI ethics, hybrid intelligence, comparative analysis

*Primerjava lastnosti človeške kognicije in umetne inteligence
†Monika Jamšek, Rok Smodiš, Marko Jordan, Matjaž Gams

Information Society 2025, 6–10 October 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.cogni.5

1 Uvod

Raziskovanje človeške kognicije, od zaznavanja in spomina do učenja, jezika, kreativnosti in čustev, je temeljno področje psihologije, nevroznanosti in kognitivne znanosti. Neisser [1] kognicijo opredeli kot procese obdelave informacij za prilagodljivo vedenje, sodobni modeli pa poudarijo integracijo izvršilnih funkcij, čustev in motivacije [2, 3]. Tononi [4] obravnava zavest skozi pet informacijskih aksiomov in teoremov.

Vzporedno je umetna inteligenca, posebej veliki jezikovni modeli (LLM-ji), v zadnjih letih dosegla izjemen napredek: sistematično izboljšujejo rezultate na standardiziranih nalogah, obvladujejo kompleksne dialoge in v številnih neformalnih preizkusih v duhu Turingovega testa sogovorniki pogosto ne prepoznajo, da komunicirajo s strojem (npr. [5]). Ta navzven vidna kompetentnost krepi vtis približevanja človeški ravni. Kljub temu pa je napredek pri "človeških" lastnostih, kot so zavest, fenomenalno doživljanje, globoko semantično sidranje in

metakognitivna samoregulacija, bistveno skromnejši. Potencialna razlaga je, da sodobni modeli še niso razrešili problema mnogoterega znanja – v Gamsovem principu mnogoterega znanja: kako v enoten, konsistenten in samoreferencialen model sveta povezati heterogene oblike znanja (statistične vzorce, simbolne strukture, kavzalne relacije, epizodične in utelešene reprezentacije) ter nad njimi izvajati stabilno vsebinsko procesiranje, razlago pomena in namere. Dokler ta integracijski problem ostaja odprt, lahko UI dosega vrhunske rezultate na nalogah, vendar ne kaže jasnega napredka proti lastnostim, ki jih povezujemo z zavestnim človeškim mišljenjem.

Članek sistematično obravnava, kako daleč ChatGPT (LLM) seže v smeri "kognitivnih" lastnosti, še posebej glede zmožnosti prehoda Turingovega testa in vprašanja zavesti. Avtorja najprej predstavita hierarhijo Turingovih testov – od klasičnega kratkega besednega testa do ekspertno/adverzarnega, fizičnega, "totalnega" in "truly total" testa – ter argumentirata, da se ChatGPT lahko približa uspehu predvsem pri naivnih spraševalcih, medtem ko ga izkušeni/ekspertni spraševalci zlahka razkrijejo.

Ta referat je v nekem smislu nadaljevanje članka [6], ki analizira in podaja oceno "zavesti" po Tononijevi teoriji integrirane informacije (IIT) [4]. Po kratki predstavitvi IIT aksiomov in teoremov je narejena primerjava GPT-jev in ljudi. Zaključek je, da ChatGPT sicer presega starejše AI-sisteme po jezikovni kompetenci in širini znanja, vendar občutno zaostaja za biološkimi sistemi v integraciji informacij, zato nima fenomenalne zavesti; gre za napredno orodje, ne za "informacijsko živo bitje".

Metodološko članek združi pregled literature o Turingovem testu in teorijah zavesti z avtorjevimi lastnimi številnimi ekspertno zasnovanimi dialognimi preizkusi, kjer se pokažejo omejitve modela (npr. pomanjkanje stabilne semantične sidranosti/intencionalnosti). Zaključek poudari praktično implikacijo: LLM-je je smiselno obravnavati kot zelo zmogljiva, a ne-čuteča orodja, s potrebo po previdnosti pred antropomorfizacijo in zavedanju meja pri presoji pomena in namere. Hkrati pa je narejena analiza ogromnih napredkov AI na nekaterih področjih. Ključno je vprašanje, ali se je razvoj ustavil, ali pa se nadaljuje z nezmanjšanim tempom.

V tem prispevku primerjalno obravnavamo ključne kognitivne funkcije človeka in njihove analoge v UI, pri čemer izrecno ločujemo zunanje zmogljivosti na nalogah od notranjih mehanizmov in lastnosti. Predstavimo pregledni okvir, posodobljeno primerjalno tabelo ter razpravo o hibridni inteligenci, kjer kombinacija človeških in umetnih zmožnosti ponuja praktično pot naprej – ob hkratnem opozorilu na etične razsežnosti in tveganje kognitivne atrofije zaradi pretirane rabe UI.

2 Kognitivne funkcije človeka

Človeška kognicija zajema širok spekter medsebojno povezanih procesov, ki skupaj omogočajo prilagodljivo vedenje in kompleksno obdelavo informacij. V tem razdelku sintetiziramo človeške mehanizme kot referenčni okvir za poznejšo primerjavo z UI. Eden od temeljnih mehanizmov je pozornost, ki omogoča usmerjanje miselnih virov na relevantne dražljaje; nevroznanstvene raziskave kažejo, da pri tem ključno vlogo igrajo frontalno-parietalne možganske mreže [7]. Nepogrešljiv del kognitivnega sistema je tudi spomin, ki ga delimo na senzornega, kratkoročnega, delovnega in dolgoročnega. Baddeley [8] je delovni spomin razčlenil na fonološko zanko, vizuoprostorski blok in izvršilni mehanizem, medtem ko Cowan [9] ugotavlja, da kratkoročni spomin praviloma ne preseže štirih enot. Ti mehanizmi tvorijo osnovo za učenje in sklepanje, pri čemer človek pogosto uči iz malo podatkov ter kompozicionalno posplošuje na nove okoliščine. Ravenove matrice so pokazatelj sposobnosti abstraktnega mišljenja [10], medtem ko Stanovich [11] opozarja, da tradicionalni testi inteligence ne zajamejo racionalnosti, ključne za učinkovito odločanje.

Posebno mesto v človeški kogniciji ima jezik, saj omogoča simbolno reprezentacijo, socialno koordinacijo in prenos kulturnega znanja med generacijami. Jezikovne sposobnosti so semantično zasidrane v utelešeni in socialni izkušnji, z jezikom pa je tesno povezana tudi kreativnost, ki presega zgolj divergentno mišljenje in je rezultat interakcije posameznika, njegovega domenskega znanja ter širšega sociokulturnega okolja [12, 13]. Vendar pa na kognitivne procese močno vplivajo tudi čustva in motivacija. Damasio [3] je pokazal, da brez čustvene komponente človeška racionalnost oslabi, medtem ko Immordino-Yang [14] poudarja, da so čustva neločljivo povezana z učenjem, odločanjem in socialnim vedenjem.

Dodatno raven kompleksnosti predstavljata metakognicija in zavest, ki posamezniku omogočata samorefleksijo ter strateško prilagajanje učenja, vključno s fenomenalnim doživljanjem, kar je bistveno za učinkovito regulacijo lastnih miselnih procesov. Nenazadnje pa ima ključno vlogo tudi socialna kognicija, ki vključuje sposobnost razumevanja drugih preko t. i. teorije uma [15], kar je osnova za uspešno sodelovanje, empatijo in gradnjo družbenih odnosov. Ta človeška arhitektura mnogoterega znanja, od senzorno-motoričnih in epizodičnih do simbolnih in kavzalnih reprezentacij, bo v nadaljevanju služila za pojasnilo, zakaj današnji LLM-ji kljub hitremu napredku na nalogah pri lastnostih, kot sta zavest in metakognicija, napredujejo bistveno počasneje [6].

3 Kognitivne lastnosti umetne inteligence

Kognitivne lastnosti obravnavamo kot funkcionalne analoge človeških sposobnosti; pri UI so te uresničene prek drugačnih mehanizmov (statistične reprezentacije, optimizacija, vzorčno ujemanje) kot pri ljudeh, zato ločujemo zmogljivost reševanja nalog od notranjih lastnosti (zavest, fenomenalno doživljanje, metakognitivna samoregulacija).

Zaznavanje in pozornost. V računalniškem vidu so globoke mreže dosegle (in na nekaterih nalogah presegle) človeško raven prepoznave [16]. V jezikovni obdelavi mehanizem porazdeljene pozornosti omogoča učinkovito kontekstno sledenje in gradnjo hierarhičnih reprezentacij [17], vendar ostaja to algoritemska selekcija informacij, ne zavestno usmerjena pozornost z namero [18].

Spomin. LLM-ji shranjujejo »znanje« v parametrih in trenutnih kontekstih; zunanji pomnilniki (RAG ipd.) razširijo doseg, a ne tvorijo epizodičnega ali utelešenega spomina v človeškem smislu [5, 18]. Semantično sidranje je posredno in nestabilno.

Učenje. Napredek temelji na masivnih korpusih in gradientni optimizaciji; zmožnosti few-shot in učenja v kontekstu kažejo na

vzorec-odvisno generalizacijo, toda prenos med domenami in robustnost zunaj učne porazdelitve ostajata omejena [19, 20, 21].

Sklepanje in kavzalnost. Na testih znanja in razumevanja (MMLU) so rezultati visoki [22] ter tudi na problemskih zbirkah (ARC) [21], a zdravorazumsko/kavzalno sklepanje ostaja krhko; to se vidi tudi na družini izpeljank Winograd/Schema [23]. Verbalizacija korakov ne implicira resnične kavzalne strukture [20].

Jezik. Tekočnost, slog in koherenca so blizu človeški ravni, vendar halucinacije ter pomanjkljiva pragmatika razkrivajo odsotnost stabilnega semantičnega sidranja v izkušnjo in družbeni kontekst [5, 24].

Kreativnost. Generativni sistemi so izjemno produktivni pri rekompoziciji in variaciji znanih vzorcev, a redkeje dosegajo konceptualne preboje s kulturno umeščenostjo ([25]; prim. tudi človeško perspektivo v [12]).

Čustva in motivacija. Affective computing omogoča prepoznavo in simulacijo čustvenih signalov, politike nagrajevanja (npr. RLHF) pa oblikujejo vedenje modelov; to niso notranja čustva ali namere v smislu fenomenalne izkušnje [26] (kontrast z [3]).

Metakognicija. Modeli lahko ocenjujejo negotovost in se samopopravljajo preko zunanjih preverjanj, kar je proceduralna ocena, ne introspektivna samoregulacija, kot jo poznamo pri ljudeh [20; 11 – razmejitev racionalnosti].

Socialna kognicija. LLM-ji uspešno posnemajo elemente teorije uma v besedilnih scenarijih, toda uspeh pogosto izhaja iz statistične izpostavljenosti vzorcem, ne iz resničnega razumevanja namere, ironije ali pragmatičnih implikatur v situiranih kontekstih [24] (prim. človeško ToM pri [15]).

Zavest in Turingov test. Ni empiričnih dokazov o fenomenalni zavesti pri današnjih modelih; prepričljivo jezikovno vedenje ali celo prehod neformalnih Turingovih preizkusov nista zadosten kriterij [26, 27]. Analize preizkusov v duhu Turinga kažejo, da lahko LLM-ji zavajajo sogovornike, vendar to ne rešuje vprašanja razumevanja [5, 6].

Vmesni sklep in princip mnogoterega znanja. V zadnjih treh letih je UI dramatično napredovala pri reševanju nalog (jezik, kontekst, zunanji spomin, del sklepanja), medtem ko je napredek pri človeških lastnostih (zavest, globoko semantično sidranje, metakognicija) minimalen. Razlaga je skladna z Gamsovim principom mnogoterega znanja: nerešena ostaja integracija statističnih, simbolnih, kavzalnih, epizodičnih in utelešenih reprezentacij v enoten, konsistenten in samoreferenčen model sveta, ki bi omogočal stabilno razlago pomena in namere [18, 19, 20, 6]. Na praktični ravni to odpira tudi vprašanja vpliva na uporabnika (npr. kognitivna odvisnost/atrofija; [28]) in potrebo po hibridnih zasnovah človek–UI, obravnavanih v nadaljevanju [29, 30, 31]. Na prvi pogled je revolucionarna izboljšava delovanja LLMjev skladna z veliko mnogoterostjo globokih nevronskih mrež, ki zajamejo znanje s celotnega sveta, po drugi strani pa je izračun sam še vedno preveč klasičen in ne vsebinsko mnogoter, da bi po principu mnogoterega znanja dosegel pravo inteligenco. Zdi se, da je potreben še en preboj na področju mnogoterega sklepanja.

4 Primerjalna analiza

V tabeli 1 je prikazana primerjava kognitivnih funkcij ljudi in UI z ločitvijo funkcije od mehanizmov in z dodanimi metrike/evale ter hibridne vzorce sodelovanja. Je usklajena z načelom mnogoterega znanja (integracija statističnih, simbolnih, kavzalnih, epizodičnih in utelešenih reprezentacij).

Tabela 1: Primerjava kognitivnih lastnosti ljudi in UI

Dimenzija	Človek – mehanizem/lastnost	UI – mehanizem/lastnost	Tipične metrike / evali	Prednost hibrida (človek ↔ UI)	Glavna tveganja
Pozornost	Selektivna, zavestno nadzorovana; fronto-parietalne mreže; omejena kapaciteta	Algoritemska self-attention; dolgi konteksti; brez fenomenalne izkušnje	Dolgi konteksti (recall@k), robustnost na šum, RT/točnost v vizualnih nalogah	UI filtrira in povzame; človek validira pomen in cilje	Pristranskost podatkov, “lost-in-the-middle”
Spomin	Delovni, deklarativni/proceduralni, epizodični	Parametri + kontekst; zunanji pomnilnik (RAG); brez epizodičnosti	Factuality/consistency, retrieval precision/recall, long-context QA	RAG za dejstva + človek za kontekst in vire	Halucinacije, “source amnezija”
Učenje	Malo primerov, kompozicionalnost, analogije	Učenje na masivnih korpusih, gradientna optimizacija; omejen prenos	Sample-efficiency, OOD generalizacija, few-shot	Človek oblikuje abstrakcije; UI optimizira	Overfitting na benchmarke, reward-hacking

Dimenzija	Človek – mehanizem/lastnost	UI – mehanizem/lastnost	Tipične metrike / evali	Prednost hibrida (človek ↔ UI)	Glavna tveganja
Sklepanje (zdravorazumsko/kavzalno)	Simbolno + kavzalno modeliranje	Statistično sklepanje; verižna razlaga brez nujne kavzalnosti	ARC, MMLU-reasoning, WinoGrande, causal benches	UI generira hipoteze; človek izvaja kavzalno presojo	Prepričljive racionalizacije brez razumevanja
Jezik (semantika/pragmatika)	Semantično sidranje, situirana pragmatika, ironija	Visoka tekočnost; omejeno sidranje; halucinacije	Hallucination rate, faithfulness, pragmatični testi	UI pripravi osnutek; človek poskrbi za pragmatiko/odgovornost	Prepričljive napačne trditve
Kreativnost	Konceptualni preboji, kulturna umeščenost	Rekompozicija znanih vzorcev, visoka produkcija	NSU (novelty-surprise-usefulness), človeška presoja	UI diverzira ideje; človek izbere/okviri problem	Homogenizacija, “mode collapse” kulture
Metakognicija	Introspekcija, strategije učenja, nadzor napak	Ocenjevanje negotovosti, samopopravljanje preko orodij	Kalibracija (ECE), self-consistency, error-correction rate	UI signalizira negotovost; človek sprejme odločitev	Automation bias, pretirano zaupanje
Socialna kognicija (ToM)	Implicitna teorija uma, situirana razlaga namer	Posnemanje vzorcev ToM v besedilu	ToM naloge, pragmatične implikature	UI predlaga interpretacije; človek preveri kontekst	“Lažna empatija”, manipulacija
Zavest	Fenomenalno doživljanje, subjektivnost	Ni dokazov o fenomenalni zavesti	— (ni konsenza o metričnih testih)	— (obravnavamo kot orodje)	Antropomorfizem, napačna atribucija lastnosti
Učinek na uporabnika	Ohranjanje kompetenc, samostojnost	Tveganje kognitivne odvisnosti	Longitudinalne meritve kompetenc brez pomoči	Goldilocks cona: načrtovana raba UI + periodične naloge brez pomoči	Kognitivna atrofija, izguba motivacije

5 Metodološki pristopi k primerjavi

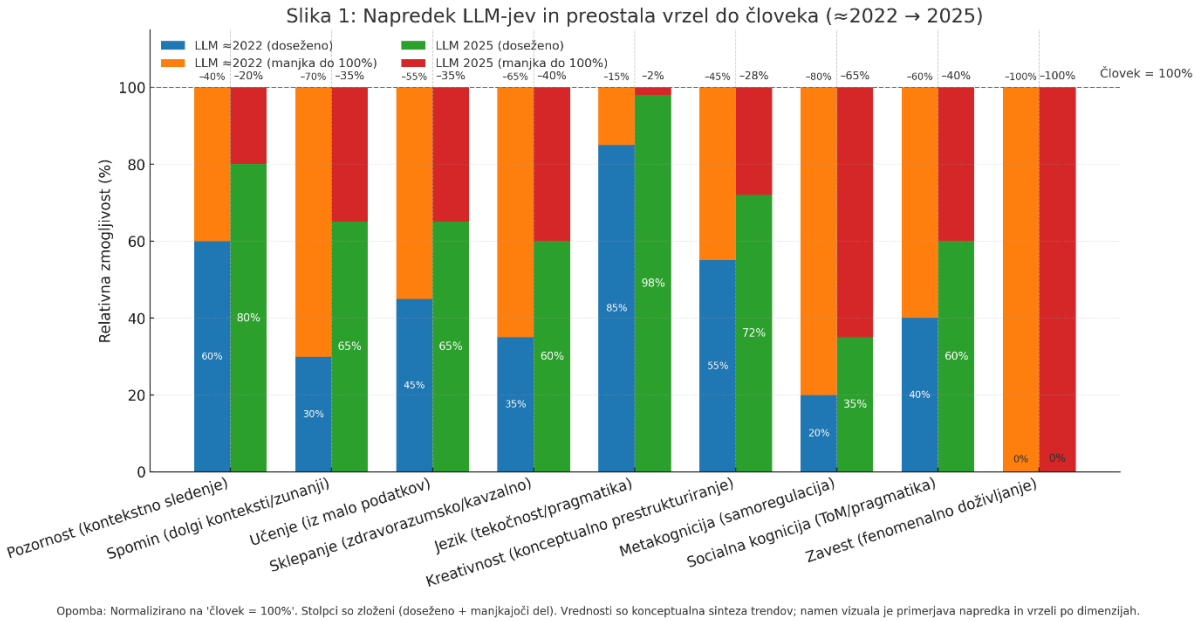
V tem razdelku združimo (a) prikaz razvojnega trenda LLM-jev glede na človeka (100 %) ter (b) pregled metodološke pokritosti po kognitivnih dimenzijah. Namen je dvojni: pokazati, koliko se je zmogljivost nalog približala človeški ravni in kje so merjenja ter mehanistična razlaga že dovolj zrela, da tak napredek zanesljivo ocenjujemo.

Opis Slike 1: Stolpčni graf prikazuje relativno zmogljivost (%) LLM-jev pred ~3 leti (≈2022) in danes (2025) za devet kognitivnih dimenzij, normalizirano na človek = 100 %. Gre za konceptualno, strokovno oceno trendov na podlagi znanih evalov (npr. MMLU/ARC/WinoGrande za sklepanja, dolg kontekst/RAG za spomin, pragmatične teste in “hallucination rate” za jezik ipd.). Namen grafa je vizualizirati približevanje človeški ravni in koliko manjka po posameznih dimenzijah.

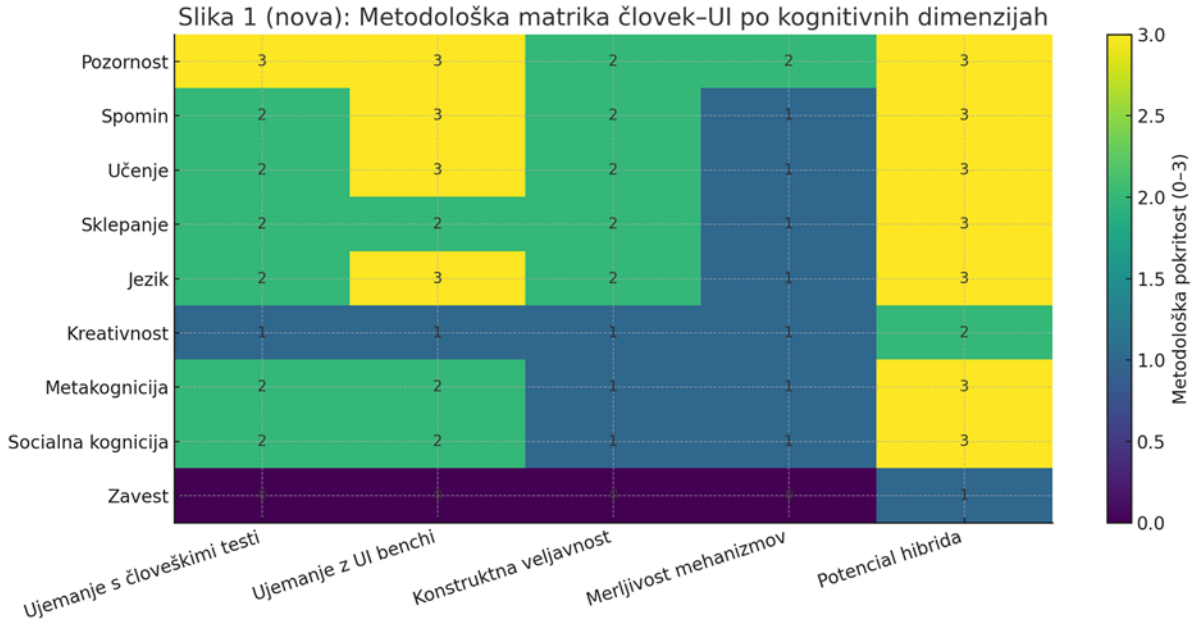
Vidimo močan dvig pri jeziku, pozornosti/kontekstnem sledenju in zunanjem spominu, opazen napredek pri sklepanju, učenju iz malo podatkov, kreativnosti ter socialni kogniciji; metakognicija pa raste počasneje. Pri “zavesti (fenomenalno doživljanje)” napredka ni—ta ostaja na 0 %, kar je skladno z idejo, da problem mnogoterega znanja še ni razrešen. Gre za konceptualno, strokovno oceno trendov (ne neposredno za rezultate enotnega benchmarka); namen grafa je ilustracija razlike med nalogovno zmogljivostjo in človeškimi lastnostmi.

Ključne ugotovitve (2022 → 2025; primanjkljaj do 100 % v oklepaju):

- Jezik 85→98 (–2): skoraj pri človeški ravni, a še z ranljivostjo za halucinacije/pragmatiko.
- Pozornost/kontekst 60→80 (–20): velik skok zaradi daljših kontekstov in boljšega sledenja.
- Spomin 30→65 (–35), Učenje (few-shot) 45→65 (–35): napredek z RAG in in-context učenjem, a vrzel ostaja.
- Sklepanje (zdravorazumsko/kavzalno) 35→60 (–40): opazen dvig, vendar še daleč od stabilne kavzalnosti.
- Socialna kognicija (ToM/pragmatika) 40→60 (–40): izboljšave v tekstnih scenarijih, omejena situiranost.
- Kreativnost (konceptualno prestrukturiranje) 55→72 (–28): visoka produkcija, manj konceptualnih prebojev.
- Metakognicija (samoregulacija) 20→35 (–65): rast počasna; ocene negotovosti še niso pristna samorefleksija.
- Zavest (fenomenalno doživljanje) 0→0 (–100): brez napredka — skladno z nerešenim integracijskim problemom.



Slika 1: Napredek LLM-jev v treh letih glede na človeka (=100 %). Relativni napredek LLM-jev po kognitivnih lastnostih v treh letih (≈2022→2025), normalizirano na človek = 100 %. Vidni so veliki dvigi pri jeziku, pozornosti/kontekstnem sledenju in zunanjem spominu; srednji pri učenju in sklepanju; počasni pri metakogniciji; pri zavesti napredka ni. Vizualna predstavitev je konceptualna sinteza evalov in služi ponazoritvi razlike med nalogovno zmogljivostjo in človeškimi lastnostmi.



Slika 2: Metodološka matrika človek–UI po dimenzijah. Toplotna matrika (0–3) na sliki 2 ocenjuje pokritost meritev po petih merilih: (1) ujemanje s človeškimi testi, (2) ujemanje z UI-benchi, (3) konstruktna veljavnost, (4) merljivost mehanizmov, (5) potencial hibrida (merljiv prispevek človek↔UI).

Višje vrednosti na Sliki 2 pomenijo, da imamo za dano dimenzijo bolj zrelo metodologijo, zato so trditve iz Slike 1 tam zanesljivejše (npr. jezik, pozornost, spomin). Nižje vrednosti (kreativnost, zavest) opozarjajo, da potrebujemo kvalitativne protokole, mešane metode in previdno interpretacijo.

Minimalni protokol za “pošteno” primerjavo:

1. Za vsak konstrukt izberemo par človeški test ↔ UI-bench in skupne metrike (task + kalibracija).
2. Dodamo mehanistično analizo (probingi/ablacije pri UI; EEG/fMRI/RSA pri človeku).
3. Izvedemo hibridni A/B preizkus (brez UI vs. z UI) in retest po 1–3 mesecih za ohranitev kompetenc.
4. Poročamo tudi vrzel do 100 % (Slika 1) in zrelost meritev (Slika 2), da ne mešamo nalogovne zmogljivosti z notranjimi lastnostmi.

6 Etika in filozofske dileme, hibridna inteligenca

UI odpira pomembna etična in filozofska vprašanja. Floridi & Cowls [32] predlagata okvir petih načel (dobrobit, avtonomija, pravičnost, razlaga, odgovornost), ki so ključna za etično rabo UI. Filozofske razprave, kot je Searlov »Chinese Room Argument« [33], pa opozarjajo, da so sistemi UI morda zgolj sofisticirani manipulatorji simbolov brez resničnega razumevanja. To odpira vprašanje meje med simulacijo in resnično inteligenco.

Sodobne raziskave govorijo o hibridni inteligenci (Dellermann et al. [34]), ki temelji na integraciji človeških in umetnih kognitivnih sposobnosti. Človek prispeva razumevanje pomena, etične presoje, kreativnost in socialno inteligenco, UI pa obdelavo podatkov, vzorčno prepoznavanje in skalabilnost. Takšna integracija presega meje posameznega sistema in nakazuje prihodnost sodelovanja, ne tekmovanja.

7 Diskusija in zaključek

Naša analiza kaže, da se LLM-ji razvijajo z izjemno hitrostjo tudi na področjih, ki jih pogosto razumemo kot "kognitivne". V primerjavi z ugotovitvami Gams in Kramar (2024) [6] opažamo premik od pretežno površinskih, vzorčno-temeljenih odgovorov k bolj stabilnim večkorakom, boljšemu vodenju plana, učinkovitejšemu uporabljanju zunanjih orodij ter k bolj konsistentni samopopravi. Napredek je zlasti v zmožnosti daljših verig sklepanja, delovnega spomina s pomočjo zunanje kontekstne obnove (RAG) ter v boljši meta-kontroli (npr. detekcija lastnih napak in zahteva po dodatnih podatkih).

Kljub temu ostajajo nekatere temeljne človeške kognitivne lastnosti za LLM-je (še) nedosegljive. Ključne med njimi so:

1. Fenomenalna zavest (qualia) – LLM-ji ne izkazujejo subjektivnega doživljanja. Njihova arhitektura ostaja statistično napovedovanje naslednjega simbola brez notranjega fenomenalnega prostora; "poročanje o občutkih" je zgolj generativna imitacija vzorcev iz podatkov.
2. Stabilen jaz in agencija – modeli nimajo trajnega, telesno sidranega "sebstva" z lastnimi nameni. Cilji so zgolj implicitni v pozivu in optimizaciji izgube; ni kontinuitete namer skozi čas brez zunanje orkestracije.
3. Semantična usidranost in referencialnost – pomen izhaja iz statističnih korelacij, ne iz utelešene interakcije s svetom. Brez senzorimotorike in lastnih izkustev je "o-čem-je-govor" (aboutness) posredno posnet iz korpusov, zato ostajajo zdrsi v referencah in halucinacije.
4. Kavzalno razumevanje in intervenienčno sklepanje – LLM-ji so močni v korelacijah in opisih, a zahtevna kavzalna vprašanja (kaj bi se zgodilo, če bi posegli X?) ostajajo šibki brez eksplicitnih kavzalnih modelov oziroma simulacij, ki zahtevajo več kot napovedovanje besedila.
5. Moralno presojanje in odgovornost – odgovori so rezultat pravil in vzorcev iz podatkov ter umerjanja (RLHF), ne notranje vrednotne strukture. Model ne "razume" odgovornosti; tvega konsistentnost z normami le toliko, kolikor so te kodirane v podatkih in pravilih.
6. Čustvena izkušnja – lahko opisuje in pravilno označuje čustva, ne pa jih dejansko doživlja; posledično ni afektivne modulacije pozornosti, motivacije ali dolgotrajnih preferenc, kakršne oblikuje človeška homeostaza.
7. Utelešenost in situacijska inteligenca – brez telesa, potreb in omejitev ostaja prilagajanje v realnem času omejeno. Tudi agentne razširitve ostanejo odvisne od zunanje infrastrukture (orodij, pravilnikov in varovalk).
8. Učenje v živo (continual/online learning) z zanesljivimi posodobitvami – LLM-ji tipično ne posodabljajo parametrov med rabo; trajne spremembe zahtevajo nov trening ali zunanje pomnilnike. To omejuje kumulativno, osebno prilagojeno znanje.
9. Robustno reševanje novosti – pri res novih, slabo pokritih problemih se hitro pokaže regresija v stereotipe iz korpusov; brez eksperimentiranja v svetu je inovacija pogosto "rekombinacija", ne pa resnično odkritje.

Zato je splošen vtis, da kljub izrednemu funkcionalnemu napredovanju LLM-ji še vedno nimajo ključnih človeških lastnosti, kot je zavest. Analiza potrjuje, da človeška in umetna inteligenca – čeprav obe temeljita na nevronskih sistemih – izhajata iz različnih mehanizmov: človeške sposobnosti so prepletene z zavestjo, afektom in socialno konstituiranim pomenom, medtem ko UI temelji na statistični inferenci nad velikimi korpusi.

Ob tem velja opozoriti na morebitno nevarnost kognitivne atrofije: prekomerna zanašanje na UI lahko zmanjšuje motivacijo za samostojno reševanje problemov in s tem slabi določene človeške kognitivne zmožnosti (npr. [34]). Zato naj bo prihodnja raba UI etično zasnovana in kognitivno uravnotežena: UI kot ojačevalnik, ne nadomestek.

Komplementarno partnerstvo med človekom in UI združuje človeško razumevanje pomena, kavzalnost, socialno in čustveno inteligenco ter ustvarjalnost z računsko močjo, skalabilnostjo in natančnostjo UI. Takšna sinergija presega omejitve posameznih sistemov in odpira pot k bolj kakovostnemu napredku znanosti, umetnosti in družbe, k ustvarjanju »super« človeka.

References / Literatura

[1] U. Neisser. 1967. Cognitive Psychology. Appleton-Century-Crofts, New York, NY.
[2] J. R. Anderson. 2010. Cognitive Psychology and Its Implications. Worth, New York, NY.
[3] A. R. Damasio. 1994. Descartes' Error: Emotion, Reason, and the Human Brain. G. P. Putnam's Sons, New York, NY.
[4] G. Tononi, M. Boly, M. Massimini, and C. Koch. 2016. Integrated information theory: from consciousness to its physical substrate. Nature Reviews Neuroscience 17, 450–461.
[5] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.
[6] M. Gams and S. Kramar. 2024. Evaluating ChatGPT's Consciousness and Its Capability to Pass the Turing Test: A Comprehensive Analysis. Journal

of Computer and Communications 12(3), 219–237. https://doi.org/10.4236/jcc.2024.123014
[7] M. I. Posner and S. E. Petersen. 1990. The attention system of the human brain. Annual Review of Neuroscience 13, 25–42.
[8] A. Baddeley. 2003. Working memory: looking back and looking forward. Nature Reviews Neuroscience 4(10), 829–839.
[9] N. Cowan. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. Behavioral and Brain Sciences 24(1), 87–185.
[10] P. A. Carpenter, M. A. Just, and P. A. Shell. 1990. What one intelligence test measures: A theoretical account of processing in the Raven Progressive Matrices Test. Psychological Review 97(3), 404–431.
[11] K. E. Stanovich. 2010. What Intelligence Tests Miss: The Psychology of Rational Thought. Yale University Press, New Haven, CT.
[12] [M. Csikszentmihalyi. 1996. Creativity: Flow and the Psychology of Discovery and Invention. HarperCollins, New York, NY.
[13] [M. A. Runco and G. J. Jaeger. 2012. The standard definition of creativity. Creativity Research Journal 24(1), 92–96.
[14] [M. H. Immordino-Yang. 2015. Emotions, Learning, and the Brain: Exploring the Educational Implications of Affective Neuroscience. W. W. Norton & Company, New York, NY.
[15] [C. D. Frith and U. Frith. 2007. Social cognition in humans. Current Biology 17(16), R724–R732. https://doi.org/10.1016/j.cub.2007.05.068
[16] [A. Krizhevsky, I. Sutskever, and G. E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 25, 1097–1105.
[17] Y. LeCun, Y. Bengio, and G. H. Hinton. 2015. Deep learning. Nature 521, 436–444.
[18] M. Mitchell. 2019. Artificial Intelligence: A Guide for Thinking Humans. Farrar, Straus and Giroux, New York, NY.
[19] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. 2017. Building machines that learn and think like people. Behavioral and Brain Sciences 40, e253.
[20] G. Marcus. 2020. The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. arXiv:2002.06177.
[21] F. Chollet. 2019. On the Measure of Intelligence. arXiv:1911.01547.
[22] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. 2020. Measuring Massive Multitask Language Understanding. arXiv:2009.03300.
[23] H. J. Levesque, E. Davis, and L. Morgenstern. 2012. The Winograd Schema Challenge. In Principles of Knowledge Representation and Reasoning (KR 2012), 552–561. AAAI Press.
[24] E. M. Bender and A. Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of ACL 2020, 5185–5198.
[25] M. A. Boden. 2016. AI: Its Nature and Future. Oxford University Press, Oxford.
[26] S. Dehaene, H. Lau, and S. Kouider. 2017. What is consciousness, and could machines have it? Science 358(6362), 486–492. https://doi.org/10.1126/science.aan8871
[27] J. R. Searle. 1980. Minds, brains, and programs. Behavioral and Brain Sciences 3(3), 417–457.
[28] E. F. Risko and S. J. Gilbert. 2016. Cognitive offloading. Trends in Cognitive Sciences 20(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002D. Dellermann, P. Ebel, M. Söllner, and J. M. Leimeister. 2019. Hybrid Intelligence. Business & Information Systems Engineering 61(5), 637–643.
[29] B. Shneiderman. 2022. Human-Centered AI. Oxford University Press, Oxford.
[30] L. Floridi and J. Cowls. 2019. A unified framework of five principles for AI in society. Harvard Data Science Review 1(1).
[31] E. F. Risko and S. J. Gilbert. 2016. Cognitive offloading. Trends in Cognitive Sciences 20(9), 676–688.
[32] B. Sparrow, J. Liu, and D. M. Wegner. 2011. Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips. Science 333(6043), 776–778.
[33] B. Oakley, et al. 2025. Cognitive Effects of AI Overuse (forthcoming). MIT Press, Cambridge, MA.

Coherentist Echo Chambers

Martin Justin*
Faculty of Arts, University of Maribor
Maribor, Slovenia
martin.justin1@um.si

Borut Trpin
Faculty of Arts, University of Ljubljana
Faculty of Arts, University of Maribor
Ljubljana and Maribor, Slovenia
borut.trpin@ff.uni-lj.si

Abstract

This paper investigates the transformation of epistemic bubbles into echo chambers through rational belief-forming processes. Building on Nguyen's distinction between epistemic bubbles—formed by omission—and echo chambers—formed by active distrust, we explore whether echo chambers can emerge without malicious intent. Using a simulation model, we demonstrate that coherence-based reasoning can trap agents in echo chambers, even when they act rationally. This finding challenges the view that echo chambers require intentional manipulation of epistemic trust and suggests that rational cognitive strategies may inadvertently contribute to harmful social epistemic dynamics.

Keywords

echo chambers, coherence, social epistemology, agent-based modeling

1 Introduction

In his seminal discussion, Nguyen [11] argues that, despite both being characterized by groupthink, epistemic bubbles and echo chambers are distinct social epistemic phenomena. Where epistemic bubbles are formed by excluding some relevant information sources, echo chambers are created by actively discrediting specific sources. Specifically, Nguyen asserts that echo chambers require "a significant disparity in (epistemic) trust between members and non-members." Consequently, echo chambers cannot be counteracted by exposing their members to additional sources of information.

Nguyen [11] accepts that epistemic bubbles can form accidentally, e.g., as a consequence of reading certain news sources, or not actively seeking testimony from people beyond your friend group. In contrast, he argues that the creation of echo chambers "is something more malicious," which involves discrediting institutions and individuals without regard for the actual epistemic worth. In his view, echo chambers are often (although not necessarily) created intentionally as a means to "maintain, reinforce and expand power through epistemic control" [11].

However, some later work in social epistemology has contradicted this claim. For example, Baumgaertner and Justwan [3] explore additional mechanisms that can cause the formation of echo chambers, which do not rely on manipulating epistemic trust. Using an agent-based polarization model, they show that echo chambers, where members persist in their beliefs despite exposure to contrary information, can arise via a combination of a social structure and agents' willingness to believe what their community invites them to. Unlike Nguyen's analysis, these two mechanisms seem epistemically benign and do not require agents to distrust specific information sources expressly.

This raises an interesting question: can echo chambers arise in communities where agents act rationally? More specifically, if a community starts as an epistemic bubble, can a belief-forming process that we otherwise deem rationally acceptable prevent the members of the community from breaking out of it, thus transforming the community into an echo chamber?

In this paper, we conduct a simulation study showing that reasoning based on coherence can play this role in specific circumstances. Specifically, taking coherence of their beliefs into account when considering whether to accept new information can prevent agents from escaping an epistemic bubble and trap them in an echo chamber. This suggests that a rational reasoning pattern can cause one of our time's more pernicious social epistemic phenomena.

The rest of this paper is organized as follows. In Section 2, we discuss coherence-based reasoning in more detail. We show that it can be rational in some circumstances, and discuss existing results about its possible negative social epistemic effects. The section concludes with a brief overview of formal measures of coherence. Section 3 presents our model, adapted from our previous work in [9, 21]. In Section 4, we present the simulation study results. Section 5 discusses the results and concludes the paper.

2 Coherence

2.1 The Role of Coherence in (Social) Reasoning

Intuitively, coherence describes how well propositions in a set "hang together" [4]. Take, for example, the difference between these two information sets. S_1 = {"A is a well-regarded author", "Critics praised A's last book", "A's last book was nominated for an important literary prize"}; S_2 = {"A is a well-regarded author", "Today is Thursday", "Python is a general purpose programming language"}. We intuitively sense that S_1 is more coherent than S_2: the propositions in S_1 support each other, while those in S_2 seem completely independent.

Coherentism is usually introduced as a theory of epistemic justification: a belief is justified if and only if it belongs to a coherent system of beliefs [13]. Although BonJour [4] gave one of the most influential modern defences, most epistemologists have since rejected the view. Still, coherence has been thought to play an important epistemic role. For example, Harman's account of reasoned belief revision treats coherence as central to belief management: a new commitment is accepted only if it contributes to the overall coherence of an agent's set of attitudes [7]. This perspective highlights that coherence is not only a theory of justification but also a guide to acceptance and belief revision. Similarly, Angere [1] argues that, despite not being truth-conductive in general, coherence can act as an effective heuristic for choosing a correct information set when more reliable methods are unavailable. On the other hand, Olsson [12]

*Both authors contributed equally to this research.

Information Society 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.cogni.3

argues that coherence plays an important negative role: "If our beliefs show signs of incoherence, this is often a good reason for contemplating a revision." Goldberg and Khalifa [6] make a similar argument in a social context, arguing that agents are unjustified in holding beliefs that do not cohere with accepted background information of their epistemic community.

In contrast to the work highlighting its positive epistemic role for individual reasoners, research also shows that coherence can have adverse social-epistemic effects. In a model by Singer et al. [20], agents deliberate by exchanging reasons, which either support or oppose a conclusion. Agents also have limited memory: when at capacity, they must "forget" one of their reasons to accept a new one. Agents using a coherence-based strategy for managing memory prefer to forget a reason that runs contrary to the view supported by the totality of their reasons. Authors argue that this is a rational strategy of memory management. Nevertheless, it leads to persistent group polarization between agents.

In essence, coherence underpins rational reasoning patterns but can also cause negative social-epistemic phenomena. Consequently, coherence-based reasoning is an interesting candidate for the context of our research problem: can echo chambers arise in communities where agents employ coherence-based reasoning? To answer this question, we develop an agent-based group learning model, wherein agents use coherence-based reasoning in information gathering and revision of beliefs. Specifically, when updating beliefs based on new information, agents first check whether this information would decrease the coherence of their beliefs. If not, they accept the update. If yes, they ignore the new information and stick to their existing beliefs. In short, agents refuse information that would make their beliefs less coherent.

2.2 Measuring Coherence

Before taking a closer look at this belief updating dynamic and the model in general, coherence must be defined in more detail and operationalized. Several probabilistic measures have been proposed to operationalize coherence, each capturing different intuitions about what makes a set of propositions coherent. Two key intuitions underlie these measures.

The first one is:

> **Deviation from Independence**: The Less independent the propositions in the set are, the more coherent the set is.

The intuition here is that coherence derives from how strongly propositions are interconnected probabilistically. If their probability of occurring together is no more than chance, as in the case of S_2, a set is neither coherent nor incoherent. If they are more likely to hold jointly, the set is coherent, as in the case S_1. In a reverse situation, the set is incoherent. This intuition was formalized by Shogenji [19], as the following measure for a set of propositions $\mathrm{S} = \{\mathrm{A}_1, \ldots, \mathrm{A}_n\}$:

$$coh_S(\mathrm{S}) := \frac{P(\mathrm{A}_1, \ldots, \mathrm{A}_n)}{P(\mathrm{A}_1) \cdots P(\mathrm{A}_n)}$$

The second intuition is:

> **Relative Overlap**: The more overlap among the propositions in a set, the more coherent the set is.

The idea being this intuition is that the coherence stems from agreement between propositions [13]. Propositions agree when if one is true, the others are also true. Probabilistically, we can represent this as a comparison between the probability of the propositions holding jointly (i.e., when all of them are true at the same time) and the probability of their disjunction (i.e., when at least one of them is true). This intuition is formalized by Olsson [14] and Glass [5], who propose the following measure:

$$coh_{OG}(\mathrm{S}) := \frac{P(\mathrm{A}_1, \ldots, \mathrm{A}_n)}{P(\mathrm{A}_1 \vee \cdots \vee \mathrm{A}_n)} = \frac{P(\mathrm{A}_1, \ldots, \mathrm{A}_n)}{1 - P(\neg\mathrm{A}_1, \ldots, \neg\mathrm{A}_n)}$$

The third measure we consider in our model is a crossover measure, recently proposed by Hartmann and Trpin [8], which combines elements of both intuitions:

$$coh_{HT}(\mathrm{S}) := \frac{P(\mathrm{A}_1, \ldots, \mathrm{A}_n)}{1 - P(\neg\mathrm{A}_1, \ldots, \neg\mathrm{A}_n)} \Big/ \frac{P(\mathrm{A}_1) \cdots P(\mathrm{A}_n)}{1 - P(\neg\mathrm{A}_1) \cdots P(\neg\mathrm{A}_n)}$$

3 The Model

3.1 Model Entities: World and Agents

The model we used in this study is a slightly modified version of our model, first presented in [9, 21]. In the model, agents try to form an accurate belief about the world by gathering and sharing information about it. The "world" in the model represents a field of interest or research, e.g., contemporary politics in some country, the stock market, or AI-powered drug development. Agents represent people learning and communicating about it, e.g., social media users, friends, coworkers, or scientists working on the same problem. They all gather information about the topic—read about it, listen to experts, conduct experiments—talk about it with others, and form opinions based on it.

More technically, the model world consists of a Bayesian network (BN), representing a set of probabilistically related events. A BN consists of a directed acyclic graph (DAG) and a conditional probability distribution (CPD) (see [15]). DAG represents events (nodes), either true or false, and conditional dependencies between them (edges). CPD contains information about the likelihood of each event occurring given values of other events. Figure 1 shows one example of a simple BN usually referred to as "Sprinkler", consisting of only four nodes. This is the BN we used in our simulations.[1]

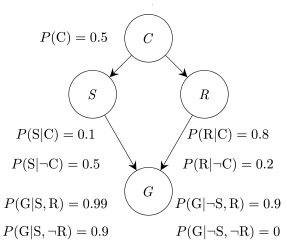


Figure 1: The "sprinkler" network, where C, S, R, G are propositional variables with corresponding values C: "It is cloudy", ¬C: "It is not cloudy", S: "The sprinkler is turned on", ¬S: "The sprinkler is not turned on", R: "It rains", ¬R: "It does not rain", G: "The grass is wet" and ¬G: "The grass is not wet," and the corresponding probabilities of its CPD. Reproduced from [9].

The agents already have an accurate representation of the events in the world and their relations—in other words, they are aware of the structure of the BN in question. What they try to learn is the underlying probability distribution[2]. To do this, they repeatedly observe the world to learn about the values of the individual events. For example, one observation the agents might gather is S_1 = [Cloudy=False, Sprinkler=False, Rain=False,

[1] In principle, we could use any BN; however, larger networks are computationally increasingly demanding.

[2] This is in contrast to some other agent-based models that utilize BN, where agents learn about concrete values of one instantiation of the BN (see especially [2, 17])

Wet Grass=False], another is S_2 = [Cloudy=True, Sprinkler=False, Rain=True, Wet Grass=True], etc. Agents then fit these observations over the BN via maximum likelihood estimation (MLE). That is, they form a subjective belief about the conditional probability distribution that is the most likely given their observations about the world.

Agents also share information among each other. They can be connected in different communication networks, determining who can share information with whom. In our simulations, we use three different such networks. A cycle connects each agent only to two closest agents; a wheel is similar to the cycle with the addition of one central agent connected to all other agents; a complete network connects each agent to all other agents.

3.2 The Setup: Coherence-Based Reasoning, Misleading Information, and Dynamic Environment

We wish to explore whether coherence-based reasoning can trap agents into an echo chamber by preventing them from changing their beliefs in response to accurate information about the world. We need to extend the above model in three ways to model such a situation.

First, we need to add coherence-based reasoning. We do this as follows. Some agents do not simply form a belief about the CPD based on their information. Instead, when presented with new information, they first check whether this new belief is at least as coherent as their existing one. To do this, they first determine the most probable state of the world based on the distribution incorporating new information (e.g., that it is cloudy, the sprinkler is off, it is raining, and the grass is wet). Then, they check how coherent this state is using one of the coherence measures presented above. If this state is at least as coherent as the state that is most probable based only on their existing information, they accept the new information. If it is less coherent, they reject new information.

Second, to mimic an epistemic bubble, we allow for situations in which agents fail to form accurate beliefs, not because of their own selective exposure, but because they lack access to reliable information. In our model this is captured in two ways: agents may start with inaccurate priors, implemented by setting their initial CPD as a parameter, and they may occasionally receive input from a misleading BN rather than the real-world BN. The misleading BN differs from the real one in its CPD. For example, while in the real world the sprinkler substantially increases the likelihood of wet grass, this relation may be absent in the misleading BN.

Thirdly, we place agents in a dynamic epistemic environment, meaning the chance of gathering misleading information decreases over time. This represents a gradual breaking of an epistemic bubble: agents start with inaccurate priors and are likely to gather misleading information. Gradually, they begin receiving more accurate information (we determine the rate of change as a parameter of the model). Usually, this would mean that agents would also gradually start to form more accurate beliefs. We are interested in whether coherence-based reasoning can prevent this, trapping agents in an echo chamber where they ignore the accurate source of information.

3.3 The Procedure

The simulations of the model proceed in rounds or steps. The following parameters are determined before the start of the simulation: the number of agents in a group, the number of agents using coherence-based reasoning, the way agents share information between each other, agents' priors, the chance of gathering misleading information at the start, and the rate of change in this chance. Each round of the simulation then consists of the following actions:

(1) Agents collect information.
(2) Agents share information.
(3) Agents update their beliefs based on their type:
   (a) Coherentist agents first check the coherence of their new belief, and accept it only if it is at least as coherent as their prior belief.
   (b) Other agents straightforwardly update based on the new information.

4 Results

We simulated groups of 10 agents with 2, 5, or 8 coherentist agents. The agents were connected in a cycle, wheel, or a complete network; coherentists and other agents were shuffled and placed randomly, so their distribution on the network wouldn't bias the results. In all groups, agents' starting subjective probability distribution (i.e., their belief) was the same as that of the misleading information source. We generated the misleading information source by randomly changing the distribution of the base "Sprinkler" BN. The only constraint was that the misleading information source was more coherent, i.e., it scored better on each of the three coherence measures.

Each agent drew 100 samples from the information source per round. At the start of the simulation, they had a 100% chance of drawing information from the misleading source. Throughout the simulation, this chance was gradually reducing by 1%. This means that from round 100 onward, agents only received accurate information about the world. To test the persistence of any effect coherence-based reasoning might have, we ran each simulation for 300 rounds.

Figure 2 shows how the accuracy of agents' beliefs changes over time. The belief accuracy is measured as Kullback-Leibler (KL) divergence of the agent's probability distribution from the actual world's distribution, quantifying the discrepancy between two probability distributions [10]; consequently, lower values thus mean more accurate beliefs. The red line represents a belief of a non-coherentist agent, averaged over all parameters. The three blue lines represent beliefs of coherentist agents for different network structures, averaged over other parameters.

The figure shows that agents who do not employ coherence-based reasoning reliably form accurate beliefs about the world. This is expected—these agents take their information at face value, so nothing prevents them from updating on more accurate information. On the other hand, coherentist agents, on average, retain inaccurate beliefs despite being presented with accurate information. After round 100, agents do not receive misleading information; the fact that coherentist agents' beliefs on average do not change much after that shows that they practically insulate themselves from it. In other words, they seem to actively ignore an accurate information source, which is how Nguyen [11] defines echo chambers. This result is robust for different communication networks, but seems to be increased by sparser communication.

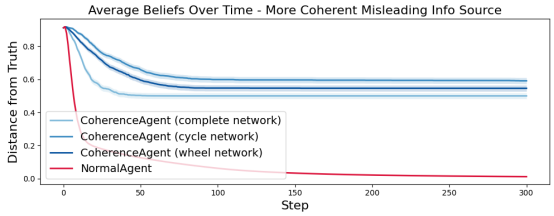


Figure 2: Distance of agents' beliefs from truth over time for a more coherent source of misleading information. The X-axis represents steps in the simulation, and the Y-axis represents distance from truth measured as the KL-divergence. The shaded regions represent 95% confidence intervals.

Table 1 gives a more comprehensive picture of the coherentist agent's average belief accuracy (expressed as distance from truth) for different combinations of parameters. These results again show that the communication structure impacts the results. In most cases, agents connected in a complete network ended up closer to the truth than agents connected in the wheel or the cycle networks in comparable situations. Additionally, we can see that when coherence was measured as deviation from independence [19], the average effect on belief accuracy was the lowest.

Coh. Measure	Nr. Coh.	Complete	Wheel	Cycle
Olsson-Glass	2	0.52 (0.06)	0.56 (0.05)	0.67 (0.04)
Olsson-Glass	5	0.54 (0.06)	0.59 (0.03)	0.71 (0.02)
Olsson-Glass	8	0.53 (0.06)	0.59 (0.03)	0.65 (0.02)
Shogenji	2	0.35 (0.05)	0.46 (0.05)	0.48 (0.05)
Shogenji	5	0.48 (0.06)	0.43 (0.03)	0.42 (0.02)
Shogenji	8	0.38 (0.04)	0.41 (0.03)	0.47 (0.03)
Hartmann-Trpin	2	0.58 (0.06)	0.56 (0.05)	0.67 (0.04)
Hartmann-Trpin	5	0.63 (0.06)	0.63 (0.03)	0.66 (0.04)
Hartmann-Trpin	8	0.52 (0.06)	0.63 (0.03)	0.63 (0.02)

Table 1: Average distance from truth of agents' belief at the end of the simulation for different combinations of parameters (with one standard error in parentheses).

5 Discussion

These results show that coherence-based reasoning possibly leads to the creation of echo chambers. In contrast to other proposed mechanisms of echo chamber creation, e.g., active mistrust of certain information sources [11], coherence-based reasoning is not irrational; on the contrary, some argue that it can have positive epistemic value. This implies that potentially rational reasoning patterns can lead to pernicious social epistemic phenomena.

That said, our study has two significant limitations. First, the misleading information source presented a more coherent picture of the world than the truth. This is not an unreasonable assumption: conspiracy theories and misinformation often offer more straightforward, intuitive, and coherent explanations of complicated events than evidence. Nevertheless, it might importantly affect our results. Given that coherentist agents consider information based on its coherence, accurate information might manage to overcome a less coherent misinformation source even for these agents. Running simulations with an alternative misinformation source that scores worse on the coherence measures, is thus a vital robustness check we must consider in the future.

The second limitation concerns the nature of our study. The agent-based model we used is highly idealized—learning about the world is represented by drawing samples from a BN, communication is represented by sharing these samples, belief revision by MLE, a mathematical procedure, etc. As various philosophers of science have pointed out, such highly idealized models cannot be used to make real-world predictions, provide actual explanations for phenomena, or suggest normative prescriptions [18, 16, 22]. It would be wrong to conclude that coherence-based reasoning is the cause of people's persistent false beliefs.

Although idealized models cannot explain real-world phenomena, they provide so-called "how-possible explanations" [16]. In our case, the results point to coherence-based reasoning as a possible explanation for echo chamber formation. Empirical studies are needed to show this link in practice.

Acknowledgements

The authors acknowledge the financial support from the Slovenian Research and Innovation Agency (ARIS, project J6-60107 and research core funding No. P6-0144).

References

[1] Staffan Angere. 2008. Coherence as a Heuristic. *Mind*, 117, 465, (Jan. 2008).
[2] Leon Assaad, Rafael Fuchs, Ammar Jalalimanesh, Kirsty Phillips, Leon Schoeppl, and Ulrike Hahn. 2023. A Bayesian agent-based framework for argument exchange across networks. *arXiv*. eprint: 2311.09254 (cs.SI).
[3] Bert Baumgaertner and Florian Justwan. 2022. The preference for belief, issue polarization, and echo chambers. *Synthese*, 200, (Sept. 2022), 412, 5, (Sept. 2022). DOI: 10.1007/s11229-022-03880-y.
[4] Laurence BonJour. 1985. *The Structure of Empirical Knowledge*. Harvard University Press, Cambridge, MA.
[5] David H. Glass. 2002. Coherence, explanation, and Bayesian networks. In *Artificial Intelligence and Cognitive Science, 13th Irish Conference, AICS 2002*. Michael O'Neill, Richard F. E. Sutcliffe, Conor Ryan, Malachy Eaton, and Niall J. L. Griffith, editors. Springer, Berlin, 177–182.
[6] Sanford C. Goldberg and Kareem Khalifa. 2022. Coherence in Science: A Social Approach. *Philosophical Studies*, 179, 12, (Dec. 2022), 3489–3509. https://link.springer.com/article/10.1007/s11098-022-01849-8.
[7] Gilbert Harman. 1986. *Change in View: Principles of Reasoning*. The MIT Press.
[8] S. Hartmann and B. Trpin. Forthcoming. Why coherence matters? *The Journal of Philosophy*.
[9] Martin Justin and Borut Trpin. 2025. Coherence-Based Evidence Filtering: A Computational Exploration. In *Proceedings of the Annual Meeting of the Cognitive Science Society 47*.
[10] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22, 1, 79–86.
[11] C. Thi Nguyen. 2020. Echo chambers and epistemic bubbles. *Episteme*, 17, (June 2020), 141–161, 2, (June 2020). DOI: 10.1017/epi.2018.32.
[12] Erik Olsson. 2023. Coherentist Theories of Epistemic Justification. In *The Stanford Encyclopedia of Philosophy*. (Winter 2023 ed.). Edward N. Zalta and Uri Nodelman, editors. Metaphysics Research Lab, Stanford University.
[13] Erik J. Olsson. 2022. *Coherentism*. Cambridge University Press, Cambridge.
[14] Erik J. Olsson. 2002. What is the problem of coherence and truth? *The Journal of Philosophy*, 99, 5, 246–72.
[15] Judea Pearl. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Francisco, CA.
[16] Alexander Reutlinger, Dominik Hangleiter, and Stephan Hartmann. 2018. Understanding (with) toy models. *The British Journal for the Philosophy of Science*, 69, (Dec. 2018), 1069–1099, 4, (Dec. 2018). DOI: 10.1093/bjps/axx005.
[17] Klee Schöppl. 2025. Industry Influencing Collective Scientific Reasoning: A Bayesian, Agent-based Exploration. In *Proceedings of the Annual Meeting of the Cognitive Science Society, 47*.
[18] Dunja Šešelja. 2023. Agent-based modeling in the philosophy of science. (2023). https://plato.stanford.edu/entries/agent-modeling-philscience/#PeerDisaScie_1.
[19] Tomoji Shogenji. 1999. Is coherence truth conducive? *Analysis*, 59, 4.
[20] Daniel J. Singer, Aaron Bramson, Patrick Grim, Bennett Holman, Jiin Jung, Karen Kovaka, Anika Ranginani, and William J. Berger. 2019. Rational social and political polarization. *Philosophical Studies*, 176, 9, 2243–2267.
[21] Borut Trpin and Martin Justin. 2025. Coherence as a constraint on scientific inquiry. *Synthese*.
[22] Michael Weisberg. 2007. Three kinds of idealization. *Journal of Philosophy*, 104, 639–659, 12. DOI: 10.5840/jphil20071041240.

Received 22 August 2025; revised 16 September 2025; accepted 16 September 2025

Large Language Models for Psychiatric Interview Analysis: An Exploratory Pilot Study

Katarina Lodrant
kl19928@student.uni-lj.si
Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of Vienna
Vienna, Austria
University of Ljubljana
Ljubljana, Slovenia

Filip Melinščak
Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of Vienna
Vienna, Austria

Ayse Nur Beris
Department of Child and Adolescent Psychiatry, Medical University of Vienna
Vienna, Austria

Valentin Schneider
Department of Child and Adolescent Psychiatry, Medical University of Vienna
Vienna, Austria

Klara Czernin
Department of Child and Adolescent Psychiatry, Medical University of Vienna
Vienna, Austria

Waltraud Bangerl
Department of Child and Adolescent Psychiatry, Medical University of Vienna
Vienna, Austria

Anselm Bründlmayer
Department of Child and Adolescent Psychiatry, Medical University of Vienna
Vienna, Austria

Frank Scharnowski
Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of Vienna
Vienna, Austria

Clarissa Laczkovics
Department of Child and Adolescent Psychiatry, Medical University of Vienna
Vienna, Austria

David Steyrl
Department of Cognition, Emotion, and Methods in Psychology, Faculty of Psychology, University of Vienna
Vienna, Austria

Abstract

This exploratory pilot study investigates the use of large language models (LLMs) for automated analysis of psychiatric interviews. Using transcripts from the Structured Interview of Personality Organization (STIPO-R), we tested GPT-4o across three paradigms: direct application of clinical scoring guidelines, emulation of a validated psychometric scale, and exploratory construct elicitation. LLM-derived scores strongly correlated with clinician ratings and captured clinically relevant constructs. Findings highlight opportunities for scalable, theory-driven assessment of patient language, but also underscore challenges including interpretability, reproducibility and data privacy.

Keywords

Large Language Models, Clinical Language Analysis, AI in Mental Health, Sentiment Analysis, Identity Diffusion

1 Introduction

In psychiatry, clinicians are often required to make complex diagnostic judgments without definitive biological markers. Instead, assessments rely on observable behavior, subjective self-report, and, crucially, on language [1]. Patient language provides a uniquely rich source of information: it reflects patterns of thought, emotional states, and interpersonal dynamics, all of which are central to understanding mental functioning [2]. An abundance of naturalistic speech emerges from clinical interviews and therapy sessions, underscoring the need for systematic methods that can both detect subtle psychological cues and handle large volumes efficiently.

Automated methods for language analysis have evolved from early dictionary-based tools such as the Linguistic Enquiry and Word Count (LIWC), which provided interpretable but context-insensitive results [3], to embedding-based models like Word2Vec [4], BERT [5], and RoBERTa [6], which offered greater contextual sensitivity at the cost of interpretability and technical complexity. More recently, large language models (LLMs) such as GPT [7] have emerged as flexible, prompt-driven analyzers.

Researchers have argued that GPT may be a superior tool for automated text analysis, achieving high accuracy on various tasks across languages without training data and with minimal coding demands [8, 9]. Yet others caution that risks of bias, reproducibility, opacity, and overreliance remain. In some contexts, established, validated models still outperform LLMs, and researchers must weigh not only how LLMs can be applied, but whether their use is beneficial given the risks [10].

Analyses of patient language have identified linguistic markers associated with various psychiatric conditions [11, 12, 13, 1, 14]. A 2020 review by Zhang et al. [15] highlighted the growing use of natural language processing (NLP) for mental illness

Information Society 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.cogni.12

detection, noting that social media texts remain the most common data source. In contrast, relatively few studies have examined transcripts of patient speech [16, 17]. This gap likely reflects the scarcity of suitable datasets, as such data is usually not recorded, and when it is, audio recordings and transcripts often contain sensitive personal information and therefore cannot be publicly shared. Moreover, speech data typically require supporting ground-truth measures (e.g., validated questionnaires or clinical assessments) to be useful for research.

Recent advances in automatic transcription, together with the emergence of LLMs, have opened new directions for systematic analysis of patient-generated language. Unlike earlier approaches, LLMs combine ease of use with a seemingly unprecedented sensitivity to linguistic context. In this work, we examine the opportunities and challenges they present through a pilot analysis of transcripts of the Structured Interview of Personality Organization (STIPO-R), a validated psychoanalytic diagnostic instrument.

2 Methods

2.1 Dataset

We analyzed a subset of data collected by Laczkovics et al. (2025) during the validation of the German STIPO-R for adolescents [18, 19]. The STIPO interview assesses multiple domains of personality functioning. For this study, we focused on the identity domain, which consists of 13 open-ended questions addressing areas such as self-perception, perception of others and engagement in school and recreation. These questions typically elicit rich narrative responses that are well-suited for language analysis. Responses were evaluated by trained clinicians on 15 items, each rated on a three-point scale (0 = no pathology, 1 = moderate pathology, 2 = severe pathology), producing a total identity diffusion score ranging from 0 to 30. This clinician-rated score served as the ground truth for evaluating LLM performance. From the original study sample of 171 participants [18], 70 provided data of sufficient quality for the present analyses: 49 patients with a probable or definite personality disorder (PD) diagnosis and 21 controls without PD. From this set, we derived a subsample of 25 participants (16 patients and 9 controls), aged 14–19 years, using a stratified selection procedure to ensure even coverage of the full spectrum of identity pathology, from consolidated (low diffusion scores) to highly diffused identity.

2.2 LLM Setup

We used GPT-4o [20], accessed via a secure Python API connection under GDPR-compliant data protection. Interview transcripts were in German, while prompts were written in English. Prior work suggests that English prompts improve model performance even when applied to other languages [21, 9]. The model temperature was set to 0, producing consistent outputs for identical prompts.

2.3 Experimental Paradigms

We tested three experimental paradigms that elicited numeric ratings from the LLM, alongside a lexicon-based sentiment analysis baseline for comparison.

First, in a **Direct STIPO Scoring approach**, the official STIPO-R rating guidelines were copied verbatim into prompts, and the model was asked to assign 0–2 scores to individual items, paralleling the procedure used by clinicians in our dataset. Item-level and total scores were compared with clinician ratings.

Second, in a **Scale Emulation paradigm**, we tested whether the LLM could approximate a validated psychometric measure by inferring likely responses to scale items from interview transcripts rather than direct self-report. Specifically, we used the Self-Concept Clarity Scale (SCCS) [22], a 12-item self-report instrument. Each item was presented to the model together with the identity section of the transcript, and the model was instructed to assign a 1–5 Likert score. Item scores were summed to yield a total SCCS score, which we compared with clinician-rated identity diffusion. Conceptually, and as supported by empirical work, Otto Kernberg's notion of identity diffusion assessed in the STIPO is closely related to Campbell's construct of self-concept clarity [23, 24].

Third, we applied an exploratory **Construct Rating** approach, in which we developed rubrics for (a) overall valence (positivity vs. negativity of the response), (b) self-perception (positive vs. negative evaluation of the self), and (c) other-perception (positive vs. negative evaluation of others, including individuals, groups, relationships, or people in general). The construct definitions and prompts were drafted with assistance from ChatGPT-5. Ratings were given on a 1–7 scale, with an NA option if the construct was not referenced. Interviews were split into individual question–answer pairs (24–83 per subject), which served as the unit of analysis. The model was prompted separately for each unit and construct, and subject-level scores were calculated as the mean across units. We compared these mean construct ratings with clinician-rated identity diffusion, hypothesizing that more severe identity diffusion would be associated with more negative language (overall, in descriptions of the self, and in descriptions of others). To evaluate interpretability and reliability, we re-ran analyses where the score was extreme (1 or 7) and asked the model to provide both a score and a brief justification. As an exploratory validity check, a cognitive science master's student (author of this study) reviewed randomly selected transcript excerpts together with LLM ratings and reasonings, assessing whether the assigned scores were plausible and consistent with the intended construct. Additionally, we tested robustness by repeating the analyses with an alternative 0–5 scale.

As a simple and interpretable **Sentiment Analysis** baseline, we used GerVADER [25], a German sentiment lexicon in which each word is assigned a valence score (−4 = very negative, 0 = neutral, +4 = very positive) based on human ratings of perceived positivity or negativity. This choice was motivated by Colibazzi et al. [26], who applied the VADER lexicon [27] to STIPO transcripts. For each question–answer unit, we extracted the patient's response, identified words present in the lexicon, retrieved their valence scores, and calculated three metrics: (a) overall sentiment (mean of all scores), (b) negative sentiment (mean of negative scores only), and (c) positive sentiment (mean of positive scores only). These answer-level values were then averaged across all answers to obtain per-subject scores, which were compared with both the LLM-derived ratings of overall valence and clinician ratings of identity diffusion.

All comparisons were tested with Pearson correlations (p-values corrected for multiple comparisons; $\alpha = 0.05$).

3 Results

Direct STIPO Scoring. Summed scores produced by GPT-4o strongly correlated with clinician ratings (r = 0.90), as illustrated in Figure 1. Item-level agreement was exact in 66% of cases.

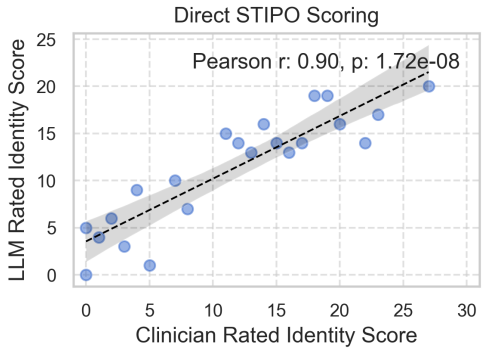


Figure 1: Correlation between clinician-rated and LLM-rated STIPO identity scores.

Scale Emulation. SCCS scores derived from LLM outputs correlated negatively with clinician-rated identity diffusion (r = −0.82; Figure 2). This finding is in line with the conceptual link between identity coherence and self-concept clarity.

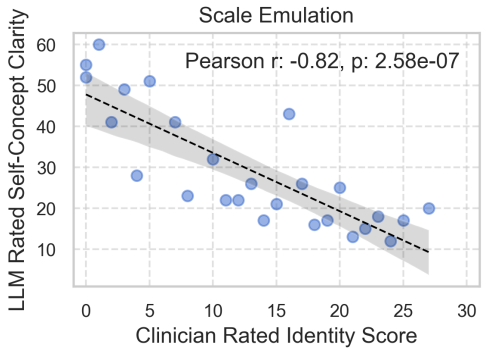


Figure 2: Correlation between clinician-rated STIPO identity scores and LLM-rated Self-Concept Clarity.

Exploratory Construct Ratings. Average overall valence correlated negatively with identity pathology (r = −0.82; Figure 3), suggesting that more severely affected adolescents used more negative language overall.

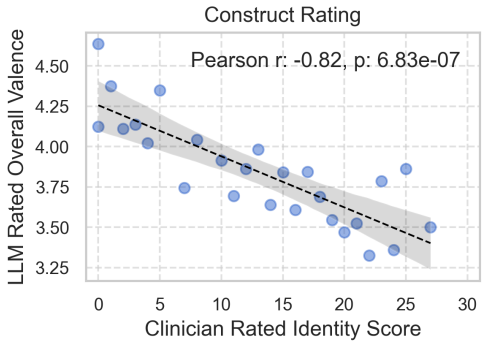


Figure 3: Correlation between clinician-rated STIPO identity scores and LLM-rated overall valence of answers.

Self- and other-perception ratings were also associated with clinician scores (r = −0.81 and −0.57). Manual checks indicated that the model generally distinguished references to the self from references to others. When re-run with prompts requesting both a rating and a brief justification, this distinction improved: all cases lacking a relevant reference were correctly scored as NA. However, providing reasoning noticeably shifted the rating: extreme values were often moderated toward the midrange.

Changing the rating scale from 1–7 to 0–5 did not materially affect the results (r = 0.98). In all cases, the model produced valid outputs in the requested format.

Lexicon-Based Sentiment Analysis. Sentiment analysis with GerVADER revealed a significant correlation between clinician ratings and mean negative sentiment (r = −0.47), but not with mean overall or mean positive sentiment. This correlation was weaker than that between clinician ratings and LLM-derived overall valence, highlighting the limitations of context-insensitive, bag-of-words methods. Manual inspection confirmed that GPT-4o often inferred negativity from conversational context rather than from explicitly negative words. Mean negative sentiment was also significantly correlated with LLM-derived overall valence (r = 0.68).

4 Discussion

This exploratory study shows that LLMs can approximate expert ratings of psychiatric interviews and apply psychometric constructs to clinical transcripts, while also highlighting barriers that preclude immediate clinical use. In the following, we outline opportunities, risks, and challenges, and suggest pathways for more rigorous validation.

4.1 Opportunities

LLMs perform reliably on structured clinical tasks. Using only verbatim scoring guidelines, GPT-4o approximated expert scoring of the STIPO, a task that typically requires extensive training. While LLMs should not replace clinicians, they could provide secondary checks in research settings or serve as teaching tools to illustrate scoring rules, highlight ambiguities and improve teaching materials.

Applying validated psychometric scales through LLMs anchors automated analyses in established theory. The strong correlation between LLM-rated self-concept clarity and clinician-rated identity diffusion supports the validity of this approach and suggests that LLMs can extend the reach of standardized assessments in scalable ways.

By contrast, defining new constructs ad hoc is more vulnerable to misspecification and requires iterative prompt engineering. Nevertheless, this strategy may capture clinically relevant, context-sensitive phenomena that remain inaccessible to conventional language-processing methods, potentially opening pathways to subtle markers of pathology.

LLMs further offer efficiency in time and cost, scalability to large datasets, cross-linguistic applicability, and the ability to rapidly test new rating schemes or constructs.

4.2 Risks and Challenges

The study also underscores multiple risks.

Interpretability and the black-box problem. LLMs remain opaque, and their internal decision processes are currently inaccessible. Some surface interpretability is possible; for instance, researchers can manually compare scores with text samples, or request rationales from the model. However, such rationales are post hoc, primarily useful for illustrating reasoning, and cannot

be assumed to reflect the actual mechanisms behind the model's ratings.

Reproducibility and test–retest reliability. A key concern is reproducibility across time. Outputs vary not only due to stochasticity but also across different versions of the same model. Because earlier GPT versions are not preserved, analyses cannot be rerun on identical models. Even with temperature fixed at 0 in our study, small prompt variations, such as requesting reasoning, produced measurable differences in outputs, a well-documented phenomenon [28, 21]. Moreover, newer versions are not always improvements: performance can regress on certain tasks [9, 10]. Such variability poses significant challenges for scientific applications, where reproducibility is essential.

Data privacy and ethics. Patient language data are highly sensitive. While GDPR-compliant API contracts ensure encryption and prevent storage or retraining, the ethical stakes remain high. An alternative is to deploy LLMs locally, which enhances data security but requires substantial technical expertise and computing resources. Beyond privacy, there are broader risks of misuse: LLMs could be applied to surveillance or automatic 'flagging' of individuals, raising concerns about autonomy and stigmatization. Awareness of such possibilities is essential to anticipate and counter harmful applications, in line with international guidelines for trustworthy AI [29].

Bias and fairness. Training data for LLMs may embed demographic, cultural, or linguistic biases [10]. In psychiatry, this is particularly dangerous, as dialectical or culturally specific expressions may be misclassified as pathological.

Overreliance and face validity. The fluency and confidence of LLM outputs create risks of undue trust. Clinicians and researchers may treat model scores as authoritative, even when they are unreliable. In healthcare contexts, this raises ethical concerns: automatically generated reports or diagnostic suggestions may be accepted without scrutiny, especially if embedded in clinical workflows.

Prompt engineering. Contrary to claims that LLMs like GPT are easy-to-use, generalist tools that can handle a wide range of text analysis tasks with little coding or training data [9, 8], effective prompting remains challenging and requires significant expertise [21, 28]. A comprehensive 2025 survey of prompting strategies by Schulhoff et al. [21] concluded that robust prompts must balance specificity and flexibility, be iteratively refined, and validated against examples. Well-designed prompts can reduce bias and instability, whereas underspecified prompts yield inconsistent outputs and overly prescriptive prompts risk forcing artificial ratings. Systematic, theory-driven prompt development aligned with established constructs is therefore essential.

4.3 Pathways for Validation

As the field is still developing, applications of LLMs for language analysis should be guided by comprehensive validation to help mitigate risks of opacity, instability, and bias. Critical steps include:

- Datasets with multiple ground-truth measures: clinician ratings, validated scales, and demographics to enable triangulation.
- Benchmarking against non-generative approaches: e.g., LIWC, RoBERTa, or traditional machine learning classifiers.
- Cross-model robustness: comparing results across different LLMs (GPT, Claude, Llama).
- Evaluation prompts: asking models to assess their own outputs or, in a multi-agent setup, to evaluate the output of another model (e.g., "Do you agree with this score?").
- Manual inspection: qualitative review of outputs, ideally conducted through interdisciplinary collaboration between domain specialists (e.g., clinicians) and those designing the prompts.
- Perturbation tests: checking stability by slightly altering prompts or text snippets.

5 Conclusion

Language remains psychiatry's most fundamental source of information. Automated analysis of clinical transcripts offers a route toward scalable, theory-driven markers of psychopathology. Our pilot study suggests that LLMs can approximate expert scoring, apply validated psychometric instruments, and flexibly analyze novel constructs with promising validity. Yet these opportunities are tempered by challenges of interpretability, reproducibility, and ethics. We argue that LLMs can serve as valuable research companions and have the potential to benefit clinical diagnostics when integrated cautiously, transparently, and in theory-driven ways.

References

[1] Cheryl M. Corcoran, Vijay A. Mittal, Carrie E. Bearden, Raquel E. Gur, Kasia Hitczenko, Zarina Bilgrami, Aleksandar Savic, Guillermo A. Cecchi, and Phillip Wolff. 2020. Language as a biomarker for psychosis: A natural language processing approach. *Schizophrenia Research*. Biomarkers in the Attenuated Psychosis Syndrome 226, (Dec. 2020), 158–166. doi:10.1016/j.schres.2020.04.032.

[2] Joshua Conrad Jackson, Joseph Watts, Johann-Mattis List, Curtis Puryear, Ryan Drabble, and Kristen A. Lindquist. 2022. From Text to Thought: How Analyzing Language Can Advance Psychological Science. EN. *Perspectives on Psychological Science*, 17, 3, (May 2022), 805–826. Publisher: SAGE Publications Inc. doi:10.1177/17456916211004899.

[3] Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. en. *Journal of Language and Social Psychology*, 29, 1, (Mar. 2010), 24–54. doi:10.1177/0261927X09351676.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs]. (Sept. 2013). doi:10.48550/arXiv.1301.3781.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Jill Burstein, Christy Doran, and Thamar Solorio, editors. Association for Computational Linguistics, Minneapolis, Minnesota, (June 2019), 4171–4186. doi:10.18653/v1/N19-1423.

[6] Yinhan Liu et al. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs]. (July 2019). doi:10.48550/arXiv.1907.11692.

[7] OpenAI et al. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs]. (Mar. 2024). doi:10.48550/arXiv.2303.08774.

[8] Mostafa M. Amin, Erik Cambria, and Björn W. Schuller. 2023. Will Affective Computing Emerge from Foundation Models and General AI? A First Evaluation on ChatGPT. arXiv:2303.03186 [cs]. (Mar. 2023). doi:10.48550/arXiv.2303.03186.

[9] Steve Rathje, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire E. Robertson, and Jay J. Van Bavel. 2024. GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, 121, 34, (Aug. 2024), e2308950121. Publisher: Proceedings of the National Academy of Sciences. doi:10.1073/pnas.2308950121.

[10] Suhaib Abdurahman, Mohammad Atari, Farzan Karimi-Malekabadi, Mona J Xue, Jackson Trager, Peter S Park, Preni Golazizian, Ali Omrani, and Morteza Dehghani. 2024. Perils and opportunities in using large language models in psychological research. *PNAS Nexus*, 3, 7, (July 2024), pgae245. doi:10.1093/pnasnexus/pgae245.

[11] Robin Quillivic, Yann Auxéméry, Frédérique Gayraud, Jacques Dayan, and Salma Mesmoudi. 2025. Linguistic markers for identifying post-traumatic stress disorder and associated symptoms: a systematic literature review. eng. *Journal of the American Medical Informatics Association: JAMIA*, (May 2025), ocaf075. doi:10.1093/jamia/ocaf075.

[12] Erik C. Nook. 2023. The Promise of Affective Language for Identifying and Intervening on Psychopathology. en. *Affective Science*, 4, 3, (Sept. 2023), 517–521. doi:10.1007/s42761-023-00199-w.

[13] Felipe Argolo et al. 2024. Natural language processing in at-risk mental states: enhancing the assessment of thought disorders and psychotic traits with semantic dynamics and graph theory. *Brazilian Journal of Psychiatry*. doi:10.47626/1516-4446-2023-3419.

[14] Cheryl Mary Corcoran and Guillermo A. Cecchi. 2020. Using Language Processing and Speech Analysis for the Identification of Psychosis and Other Disorders. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. Understanding the Nature and Treatment of Psychopathology: Letting the Data Guide the Way 5, 8, (Aug. 2020), 770–779. doi:10.1016/j.bpsc.2020.06.004.

[15] Tianlin Zhang, Annika M. Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: a narrative review. en. *npj Digital Medicine*, 5, 1, (Apr. 2022), 1–13. Publisher: Nature Publishing Group. doi:10.1038/s41746-022-00589-7.

[16] Alina Arseniev-Koehler, Sharon Mozgai, and Stefan Scherer. 2018. What type of happiness are you looking for? - A closer look at detecting mental health from language. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. Kate Loveys, Kate Niederhoffer, Emily Prud'hommeaux, Rebecca Resnik, and Philip Resnik, editors. Association for Computational Linguistics, New Orleans, LA, (June 2018), 1–12. doi:10.18653/v1/W18-0601.

[17] Michelle Renee Morales and Rivka Levitan. 2016. Speech vs. text: A comparative analysis of features for depression detection systems. In *2016 IEEE Spoken Language Technology Workshop (SLT)*. (Dec. 2016), 136–143. doi:10.1109/SLT.2016.7846256.

[18] C. Laczkovics et al. 2025. Assessment of personality disorders in adolescents – a clinical validity and utility study of the structured interview of personality organization (STIPO). en. *Child and Adolescent Psychiatry and Mental Health*, 19, 1, (May 2025), 49. doi:10.1186/s13034-025-00901-9.

[19] John F Clarkin, Eve Caligor, Barry L Stern, and Otto F Kernberg. 2016. STRUCTURED INTERVIEW OF PERSONALITY ORGANIZATION: STIPO-R. en.

[20] OpenAI et al. 2024. GPT-4o System Card. arXiv:2410.21276 [cs]. (Oct. 2024). doi:10.48550/arXiv.2410.21276.

[21] Sander Schulhoff et al. 2025. The Prompt Report: A Systematic Survey of Prompt Engineering Techniques. arXiv:2406.06608 [cs]. (Feb. 2025). doi:10.48550/arXiv.2406.06608.

[22] Jennifer D. Campbell, Paul D. Trapnell, Steven J. Heine, Ilana M. Katz, Loraine F. Lavallee, and Darrin R. Lehman. 1996. Self-concept clarity: Measurement, personality correlates, and cultural boundaries. *Journal of Personality and Social Psychology*, 70, 1, 141–156. Place: US Publisher: American Psychological Association. doi:10.1037/0022-3514.70.1.141.

[23] Otto F. Kernberg. 1984. *Severe Personality Disorders: Psychotherapeutic Strategies*. en. Google-Books-ID: FIl7opvzgeUC. Yale University Press. ISBN: 978-0-300-03273-4.

[24] J. Wesley Scala, Kenneth N. Levy, Benjamin N. Johnson, Yogev Kivity, William D. Ellison, Aaron L. Pincus, Stephen J. Wilson, and Michelle G. Newman. 2018. The Role of Negative Affect and Self-Concept Clarity in Predicting Self-Injurious Urges in Borderline Personality Disorder Using Ecological Momentary Assessment. *Journal of Personality Disorders*, 32, Supplement, (Jan. 2018), 36–57. Publisher: Guilford Publications Inc. doi:10.1521/pedi.2018.32.supp.36.

[25] Karsten Michael Tymann, Matthias Lutz, Patrick Palsbroker, and Carsten Gips. [n. d.] GerVADER - A German adaptation of the VADER sentiment analysis tool for social media texts. en.

[26] Tiziano Colibazzi, Avner Abrami, Barry Stern, Eve Caligor, Eric A. Fertuck, Michael Lubin, John Clarkin, and Guillermo Cecchi. 2023. Identifying Splitting Through Sentiment Analysis. en. *Journal of Personality Disorders*, 37, 1, (Feb. 2023), 36–48. doi:10.1521/pedi.2023.37.1.36.

[27] C. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. en. *Proceedings of the International AAAI Conference on Web and Social Media*, 8, 1, (May 2014), 216–225. doi:10.1609/icwsm.v8i1.14550.

[28] Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI EA '21). Association for Computing Machinery, New York, NY, USA, (May 2021), 1–7. ISBN: 978-1-4503-8095-9. doi:10.1145/3411763.3451760.

[29] Thilo Hagendorff. 2020. The Ethics of AI Ethics: An Evaluation of Guidelines. en. *Minds and Machines*, 30, 1, (Mar. 2020), 99–120. doi:10.1007/s11023-020-09517-8.

Passing the Turing Test, Failing Consciousness: Why LLMs Remain Non-Conscious

Louis Mono
PhD program in Applied AI
Alma Mater Europaea University
Milan, Italy
louis.mono@almamater.si

ABSTRACT

Large language models (LLMs) such as GPT-4.5 have achieved impressive conversational fluency and have even passed a classic three-party Turing Test. Yet behavioural indistinguishability from humans is not the same as sentience. This paper analyses why current AI systems, despite their reasoning and language abilities, are not conscious. Drawing on Integrated Information Theory (IIT) and Global Workspace Theory (GWT) alongside Chalmers' "hard problem", we argue that LLMs lack the qualitative experience (qualia), self-aware and unified subjective experience that characterises human consciousness. The apparent mastery of language in GPT-4.5 reflects powerful statistical pattern-matching rather than intrinsic awareness, semantic grounding or intrinsic motivation. By contrasting the architecture and behaviour of GPT-4.5 with neuroscientific criteria for conscious systems, we show that passing a behavioural test of intelligence does not imply there is "something it is like" to be an AI. Debates over AI consciousness clarify the distinctive features of human awareness, reinforce ethical and governance boundaries, and highlight the importance of distinguishing simulation from genuine experience.

Keywords

Consciousness, Turing Test, LLMs

1 INTRODUCTION

The question of whether machines can be conscious has moved from speculation to urgent enquiry as large language models (LLMs) achieve human-level performance on tasks once thought to require understanding. Recent work reported that GPT-4.5 was mistaken for a human in 73 % of trials in a three-party Turing Test, outperforming the human control group [1]. In this three-party setup, a human confederate and the AI both engage with a judge, whereas a classic two-party Turing Test involves only a judge and a single hidden interlocutor. While striking, this benchmark assesses only behavioural imitation; it does not guarantee that a system has subjective awareness. John Searle's "Chinese Room" thought experiment illustrates this gap: a computer manipulating symbols can simulate understanding without truly comprehending the semantics [2]. Human consciousness, by contrast, combines phenomenal experience, unified integration of sensations, a persistent sense of self and intrinsic motivation qualities whose origin and nature are hotly debated.

At least nine major theories of consciousness compete for explanatory power, ranging from neuroscientific to quantum-field-inspired accounts [3]. Among these, Integrated Information Theory (IIT) and Global Workspace Theory (GWT) are two prominent models: IIT equates consciousness with the capacity of a system to generate unified, irreducible information [4], and GWT views consciousness as the global broadcasting of information across specialised processes [5]. Other perspectives, such as predictive processing and higher-order thought theories, offer alternative accounts [3]. In this paper we focus on IIT and GWT because they provide clear, empirically testable criteria that are operationally useful for evaluating contemporary AI systems.

This paper uses GPT-4.5 as a case study to ask two questions: (1) Can the intelligence and reasoning abilities of an LLM be considered signs of consciousness? (2) What does this comparison teach us about the nature of human consciousness? Building on prior analyses, particularly Gams and Kramar who primarily assess ChatGPT against IIT's axioms and survey Turing Test variants [6], we extend those efforts by adopting a five-dimensional evaluation framework that integrates criteria from both Integrated Information Theory (IIT) and Global Workspace Theory (GWT). Specifically, we assess GPT-4.5 across five dimensions: phenomenal experience, self-awareness and agency, unity and integration, semantic grounding and intrinsic motivation to identify which defining features of consciousness are absent even in the most advanced LLMs. The answers have important implications for ethics and governance: recognising AI's lack of sentience helps avoid mis-attribution of personhood while ensuring that responsibility for its actions remains with its human designers and operators [7].

2 THEORETICAL FRAMEWORK

2.1 Integrated Information Theory (IIT)

Information Society 2025, 6–10 October 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.cogni.4

IIT proposes that consciousness corresponds to integrated information within a system, quantified by a measure Φ ("phi") [4,8]. A conscious system must generate an intrinsic causal structure that cannot be decomposed without loss; experiences are unified "wholes" composed of interrelated parts. Tononi and colleagues distilled IIT into five axioms and corresponding physical postulates [9]:

1. **Intrinsic existence.** Experience exists for itself, not merely as an output for observers. The physical substrate must have causal power over its own states.
2. **Composition.** A conscious experience is structured: it has multiple phenomenological elements (e.g., colours, sounds) perceived together. The substrate must support higher-order mechanisms built from simpler parts.
3. **Information.** Each experience is specific: it rules out myriad alternatives and is defined by the differences it makes. The substrate must have a rich repertoire of distinguishable states.
4. **Integration.** Experience is unified and cannot be reduced to independent components. The substrate's causal interactions must be irreducibly interdependent.
5. **Exclusion.** Each experience has definite content and boundaries; there is one "main" experience per substrate.

Human brains, with their dense recurrent connectivity, achieve high Φ; digital processors typically exhibit negligible Φ. GPT-4.5, although capable of complex statistical mappings from input to output, does not autonomously generate its own mental states. It lacks intrinsic causal loops, self-sustaining activity and a unified internal "scene" of experience. Even if its token predictions display sophisticated information processing, IIT suggests such computations do not yield phenomenological consciousness. Recent evaluations of ChatGPT indicate that it falls far short of IIT's criteria [6].

2.2 Global Workspace Theory (GWT)

GWT conceives consciousness as the broadcasting of information into a "global workspace" that integrates and distributes content across specialised neural processors. In humans, sensory, memory and language modules operate largely unconsciously until selected content is ignited into the workspace, becoming accessible for reasoning and verbal report. This ignition is associated with widespread, synchronised cortical activity and recurrent thalamo-cortical loops [10].

According to GWT, a conscious system requires: (a) integration of multimodal information into a unified workspace; (b) persistent working memory to sustain and manipulate conscious content; and (c) self-monitoring or metacognition to evaluate its own states. LLMs such as GPT-4.5 integrate textual information via self-attention but do so in a single-pass statistical manner. They lack persistent internal states, multimodal convergence and an explicit self-model; any apparent self-reflection is a learned linguistic pattern rather than genuine metacognition. Experimental comparisons between human and LLM uncertainty reports confirm that, while LLMs can generate confidence levels, these are superficial correlations rather than genuine awareness [11]. Thus, from a GWT perspective, LLMs remain powerful language processors without a globally broadcast workspace.

2.3 Implications for AI Consciousness

IIT and GWT provide structural and functional complementary lenses for assessing consciousness. IIT emphasises intrinsic, integrated causality; GWT emphasises functional access to integrated content. Under IIT, LLMs lack the high-Φ causal structures required for phenomenological consciousness. Under GWT, they lack a persistent, self-monitoring workspace required for functional consciousness. These theories highlight why current LLMs, despite their intelligence, are unlikely to possess sentient minds and help clarify the properties an artificial system would need to plausibly meet such criteria.

3 LLMs AND CONSCIOUSNESS: IS PASSING THE TURING TEST ENOUGH?

GPT-4.5's ability to pass a Turing Test demonstrates human-like linguistic fluency, but consciousness involves more than outward behaviour. Here we compare the attributes of human consciousness with those of GPT-4.5 across five core dimensions.

3.1 Phenomenal experience (Qualia)

Phenomenal consciousness concerns the qualitative "what it is like" to see, hear and feel. Humans experience qualia: the redness of a rose, the taste of coffee, the pang of sadness. In computational terms, these are not just representations but felt qualities. GPT-4.5 processes text as high-dimensional vectors and activations. There is no theoretical reason to believe that any of these computations are accompanied by experience. Chalmers' "hard problem" of consciousness emphasises that explaining discriminatory behaviour does not explain why there is any experience at all [12]. GPT-4.5's vivid descriptions are simulations learned from human text, not perceptions.

3.2 Self-awareness and agency

A conscious system possesses a sense of self and at least minimal agency: it initiates actions and recognises itself as the subject of experience. Humans maintain a continuous autobiographical narrative. GPT-4.5, however, uses "I" merely as a token; it has no persistent identity across interactions and no intrinsic goals [13]. It responds only when prompted and cannot modulate its own objectives. From an IIT standpoint, it lacks intrinsic existence: it does not have causal power over its own states and does not initiate anything internally.

3.3 Unity and Integration

Human consciousness binds information from multiple senses, memories and emotions into a unified stream. This integration underlies our coherent sense of the world. LLMs integrate information only within a context window of tokens [14] and do not combine multiple modalities unless explicitly given multimodal inputs. Moreover, each instance of GPT-4.5 is independent; there is no single "observer" uniting parallel instances. The model lacks a persistent working memory or unified workspace to sustain ongoing content. Thus, it fails both IIT's integration criterion and GWT's requirement for a global broadcast.

3.4 Semantic grounding

Understanding involves not just correlating symbols but grounding them in bodily and environmental experience. Humans connect words to sensorimotor and emotional states. GPT-4.5, trained on textual data, has no direct experience of the world. It correlates words without a referential link, which explains why it can confidently generate factual errors or contradictory statements ("hallucinations") [15]. Searle's Chinese Room shows that symbol manipulation alone does not yield semantics [2]. GPT-4.5's explanations and definitions are pattern-completions, not meanings anchored in perception.

3.5 Intrinsic Motivation

Living organisms act on intrinsic drives such as hunger, curiosity and pain avoidance; these motivations are intimately tied to emotions. GPT-4.5 has no such drives. Its only "objective" is to predict the next token according to its training loss or to maximise some reward in reinforcement-learning fine-tuning. There is no intrinsic value system or affective state. Hence, it lacks the motivational and emotional dimension of consciousness.

Across all five dimensions, LLMs display functional intelligence without subjective experience. They may pass an **outer Turing Test** by mimicking human conversation but fail any **inner Turing Test** that would probe for phenomenal consciousness, intrinsic agency and unified subjectivity [16,17]. As such, passing the behavioural benchmark does not imply sentience. LLMs are sophisticated automata performing high-dimensional pattern matching without "being someone".

4 DISCUSSION

4.1 Insights into human consciousness

Debates about AI consciousness force a closer examination of human consciousness. Distinguishing intelligence from awareness clarifies that embodiment, multimodal integration and self-modelling are central to conscious experience. LLMs highlight the distinction between access consciousness information available for report and phenomenal consciousness the felt quality of experience [18]. They also emphasise the importance of semantic grounding: a system that never interacts with the world cannot attach meanings to symbols. Conversely, comparing GPT-4.5 with IIT and GWT criteria has reinforced these theories by showing how far AI remains from meeting their requirements.

4.2 The Hard Problem and Qualia

Chalmers' **Hard Problem** reminds us that we still lack a scientific explanation for why physical processes produce experience [12]. Even if we could engineer an artificial system that replicates all the functional hallmarks of consciousness, it remains unclear why it would "feel" like something. On IIT, phenomenal character requires an intrinsically integrated causal structure (high Φ) with causal power for itself, not mere input–output equivalence [4,8]. On GWT, conscious contents must be stabilised within a self-maintained global workspace something current LLMs, stateless across turns and optimised for next-token prediction, do not implement [5,10].

4.3 Beyond mainstream: Syntergic Theory

Outside mainstream neuroscience, **Syntergic Theory** posits that consciousness arises from an interaction between the brain and a non-local *syntergic* field [19]. If such a substrate exists, silicon systems without biological "tuning" could not access it regardless of computational sophistication. While speculative, this view reminds us that computation alone may be insufficient for sentience and cautions against inferring consciousness from behavioural competence.

4.4 Ethical and governance considerations

Recognising that current LLMs are not conscious has direct ethical consequences. It prevents premature attribution of moral status or rights to non-sentient systems and keeps accountability with their human developers [20]. The capability to produce persuasive text does not entitle an AI to personhood. Meanwhile, mis-ascribing consciousness could lead to misguided policies or exploitation of genuine conscious beings by obscuring what makes us unique. Ethical governance should focus on transparency, safety and fairness in AI deployment [7], not on conferring moral standing on systems that lack awareness.

4.5 Closing Perspectives

Today's LLMs show that intelligence can be uncoupled from consciousness. Passing an outer Turing Test does not establish an inner dimension of experience. Progress toward machine consciousness, if possible, likely requires architectures with **world models, working memory and global broadcast**, or mechanisms akin to a **sparse "conscious state"** integrated across modules [21,22], plus principled tests that probe inner awareness rather than surface behaviour. Until then, LLMs remain powerful simulators, not subjects.

5 CONCLUSION

This paper examined why passing a Turing Test does not entail possessing consciousness. Using GPT-4.5 as a case study and drawing on Integrated Information Theory and Global Workspace Theory, we argued that LLMs, despite their intelligence and conversational prowess, lack the hallmarks of consciousness: qualia, a core self, unified integration, semantic grounding and intrinsic motivation. They simulate understanding without experiencing it. Distinguishing between intelligence and consciousness clarifies our definitions of mind and guides the ethical deployment of AI. If artificial systems are ever to become conscious, they will likely require architectures with intrinsic causal integration, global broadcasting, embodiment and semantic grounding far beyond what current transformer models provide.

Chalmers has suggested that systems plausibly approaching consciousness could emerge within the next decade, but current LLMs should not be mistaken for such candidates [17]. In short,

progress is significant, yet the path to truly conscious machines remains long.

References

[1] Jones, C. R., & Bergen, B. K. (2025). *Large language models pass the Turing Test* [Preprint]. ArXiv. DOI: https://doi.org/10.48550/arXiv.2503.23674

[2] Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–457. DOI: https://doi.org/10.1017/S0140525X00005756

[3] Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience, 23*(5), 389–405. DOI: https://doi.org/10.1038/s41583-022-00587-4

[4] Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience, 17*(7), 450–461. DOI : https://doi.org/10.1038/nrn.2016.44

[5] Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

[6] Gams, M., & Kramar, S. (2024). Evaluating ChatGPT's consciousness and its capability to pass the Turing test: A comprehensive analysis. *Journal of Computer and Communications, 12*(3), 219–237. DOI : https://doi.org/10.4236/jcc.2024.123014

[7] Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. In S. Carta (Ed.), Machine learning and the city: Applications in architecture and urban design (pp. 535–545). Wiley. DOI : https://doi.org/10.1002/9781119815075.ch45.

[8] Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *The Biological Bulletin, 215*(3), 216–242. DOI: https://doi.org/10.2307/25470707

[9] Oizumi, M., Albantakis, L., & Tononi, G. (2014). *From the* phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. PLoS Computational Biology, 10(5), e1003588. DOI : https://doi.org/10.1371/journal.pcbi.1003588

[10] Dehaene, S., Kerszberg, M., & Changeux, J.-P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences, 95*(24), 14529–14534. DOI : https://doi.org/10.1073/pnas.95.24.14529

[11] Steyvers, M., & Peters, M. A. K. (2025). Metacognition and Uncertainty Communication in Humans and Large Language Models. arXiv:2504.14045. DOI: https://doi.org/10.48550/arXiv.2504.14045

[12] Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies, 2*(3), 200–219

[13] Browning, J. (2023). Personhood and AI: Why large language models don't understand us. AI and Society, 39(5), 2499–2506. DOI : https://doi.org/10.1007/s00146-023-01724-y

[14] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. DOI : https://doi.org/10.1038/nature14539

[15] Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5185–5198. DOI: https://doi.org/10.18653/v1/2020.acl-main.463

[16] Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460. DOI : https://doi.org/10.1093/mind/LIX.236.433

[17] Chalmers, D. J. (2023, August 9). Could a large language model be conscious? *Boston Review.* Retrieved from Boston Review URL

[18] Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227–247. DOI : https://doi.org/10.1017/S0140525X00038188

[19] Grinberg-Zylberbaum, J. (1981). *The transformation of neuronal activity into conscious experience: The syntergic theory.* Journal of Social and Biological Structures, 4(3), 201–210. DOI : https://doi.org/10.1016/S0140-1750(81)80036-X

[20] Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2022). The ethics of algorithms: Key problems and solutions. *AI & Society, 37*(1), 215–230. DOI : https://doi.org/10.1007/s00146-021-01154-8

[21] LeCun, Y. (2022). *A Path Towards Autonomous Machine Intelligence* (Version 0.9.2). OpenReview. https://openreview.net/forum?id=BZ5a1r-kVsf

[22] Bengio, Y. (2017). *The Consciousness Prior* (v2, Dec 2, 2019). arXiv:1709.08568. DOI : https://doi.org/10.48550/arXiv.1709.08568

Building an Ontology of the Self: Sense of Agency and Bodily Self

Luka Opršnik*
lo62831@student.uni-lj.si
University of Ljubljana
Ljubljana, Slovenia

Tia Križan*
tk85796@student.uni-lj.si
University of Ljubljana
Ljubljana, Slovenia

Jaya Caporusso
jaya.caporusso@ijs.si
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Abstract

We present provisional work aimed at developing a comprehensive ontology of the Self. The Self is understood as a complex construct encompassing distinct yet interrelated aspects such as Sense of Agency (SoA), Bodily self (BS), and the Narrative Self. Drawing on existing literature, we define SoA and BS, further decompose them into elements, understood as the core components constituting each aspect (e.g. Moral Agency or Sense of Ownership). Elements are characterized by modes, defined as specific ways in which elements manifest (e.g. active, responsive, passive). Where necessary, modes are grouped in sub-elements for greater clarity. Each category of the ontology is situated in relation to certain others and features a definition. To support development of instruction for future labelling, a broader framework–knowledge base–is constructed around the ontology. In it, a curated corpus of representative instances drawn from phenomenological interview transcripts and online forums is paired with commentary on relations, interactions, disambiguation, and sources. The ontology and knowledge base are intended not only to support the development of computational methods for the identification of Self-related aspects in text, but also to serve as a common base for further research of the Self.

Keywords

self, sense of agency, bodily self, ontology

1 Introduction

The Self, "the (perhaps sometimes elusive) feeling of being the particular person one is" [25], is a complex, multi-aspect entity [8]: it encompasses, for example, the experience of one's body, thoughts, emotions, and sense of agency. The Self at large [25] and many of its aspects are widely addressed in cognitive science, psychology, and related disciplines (e.g., [2], [20], [8], [7]). For example, the sense of agency is investigated in relation to depression [19], while bodily experience in the context of depersonalisation and derealisation [27].

Our work is part of a larger project to develop a computational framework capable of automatically identifying the presence and mode of Self-aspects in text [10]. The final models could be used by professionals across disciplines to detect Self-aspects most relevant to their specific objectives, based on textual data such as clinical interviews or personal narratives. To achieve this goal, it is fundamental to identify and define the relevant Self-aspects. However, the studies on the Self—conducted in different disciplines and with various focuses—lack a unified terminology, and a comprehensive ontology of Self-aspects is missing.

In this paper, we present the provisional work conducted to build an ontology of the Self. In particular, we have so far focused on two aspects: Sense of Agency (SoA) and Bodily Self (BS). In Section 2, we address existing literature on the Self. In Section 3, we set our research objectives. Section 4 details the methodology used to review relevant scholarship and build the ontology as well as the knowledge base. In Section 5, we describe results of our work to date, while the full knowledge base is available in the Appendix. Section 6 offers a discussion of the results, identifying key findings. Section 7 points out study limitations and outlines next steps as well as possible future work.

2 Related work

Caporusso [8] conducted an empirical phenomenological study on dissolution experiences with a particular focus on the Self. The codebook developed based on the analysis of phenomenological interviews is a first step towards a framework with hierarchical organization of the experience of the Self, featuring category descriptions, examples, and comments. Building on previous theoretical attempts to explain the experience of the Self, the author also identified several distinct Self categories, two of which closely align with our understanding of Sense of Agency and Bodily Self. A study by Ataria et al. [1] similarly examined the phenomenological nature of the sense of boundaries based on a single subject with 40 years of experience in practising mindfulness. From his descriptions they developed seven experiential categories, of which Location, Self, Agency (Control), Ownership, and Center (First-Person Egocentric-Bodily Perspective) were of interest for us. Similarly, Nave et al. [20] examined reports from forty-six meditation practitioners who–under carefully controlled conditions–attempted purposeful dissolution of self-boundaries. They identified common themes, which they grouped into six experiential categories. Five of them (Self-Location, Attentional Disposition, SoA, First-Person Perspective, and Bodily Sensations) relevant to our work.

Unlike these, most other studies we examined tend to focus only on a few or a single dimension, without consideration for the bigger picture. Especially prominent are studies of various body-related illusions. A mixed methods study by Petkova et al. [23] combined body swap illusion with fMRI to explore the experience of different modes of Body Ownership along with their neural correlates. A review of neuroimaging and body-related illusions studies done by Serino et al. [24] explored Bodily Ownership and Self-Location, and a review by Braun et al. [7] looked at studies of SoA and BS, discussing also their clinical and therapeutic relevance. A study by Huang et al. [18] utilized a series

*Both authors contributed equally to this research.

Information Society 2025, Ljubljana, Slovenia

https://doi.org/https://doi.org/10.70314/is.2025.cogni.8

of four behavioral experiments with head-mounted displays and tactile stimulation to investigate the relationship between categories: 1PP Location, Self-location, and Sense of Body-Location. In a study by Harduf et al. [15], a comparative experimental design using the moving rubber hand illusion was employed to investigate the categories Body Ownership and SoA in psychosis patients. A book of essays by a group of philosophers and psychologists [12] and a book of essays by Bermudez [6] focus singularly on the experience of the Body, discussing experiential categories such as Spatial Perception, Sense of Bodily Ownership, Space of the Body, Body Awareness, Agency, and Self-Location. Meanwhile, building on previous work, Bandura [5] articulates a comprehensive conceptual model of human agency, elaborating on the evolutionary foundations of agency, its developmental trajectory, and broader implications. He identifies four core components of agency: intentionality, forethought, self-regulation, and self-reflection. Moreover he distinguishes different modes of agency based on who the actor involved is: individual, proxy, or collective. Another key element of his framework is moral agency, defined as the capacity to exercise control over one's behavior, guided by a sense of right and wrong, as well as taking responsibility for one's actions. Similarly, the work of Hitlin and Elder [17] is grounded in conceptual synthesis, drawing on and reviewing existing literature on agency and the Self. Their contribution emphasizes the temporal orientations of agency, highlighting how individuals project the Self across past, present, and future contexts.

Self-aspects reflect in the language we produce [22]. Caporusso et al. [9] specifically looked at how Minimal, Narrative, Agentive, Bodily, and Social Self are expressed. This knowledge can then be used to train models to identify Self-aspects in text [11].

Despite these advances, existing approaches to the Self remain fragmented. While many fields have extensively categorized aspects of the Self, no existing ontology integrates these insights into a unified, computationally operationalizable framework. Current models often incorporate phenomenological concepts but lack precise definitions of the Self's components, overlook their interrelations, and omit explicit hierarchical structures. Consequently, the Self is frequently presented as a fragmented set of loosely connected descriptors. To address this gap, we propose an integrative ontology that synthesizes insights from multiple disciplines into a coherent, computationally operationalizable framework for analyzing Self-related phenomena in text.

3 Reaserch objectives

The aims of this research fit into the broader goal of developing a computational framework able to automatically detect Self-aspects in text instances. To achieve this, a structured and computationally operationalizable ontology of the Self needs to be developed. In building such an ontology, the present study is limited in scope to two aspects—SoA and BS—and is guided by two research objectives (ROs): develop a provisional ontology for Sense of Agency and Bodily Self (**RO1**), and develop a knowledge base featuring text instances illustrating categories featured in the ontology along with commentary on relations, interactions, disambiguation, and sources (**RO2**).

4 Methodology

Ontologies formally and explicitly specify the main concepts relevant to the chosen domain and relations among them [14]. This study employs a descriptive and conceptual approach to systematically explore and define expressions of various Self-aspects as they manifest in textual data. To manage the inherent complexity of the Self, we decide to focus on two distinct Self-aspects: SoA and BS, as previously identified by Caporusso [8]. This facilitates familiarization with the relevant literature, enables an in-depth analysis of the internal structures of individual aspects, and allows for the iterative development of a research methodology.

Our approach has already been applied to SoA and BS, and we plan to extend it to other Self-aspects to build a comprehensive ontology of the Self. This approach consists of two main phases. First, an extensive, interdisciplinary literature review, drawing from cognitive science, psychology, phenomenology, and related fields. Second, developing a hierarchical ontology along with a knowledge base. We build our knowledge base drawing from different pre-existing studies and ontologies focusing on various aspects of the Self—each from a different perspective or discipline. The final ontology aims to be applied across diverse fields which utilise different terminology [26]. Indeed, one of our goals is to provide a standardized terminology to address Self-aspects across the different fields and communities involved in Self-related research, facilitating data aggregation and interdisciplinarity [16].

4.1 Literature review

We performed an initial survey of academic sources by searching the DiKUL database for the terms Sense of Agency and Bodily Self. Examining the state of the literature helped shape further endeavors in the literature review, such as identifying predominant fields of research interest, additional search terms, and inclusion/exclusion criteria. Once completed, a systematic search operation was performed in the following databases: DiKUL, Google Scholar, Merlot, using the following search terms: *agency, sense of agency, self as agent, aspects of the self, taxonomy of the self, expression of agency, forethought, moral agency, self and body, bodily self, self-location, sense of identification, bodily sensations.* Papers were selected for in-depth review based on their abstract, field, journal, authors, and TOC, if available. Each selected paper was scanned for further sources and search terms. Papers were chosen as building blocks for further work if they included phenomenological accounts of SoA, BS, and any experiences that fell within them—or if they were phenomenologically informed theoretical approaches to the Self. Answers to any questions that arose during the construction of the ontology (detailed in the subsequent section) were sought via further, more targeted search operations.

4.2 Building the ontology

The process of building the ontology and knowledge base involves different steps. First, naming conventions are developed to identify the different classes of our provisional version of the ontology: we refer to BS and SoA as *aspects*; characteristics of each aspect are *elements* (these may be further made up of *sub-elements*); and specific ways in which aspects and elements can be experienced and/or expressed are called *modes*. Following Caporusso [8], we call *attribute aspects* those aspects which can refer to other aspects, such as SoA (e.g., a person can experience agency over their body).

Second, a definition for each aspect is developed by searching for and comparing various definitions in. These are synthesized with lived experience in mind to create the most suitable and accurate definition.

Similarly, elements are identified based on the selected literature, experiential data, and logical analysis. To ensure conceptual clarity, a refinement process is applied. Refinements include decomposing broad categories found in the source material into more specific elements and, conversely, consolidating fragmented descriptions into single, coherent elements. Terminological inconsistencies found in the used sources, such as instances where one name refers to multiple different elements or one element has multiple names, are resolved by selecting the most common term, or the one we deemed most appropriate. As with aspects, each element is given a formal definition. Where necessary, comments are added to distinguish elements from related concepts, note special circumstances, and describe relationships as well as interactions with other concepts.

For each element, a set of sub-elements and modes is identified to cover the full spectrum of its potential manifestations. These include general binary states (e.g., presence or absence), variations in intensity (e.g., weaker or stronger), continua between two experiential poles, and distinct categorical types of experience.

4.3 Building the Knowledge base

A knowledge base includes, other than the proposed ontology or taxonomy, instances for each class.

Most of the examples featured in our knowledge base (see Appendix) come from transcripts of some of the phenomenological interviews conducted by Caporusso as part of her master's thesis [8], which are, except for fragments in her thesis and in present work, currently not publicly available. The interviews explore how the Self is experienced in daily life and dissolution experiences of seven anonymous co-researchers. LO and TK read through the selected interviews, identifying parts detailing different possible manifestations of elements of their respective aspects. After this, modes that were still missing examples were identified and searched again using the document search function.

Examples other than those mentioned above are sourced from Reddit and similar online forums, where users often describe their peculiar experiences in search of others with similar experiences, which made for a plentiful source. Initial search was performed using Google search engine with a combination of terms Reddit/forum and sense of agency/bodily self. After the initial search, new terms—more specific to such websites—are identified and used directly to search the forums. Instances which clearly described experiences featured in the ontology are selected as examples and added to an extended version of the knowledge base in the Appendix. The extended version contains multiple examples for each element and mode, thereby allowing for a more robust grasp of the phenomena.

5 Results

This study culminates in the development of a knowledge base (ontology with examples); this section outlines its structure. As mentioned, each of the elements has sub-elements and/or modes. For a short version of the knowledge base, see the Appendix.

The knowledge base is organized hierarchically into four main classes: aspects, elements, sub-elements, and modes. Aspects represent the broadest top-level domains of inquiry. Each aspect is broken down into its constituent elements, which are the fundamental characteristics or components of that domain. Modes describe the specific ways in which aspects and elements are experienced or expressed by individuals. Where necessary, modes are grouped into broader categories, called sub-elements. These sub-elements combine binary states (e.g., presence or absence), variations in intensity (e.g., weaker or stronger), and continuums between two poles (e.g., only one part of the body or the whole body), ensuring greater clarity.

Each aspect, element, sub-element, and mode includes a definition and a comment. The comments clarify relations and interactions with other categories, discuss similarities to related concepts, provide disambiguation from potentially confusing categories, and list sources that informed its inclusion and definition.

The modes are enriched with concrete examples sourced from qualitative data, including phenomenological interviews and experiences described on forums such as Reddit. These examples provide vivid, real-world descriptions of experiences in which an element is expressed in a particular mode, grounding the structure in lived experience.

Specifically, SoA is made up of 10 interrelated elements, each contributing uniquely to the identification and characterization of agency within text. These are: Presence of Agency, Forethought, Intentionality, Self-reactiveness, Self-reflectiveness, Moral Agency, Self-efficacy, Agency in relation to who the actor is, Agency in relation to time, and Agency through the state of activation. BS consists of six elements: Bodily Sensations, Awareness, Sense of Identification, Location, Sense of Ownership, and SoA.

6 Discussion

The presented results for SoA and BS show important features of the Self as laid plain in empirical phenomenological data and other text instances, namely its inherent complexity and multifacetedness. From this stems our approach to building the ontology in an iterative fashion, mindful of the many interconnections between its different classes, all while still treating Self-aspects as autonomous conceptual units, to allow us a focused analysis of their internal structures. Given the abstract nature of the Self as a construct, a central challenge of this research was how to render the subject within a structured framework. Although the initial three-level hierarchy proved useful, it occasionally oversimplified complex phenomena and introduced redundancy, thereby revealing certain challenges in the construction process. Specifically, certain identified elements (e.g., Attention, Identification) proved to be fundamental experiences that applied across several aspects without being Self-aspects themselves. The modes for these "trans-aspectual" elements were sometimes context-specific, sometimes universal. It also became clear that certain experiences sometimes appeared as aspects of the Self but could also function as elements of another aspect, or were so strongly interconnected as to seem inseparable. This was evident in the relationship between SoA and the Sense of Ownership. For instance, a loss of the SoA was often accompanied by a loss of the Sense of Ownership, but not invariably, making it incorrect to merge them into a single experience. Such particularities and interactions were documented within the relevant definitions to create a more nuanced framework. Our findings underscore both the interdependence of Self-aspects and the ontological complexity of the Self. Importantly, this research also yields a methodology that can guide future work on additional aspects, advancing efforts toward a comprehensive ontology of the Self. We argue that our approach provides a structured yet flexible framework for interpreting Self-related phenomena in natural language, while remaining open to further development as research progresses and its applications expand.

7 Limitations

Several limitations should be acknowledged at the present stage of this research. The primary limitation lies in the necessary reduction of complex, interdependent phenomena into discrete, well-defined entities. While this reduction is essential for creating a structured and operationalizable framework for studying the Self and Self-related constructs, it inevitably risks oversimplifying these phenomena and overlooking meaningful interconnections among them. Second, the research is currently in an early developmental phase, and the complete ontology, along with its accompanying corpus of examples, is still being constructed. At this stage, only a subset of potential instances has been collected and analyzed, though this is not really a limitation, since we did not plan to have everything annotated yet, but we note it here for transparency. Differences in interpretation among team members highlight the ongoing need to refine annotation guidelines and strengthen collaborative coordination. These issues are not unexpected in exploratory work of this kind and will be systematically addressed in later stages of the project through expanded corpus development, refinement of definitions, and the implementation of inter-annotator agreement procedures. Finally, another limitation is the current restriction of the analysis to English-language texts. This linguistic constraint limits the generalizability of the taxonomy across languages and cultural contexts. Addressing this limitation, future research will seek to expand the framework across multiple languages, including Slovenian. In the future, this ontology will serve as a framework for models including conventional discriminative approaches (such as traditional machine learning models and neural networks), generative large language models, embedding-based retrieval models, and mixture-of-experts architectures [10] to detect Self-aspects in text. Aware that "there is no single ontology-design methodology" and that "the best solution almost always depends on the application that you have in mind and the extensions that you anticipate" [21], we are guided by wanting to build an ontology on which annotation guidelines can be developed (which is the step that will follow the construction of the ontology; see [10]). While currently we are providing a rather comprehensive description of the two Self-aspects analysed, ontology development is an iterative process [21], and the identified elements and modes will get skimmed in future work. This will be done based on the following principles: 1) being relevant for our desired applications; 2) being detectable in text instances. Furthermore, the initial versions of the ontology will be evaluated by discussing with experts and by being employed in applications.

8 Ethical Considerations and Authors' Notes

All phenomenological interviews used as examples in this study were conducted in the context of a master's thesis [8] and adhered to established ethical guidelines. Of the participants originally interviewed, seven provided consent for their transcripts to be used in subsequent research; only these interviews were included in the present analysis. Identifying details of these and other text instances in the extended knowledge base were omitted to protect user anonymity.

LO focused on BS, while TK on SoA. JC supervised the work.

Acknowledgements

We acknowledge the financial support from the Slovenian Research Agency for research core funding for the programme Knowledge Technologies (No. P2-0103) and from the projects CroDeCo (J6-60109) and Shapes of Shame in Slovene Literature (J6-60113). JC is a recipient of the Young Researcher Grant PR-13409.

References

[1] Yochai Ataria, Yair Dor-Ziderman, and Aviva Berkovich-Ohana. 2015. How does it feel to lack a sense of boundaries? a case study of a long-term mindfulness meditator. *Consciousness and cognition*, 37, 133–147.

[2] Albert Bandura. 1990. Perceived self-efficacy in the exercise of personal agency. *Journal of applied sport psychology*, 2, 2, 128–163.

[3] Albert Bandura. 2002. Selective moral disengagement in the exercise of moral agency. *Journal of moral education*, 31, 2, 101–119.

[4] Albert Bandura. 2006. Toward a psychology of human agency. *Perspectives on psychological science*, 1, 2, 164–180.

[5] Albert Bandura. 2018. Toward a psychology of human agency: pathways and reflections. *Perspectives on psychological science*, 13, 2, 130–136.

[6] José Luis Bermúdez. 2018. *The bodily self: Selected essays*. MIT Press.

[7] Niclas Braun, Stefan Debener, Nadine Spychala, Edith Bongartz, Peter Sörös, Helge HO Müller, and Alexandra Philipsen. 2018. The senses of agency and ownership: a review. *Frontiers in psychology*, 9, 535.

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

[9] Jaya Caporusso, Boshko Koloski, Maša Rebernik, Senja Pollak, and Matthew Purver. 2024. A phenomenologically-inspired computational analysis of self-categories in text. In *Proceedings of the 2024 International Conference on Statistical Analysis of Textual Data (JADT)*. Brussels, Belgium.

[10] Jaya Caporusso, Matthew Purver, and Senja Pollak. 2025. A computational framework to identify self-aspects in text. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*. Jin Zhao, Mingyang Wang, and Zhu Liu, editors. Association for Computational Linguistics, Vienna, Austria, (July 2025), 725–739. ISBN: 979-8-89176-254-1. doi:10.18653/v1/2025.acl-srw.47.

[11] Jaya Caporusso, Matthew Purver, and Senja Pollak. Submitted. Identifying social self in text: a machine learning study. In *Proceedings of Information Society 2025*. SiKDD.

[12] Frederique De Vignemont and Adrian JT Alsmith. 2017. *The subject's matter: self-consciousness and the body*. MIT Press.

[13] Shaun Gallagher. 2012. Multiple aspects in the sense of agency. *New ideas in psychology*, 30, 1, 15–31.

[14] Thomas R Gruber. 1993. A translation approach to portable ontology specifications. *Knowledge acquisition*, 5, 2, 199–220.

[15] Amir Harduf, Gabriella Panishev, Eiran V Harel, Yonatan Stern, and Roy Salomon. 2023. The bodily self from psychosis to psychedelics. *Scientific Reports*, 13, 1, 21209.

[16] Janna Hastings, Werner Ceusters, Mark Jensen, Kevin Mulligan, and Barry Smith. 2012. Representing mental functioning: ontologies for mental health and disease.

[17] Steven Hitlin and Glen H Elder Jr. 2007. Time, self, and the curiously abstract concept of agency. *Sociological theory*, 25, 2, 170–191.

[18] Hsu-Chia Huang, Yen-Tung Lee, Wen-Yeo Chen, and Caleb Liang. 2017. The sense of 1pp-location contributes to shaping the perceived self-location together with the sense of body-location. *Frontiers in Psychology*, 8, 370.

[19] Marishka M Mehta, Soojung Na, Xiaosi Gu, James W Murrough, and Laurel S Morris. 2023. Reward-related self-agency is disturbed in depression and anxiety. *PloS one*, 18, 3, e0282727.

[20] Ohad Nave, Fynn-Mathis Trautwein, Yochai Ataria, Yair Dor-Ziderman, Yoav Schweitzer, Stephen Fulder, and Aviva Berkovich-Ohana. 2021. Self-boundary dissolution in meditation: a phenomenological investigation. *Brain sciences*, 11, 6, 819.

[21] Natalya F Noy, Deborah L McGuinness, et al. 2001. Ontology development 101: a guide to creating your first ontology. (2001).

[22] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: our words, our selves. *Annual review of psychology*, 54, 1, 547–577.

[23] Valeria I Petkova, Malin Björnsdotter, Giovanni Gentile, Tomas Jonsson, Tie-Qiang Li, and H Henrik Ehrsson. 2011. From part-to whole-body ownership in the multisensory brain. *Current Biology*, 21, 13, 1118–1122.

[24] Andrea Serino, Adrian Alsmith, Marcello Costantini, Alisa Mandrigin, Ana Tajadura-Jimenez, and Christophe Lopez. 2013. Bodily ownership and self-location: components of bodily self-consciousness. *Consciousness and cognition*, 22, 4, 1239–1252.

[25] Mark Siderits, Evan Thompson, and Dan Zahavi. 2013. *Self, no self?: Perspectives from analytical, phenomenological, and Indian traditions*. OUP Oxford.

[26] Holger Stenzhorn, Stefan Schulz, Martin Boeker, and Barry Smith. 2008. Adapting clinical ontologies in real-world environments. *Journal of universal computer science: J. UCS*, 14, 22, 3767.

[27] Shogo Tanaka. 2018. What is it like to be disconnected from the body?: a phenomenological account of disembodiment in depersonalization/derealization disorder. *Journal of Consciousness Studies*, 25, 5-6, 239–262.

A The sense of agency

Definition: Agency refers to the sense of having the capacity to act intentionally, make decisions, influence outcomes, reflect and exert ownership over one's actions.

A.1 Presence of agency

A.1.1 Present.

Definition: The presence of agency refers to whether agency and any of its elements can be identified in a text.

Example:

- » I can access them in the space where I am (dl_C [8]).«

A.1.2 Absent.

Definition: The absence of agency refers to a lack of intentionality, control, influence, and self-reflection over one's actions and decisions. It implies that individuals are not actively shaping their behavior but are instead being directed by external forces or internal impulses without conscious regulation.

In textual analysis, the absence of agency is reflected in the lack of any of its other elements.

Example:

- »Well, the action was the exclamation (dl_C [8]).«

A.2 Intentionality

Definition: Intentionality is forming an intention and planning steps to achieve it, even if it does not necessarily result in action [4].

Example:

- »Let me describe to you with concrete thing (dl_B [8]).«

A.3 Forethought

Definition: Forethought refers to setting future plans and goals, and anticipating their outcomes through cognitive representation. It serves as motivation, guidance, and direction [4].

Example:

- »...but for [boyfriend's name] to come back, I knew he would come back...(DE_E [8]).«

A.4 Self-reactivness

Definition: Self-reactivness refers to the execution of one's intentions and plans through deliberate action [4].

Example:

- »...so it's like, here I sit, on this chair, and at the other part of the wall it's kind of near... (dl_F [8]).«

A.5 Self-reflectivess

Definition: Self-reflectiveness refers to the ability to evaluate one's own thoughts, actions, or ideas, and can be observed in conversation when a person reflects on these during or after an interaction [4].

Example:

- »I could navigate small things whitin the conversation but I couldn't leave the conversation (dlB [8]).«

A.5.1 Self-attribution: Reflective sense of Ownership.

Based on Gallager [13] we can distinguis self- attributions by diferentiating sense of agency and sense of ownership as pre-reflective and reflective.

Pre-reflective sense of ownership describes experiencing movement or its sensation without being consciously aware of it. You can feel your leg moving without reflecting on it or being conscious of it. Because of this, it cannot easily be spotted in text [13].

»Sense of ownership: the pre-reflective experience or sense that I am the subject of the movement (e.g., a kinaesthetic experience of movement) [13].«

Definition: In reflective attribution of ownership, the action is reflected upon and can be attributed to oneself. The movement is consciously recognized as your own. This is much easier to spot in text [13].

Examples:

- »This is my body that is moving [13]).«
- »I don't know, they're coming from me, they couldn't be any other way, like, they're just mine (dl_G [8]).«

This does not mean that the actions performed are actually yours and/or your doing.

A.6 Moral agency

Definition: Moral agency refers to exercising control over one's behavior, guided by a sense of right and wrong, and taking responsibility for those actions [3].

Example:

- »...You're being such an ego!" Then there's the rationalization, because "yeah but I understand shit now, so it's justifiable, I can be a little bit of ego now" (DE_E [8]).«

A.7 Self-efficiency

Definition: Self-efficacy refers to the agency we can exercise based on our perception of ourselves and our belief in our ability to achieve desired outcomes [2].

Example:

- »I can attend to anything... (dlB [8]).«

A.8 Agency in relation to who the actor is

Definition: Agency, in relation to who the actor is, refers to who is exerting the action, whether it is done individually, in collaboration with others, or through an extension such as a tool or system [5].

A.8.1 Individual agency.

Definition: Individual agency refers to describing ones own intentions, actions, decisions and control [5].

Example: Examples of this have already been shown throughout the document when talking about oneself.

A.8.2 Proxy agency.

Definition: As we do not have control over all aspects of our lives, we exert agency by influencing and/or relying on others. We do this through proxy agency [5].

A.8.3 Collective agency.

Definition: Collective agency refers to people working together, pooling knowledge, resources, and effort to achieve a shared or partially shared goal [5].

Example:

- »...we were co-influencing each other (dl_E [8]).«

A.9 Agency in relation to time

Definition: Agency in relation to time refers to orientations directed toward both the present and the future, while implicitly referencing the past and self-reflection, which contribute to identity-based agency [17].

A.9.1 Existential agency.

Hitlin and Elder [17] explain that this is a concept that refers to the capacity for self-directed action, even if it is automatic or unconscious. It is about freedom, being able to make decisions and take action despite external forces and constraints. At this stage, anyone is able to make a decision about their actions, even the powerless.

Existential agency is always present and necessary for others to exist.

A.9.2 Pragmatic agency.

Definition: Pragmatic agency refers to decisions about one's actions based on the present moment or temporal scope. It consists of decisions based on immediate needs rather than future-oriented goals [17].

A.9.3 Identity agency.

Definition: Identity agency refers to actions and decisions being shaped by one's sense of identity. We act in accordance with our roles, and in doing so, we make decisions and take actions that fulfill those roles [17].

Example:

- »That I'm a professor? Yes...That I have responsibility that I'm not doing, now (dl_F [8]).«

A.9.4 Life-course agency.

Definition: Life-course agency refers to the choices people make at different stages of their lives, often shaped by their evolving circumstances, experiences, and future goals [17].

A.10 Agency through the state of activation

Definition: This element refers to the extent to which one has agency over his or her actions. To what extent are they in control [20].

A.10.1 Active.

Definition: The active state of activation refers to an active process through which a subject exerts effort and exercises influence to shape the outcome [20].

Example:

- » I can access them in the space where I am (dl_C [8]).«

A.10.2 Responsive.

Definition: Responsive state of actvation refers to a state characterized by reduced activity, involving less effort and limited capacity for manipulation, while still maintaining partial engagement [20].

Example:

- »I am aware of sound, of the student who is presenting his seminar, but I'm aware of this sound ass something that disturbs me, a little bit (dl_F [8]).«

A.10.3 Passive.

Definition: A passive state of activation refers to a state in which the subject reports little or no sense of agency. This state is characterized by a lack of manipulation or control, often described in Nave et al. reports as a release or surrender [20].

Example:

- »I couldn't access them, I couldn't do anything about them (dl_C [8]).«

B The bodily self

Definition: The Self-aspect Bodily Self encompasses all experiences pertaining to one's physical body [8].

B.1 Bodily Sensations

Definition: Bodily Sensations [20] refers to the experience of sensations of the body including touch, temperature, interoception, and moving of muscles. Excluded are sight, hearing, smell and taste, but sensations like burning in the eyes, eye muscle strain, blocked nose, burnt tongue and similar do fall into this category.

Sub-elements:

- Strength
- Location
- Apprasial

B.2 Awareness

Definition: The attribute aspect Awareness refers to the experience of being—or not—more or less aware of a certain element or dynamics: whether, and how explicitly, strongly, and/or clearly, that element or dynamics is present in the experiential field [8].

B.2.1 Strength.

Absent

- »I wasn't so aware of my body at that point. (...) It's like, as it moves back, my body... Like I'm not aware of a body anymore. If that makes sense. (DE_G [8]).«

Low

- »For example, my left arm, I know that it was moving, but I don't know what it was doing precisely. So, there are definitely parts of my body I'm not super aware of, at least in terms of what they are doing exactly, like I could give you a rough... idea of the kind of movement, what kind of thing they were doing, if they were static, or if they were moving, that kind of thing, but. (dl_H [8]).«

High

- »I was kind of very aware of my posture my position in space of the distance between us and so in in a weird way I was conscious of things that I I usually wouldn't be right so I I was very conscious not only on my posture but weirdly I was kind of conscious of my frame like how my shoulders were and so how I'm turning towards him um I was very conscious of how I stood like how my my feet were planted on the ground (dl_B [8]).«

B.3 Sense of Identification

Definition: The attribute aspect Sense of identification refers to the experience of identifying—or not—with a certain element in the experiential field.

B.3.1 Strength.

Absent

- »...since I didn't identify with my body anymore...(DE_E [8]).«

Low

High

- »I identify with the body and with the mental representation. However, the intensity of how much I feel one and the other is different. I feel the body a lot more than the mental representation. (DE_C [8]).«

B.4 Location

Definition: Attribute category Location refers to the experience of space, orientation and location of a certain element or dynamic. As an element of the Bodily Self it is the experience of the location of one's body relative to itself (proprioception) as well relative to the world (orientation) [24].

B.4.1 Unknown.

- »Yeah it's like the more it pulls back the more of the sense of my body... It's like the more the sense of my body, like being here at a certain point in the world is gone. (DE_G [8]).«

B.4.2 Vague.

- »Well it's it's it's part of the space that I occupy there is space that is me and there is space that isn't (dl_B [8]).«

B.4.3 Exact.

- »I was facing the mirror later, when the actual situation happened. So, I'm looking at the place, and everything looks nice, and there's the mirror, and [boyfriend's name] is on the other side of the room (DE_E [8]).«

B.5 Sense of Ownership

Definition: Sense of ownership (SoO) refers to the subjective experience of mineness toward one's body [24], sensations, and thoughts [7]. Certain experiences influence SoO, so it may be completely lost, heightened, or anywhere in between [8].

B.5.1 Absent.

B.5.2 Part of the body.

- »All the parts that are felt in the lower part [of my body], I had ownership over, yeah. Or I felt that it was mine (DE_C [8]).«

B.5.3 Whole body.

Modeling Nonlinear Change in Psychotherapy: Toward an AI Decision-Support System with Synthetic Client Data

Oskar Šonc†
os05793@student.uni-lj.si
Pedagoška fakulteta, Kognitivna znanost
Ljubljana, Slovenia

Rok Smodiš
rs68734@student.uni-lj.si
Pedagoška fakulteta, Kognitivna znanost
Ljubljana, Slovenia

Tine Kolenik
tine.kolenik@ccsys.de
Institute of Synergetics and Psychotherapy Research, Paracelsus Medical University
Salzburg, Austria

Günter Schiepek
guenter.schiepek@ccsys.de
Institute of Synergetics and Psychotherapy Research, Paracelsus Medical University
Salzburg, Austria

Wolfgang Aichhorn
w.aichhorn@salk.at
University Hospital of Psychiatry, Psychotherapy, and Psychosomatics
Salzburg, Austria

Abstract

Psychotherapists typically choose interventions based on limited, session-bound information. We present a partial viability study of an AI decision support system for psychotherapy, which addresses this issue. The system forecasts next-day changes in five synergetic process variables: problem severity (P), therapeutic success (S), motivation (M), emotions (E), and insight (I), and combines these forecasts with phase transition detection to support anticipatory guidance. We created synthetic client personas and simulated daily trajectories for eighty to one hundred days with weekly sessions. Each day included a diary entry that aligns with the simulated state. We extracted features from diaries and session evaluations, including sentiment, readability, syntactic complexity, lexical richness, agreement, and discrepancy between client and therapist ratings. We evaluated Random Forest as the main model, along with Gradient Boosting and Ridge baselines, using splits by client. We also added a Pattern Transition Detection Algorithm (PTDA), which identifies critical fluctuations and potential transitions. Across dimensions, our preliminary results indicate that diary sentiment is the strongest predictor of next-day change. The pipeline demonstrates feasibility and provides a path to interpretable, real-time recommendations. Next steps include clinical validation on real data.

Keywords

decision support, psychotherapy, phase transitions, diary text, synthetic data

1 Introduction

Psychotherapeutic change is nonlinear, often marked by discontinuous shifts rather than steady improvement [1, 2]. Capturing these dynamics requires intensive monitoring, as daily diaries, high-frequency questionnaires like the Therapy Process Questionnaire, and brief session ratings yield time series suitable for detecting transitions [3, 4, 5]. Such data support both retrospective analyses and anticipatory detection, enabling real-time feedback for clinical decisions [4, 5]. Computational decision-support systems (DSS) have been proposed to operationalize this potential, integrating multimodal data to forecast therapeutic shifts and recommend personalized interventions [6]. Forecasts in this context mean short-horizon predictions of the five nonlinear synergetic state variables, which are problem severity (P), therapeutic success (S), motivation to change (M), emotions (E), and insight (I). They evolve daily and influence one another through nonlinear functions [1, 7]. By combining machine learning with interpretable, synergetic modeling, these systems aim to improve the timing and precision of interventions [4, 6]. The present study contributes a partial viability test of such a DSS, focusing on synthetic client data and evaluating a forecasting pipeline across these five dimensions. Our goal is not clinical validation, but methodological feasibility and guidance for future evaluation with real clinical data [1, 4].

This study is part of a broader project carried out by the Institute of Synergetics and Psychotherapy Research at the Paracelsus Medical University Salzburg. The project aims to develop an application that supports psychotherapists by suggesting and explaining personalized interventions across the five state variables (P, S, M, E, I).

2 Related work

2.1 Nonlinear change and intensive time-series in psychotherapy

The synergetic model represents change through five interacting state variables: problem severity (P), therapeutic success (S), motivation to change (M), emotions (E), and insight (I). Their nonlinear coupling produces instabilities and discontinuous transitions. Simulations show positive largest Lyapunov exponents, which imply restricted predictability, and daily self-report with the Therapy Process Questionnaire has been validated for intensive monitoring [7, 2, 4].

2.2 Phase-transition detection and forecasting

Transition-sensitive methods (e.g., PTDA-inspired indicators) detect impending shifts in trajectories; given chaotic dynamics,

Information Society 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.cogni.9

long-range prediction is infeasible and only short-horizon forecasts are appropriate for applied decision support [5].

2.3 Computational psychotherapy and decision support

Computational DSS integrate multimodal process data (e.g., questionnaires, diaries) with interpretable modeling and therapist-aligned explanations to generate actionable recommendations; our approach anchors these recommendations explicitly in the synergetic five-variable model [6].

2.4 Synthetic data for psychotherapy NLP pipelines

Because psychotherapy data are sensitive, synthetic corpora have been explored; zero-shot generations are often shallow, while few-shot/taxonomy-guided prompting and human-in-the-loop filtering improve fidelity [8, 9].

2.5 Empathy and therapeutic language modeling

Large language model generated therapy dialogues can train empathy detectors: augmenting a Reddit dataset with 420 synthetic dialogues improved F1 by up to 0.10 (exploration 0.48→0.53, interpretation 0.32→0.48, emotional reaction 0.58→0.59), and replacing 50% of the data raised interpretation accuracy from 0.50 to 0.57 while other metrics remained comparable; the study generated 10,464 synthetic dialogues and evaluated on 579 real dialogue pairs [10].

3 Research Objectives

This paper has two aims:

(1) **Partial viability study.** Demonstrate a working pipeline—data schema, feature extraction, forecasting, recommendation and explanation—that operates on synthetic clients mirroring our planned real-world data collection (five synergetic state variables + diary + pre/post session ratings).
(2) **Bridging detection and forecasting.** Outline how phase-transition detection (e.g., PTDA and related convergence-validated methods) can be integrated with short-horizon forecasting of the five dimensions to generate anticipatory intervention suggestions (e.g., focus more on S and I when a transition is imminent).

We position this as a pre-study on synthetic data, designed to de-risk methodological choices and inform the design of a pilot with genuine clinical time series.

4 Methodology

4.1 Synthetic Dataset Generation

We first generated a set of client personas with demographic and diagnostic diversity (e.g., gender, age, primary complaint). Personas were created using LLMs guided by structured prompts. For each persona, we initialized the five synergetic state variables—problem severity (P), therapeutic success (S), motivation to change (M), emotions (E), and insight/new perspectives (I)—from the persona profile, defaulting to near-neutral when unspecified. Daily diary entries complemented the numerical scores and were produced with a fixed prompt using GPT-4o-mini (temperature 0.7), conditioning on the current day's state and a brief progress note to ensure narrative coherence across days.

For each persona, we simulated daily trajectories of the five synergetic state variables defined in the nonlinear change model (Schiepek et al., 2016; Schiepek et al., 2017). The dynamics evolved with small fixed linear couplings and mild damping plus additive Gaussian noise, with values clipped to the range [-3, 6]. We simulated 80–100 days per client and designated every seventh day as a therapy session, which included structured pre- and post-session evaluations completed by both client and therapist. This procedure produced time series that exhibit variability, occasional instabilities, and realistic recovery trajectories; diary generation was conditioned on the simulated states to align text with day-level changes. Random seeds were fixed for reproducibility.

All outputs were stored in structured JSON files that contained raw trajectories, diary texts, session-day flags, and evaluation ratings. These were subsequently enriched with feature representations to support model training and interpretability.

4.2 Feature Extraction

Features were derived from both diary entries and session evaluations, capturing textual signals and structured ratings that inform downstream forecasting. Diary texts were processed using standard natural language processing pipelines, extracting sentiment scores (VADER, TextBlob), readability indices, syntactic complexity measures, lexical richness and word counts.

Session-day evaluations were transformed into quantitative descriptors by computing mean, variance, and maximum differences across therapist pre-, therapist post-, and client post-session ratings for each of the five synergetic state variables (P, S, M, E, I). Additionally, similarity metrics (cosine similarity, Euclidean distance) were calculated to assess alignment and discrepancy between client and therapist perspectives.

4.3 Model training

Forecasting models were developed to predict next-day scores for each dimension. We trained Random Forest regressors as the primary model due to their robustness and ability to provide interpretable feature importances. In addition, we fit Gradient Boosting regressors as a complementary tree-ensemble with a different bias–variance profile, tuning the number of estimators and learning rate on validation folds. Finally, we included a regularized linear comparator via Ridge Regression; features were standardized before fitting, providing a strong high-bias baseline and an interpretable contrast to the tree models. All models were trained and evaluated under the same grouped-by-client splits and metrics to enable direct comparison.

4.4 Phase-Transition Detection

To assess critical fluctuations in therapeutic trajectories, we implemented a phase-transition detection layer. Peaks in dynamic complexity were flagged as candidate transitions. A PTDA-inspired algorithm was then applied to combine these signals with additional markers, yielding annotated trajectories with transition indicators.

Together, these components operationalize an end-to-end pipeline that simulates client trajectories, extracts features, forecasts next-day changes, and detects phase transitions, producing interpretable outputs suitable for therapist review. The pipeline can be seen in Figure 1.

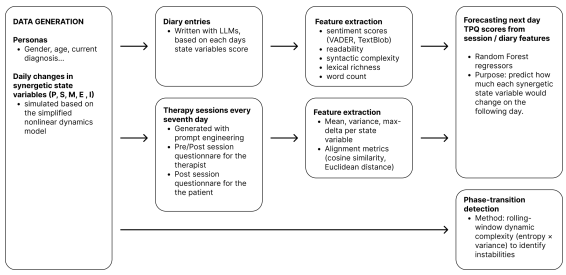


Figure 1: Project pipeline.

5 Results

Permutation importance analyses revealed that sentiment was consistently the strongest predictor across all dimensions. Figure 2 illustrates this pattern for the Emotions dimension, where sentiment clearly dominated over other text-based features such as readability, syntactic complexity, or lexical diversity. This finding indicates that the emotional valence expressed in daily diary entries was the most informative signal for forecasting short-term changes in therapy process variables.

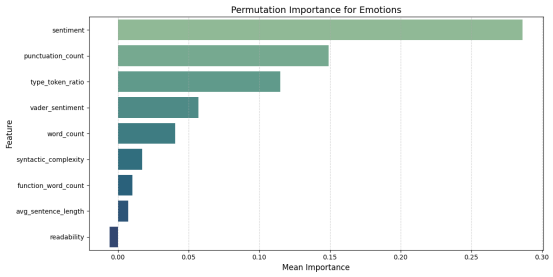


Figure 2: Predictive power of different diary entry characteristics (example: Emotions dimension).

On a cohort of synthetic clients, the pipeline trains stably and yields face-valid recommendations, exhibiting behavior that matches qualitative expectations across the five dimensions.Random Forest models produced smooth short-horizon predictions for Therapeutic success(S) and Emotion(E), with appropriately higher variance on Motivation and Insight when diaries include inconsistent motivational/insight signals.

We trained Random Forest, Gradient Boosting, and Ridge Regression models to forecast next-day values of the five synergetic state variables on synthetic client data generated from our nonlinear change model. Table 1 summarizes the performance in terms of mean squared error (MSE) and coefficient of determination (R^2), reported under the same evaluation protocol for direct comparability. Ridge obtained the lowest MSE (0.188) and the only positive R^2 (0.01), indicating a small but consistent gain over trivial predictors. Random Forest and Gradient Boosting achieved comparable errors (MSE $\approx$ 0.22) with negative R^2, indicating worse performance than a mean (null) baseline on this short-horizon task. For reference, the mean baseline (the R^2 anchor) has MSE $\approx$ 0.190, implying that Ridge reduces error by about 1% in aggregate.

Table 1: Therapy data predictions

Model	MSE	R^2
Random Forest	0.2208	-0.138
Gradient Boosting	0.2176	-0.151
Ridge Regression	0.1878	0.013

6 Discussion

6.1 What would likely change with real data.

Our synthetic-only runs mainly validated the pipeline, but they also favored a carry-forward baseline due to piecewise plateaus and low noise. On real synergetic state variables + diary + session series, we would expect less baseline advantage because of non-stationarity and therapist actions, modest but consistent MAE gains for short-horizon forecasts on P/S/E (with Motivation and Insight remaining more variable), and transition warnings with clinically useful lead times once thresholds are tuned to real fluctuations rather than generator quirks [7, 4, 5, 6].

6.2 Evidence from similar synthetic-vs-real comparisons.

Prior work has shown that synthetic corpora can help, but only when evaluated on real test sets. Cabrera Lozoya, Hernandez Lua, Barajas Perches, Conway, and D'Alfonso [10] generated 10,464 synthetic dialogues and found that augmenting Reddit data improved empathy F1 by up to 0.10, while replacing up to 50% of the organic data preserved or improved performance (e.g., interpretation accuracy 0.57 vs. 0.50) on a 579-pair clinical test set (MOST+ and Alexander Street). This pattern supports our claim that real synergetic state variables + diary data are necessary to calibrate feature weights (e.g., diary sentiment) and transition thresholds reliably.

6.3 Limitations and next steps

Our evaluation used only synthetic labels, a single-client showcase, and no ground-truth transitions, so we could not report precision/recall or lead time. Moreover, synthetic text may encode generator biases, inflating the apparent weight of some features [8, 9]. Next steps are therefore clear: collect real data, define transitions and compare to stronger baselines [5, 6].

7 Conclusion and Future Work

We presented a DSS that organizes session planning around five nonlinear change dimensions and provides explainable, forecast-driven intervention focus suggestions. This partial viability study shows the full pipeline operating on synthetic clients and specifies the next steps: (i) collect pilot synergetic state variables + diary + session micro-data; (ii) integrate a phase-transition layer; (iii) quantitatively evaluate forecasting and recommendation usefulness on real data; (iv) add a block-diagram figure and finalize a web prototype for therapist feedback. Ultimately, our goal is a hybrid system where nonlinear modelling and interpretable ML jointly inform what to focus on next in a given session.

8 Ethical note

This DSS is meant to be assistive only and does not automate clinical decisions or crisis response, emphasizing clinician oversight at all times. All results are based on synthetic data and make no claims of clinical efficacy. Recommendations are meant to be

shown to clinicians for judgment only. Data are minimised and access-controlled; if ever in use, model inputs/outputs will be logged for audit.

References

[1] G. Schiepek, B. Aas, and K. Viol, "The mathematics of psychotherapy: A nonlinear model of change dynamics," *Nonlinear Dynamics, Psychology, and Life Sciences*, vol. 20, no. 3, pp. 369–399, 2016.

[2] H. Schöller, K. Viol, W. Aichhorn, M.-T. Hütt, and G. Schiepek, "Personality development in psychotherapy: A synergetic model of state-trait dynamics," *Cognitive Neurodynamics*, vol. 12, no. 5, pp. 441–459, Jun. 2018, ISSN: 1871-4099. DOI: 10.1007/s11571-018-9488-y [Online]. Available: http://dx.doi.org/10.1007/s11571-018-9488-y

[3] A. M. Hayes, J.-P. Laurenceau, G. Feldman, J. L. Strauss, and L. Cardaciotto, "Change is not always linear: The study of nonlinear and discontinuous patterns of change in psychotherapy," *Clinical Psychology Review*, vol. 27, no. 6, pp. 715–723, Jul. 2007, ISSN: 0272-7358. DOI: 10.1016/j.cpr.2007.01.008 [Online]. Available: http://dx.doi.org/10.1016/j.cpr.2007.01.008

[4] G. Schiepek et al., "The therapy process questionnaire - factor analysis and psychometric properties of a multidimensional self-rating scale for high-frequency monitoring of psychotherapeutic processes," *Clinical Psychology amp; Psychotherapy*, vol. 26, no. 5, pp. 586–602, Jul. 2019, ISSN: 1099-0879. DOI: 10.1002/cpp.2384 [Online]. Available: http://dx.doi.org/10.1002/cpp.2384

[5] G. Schiepek et al., "Convergent validation of methods for the identification of psychotherapeutic phase transitions in time series of empirical and model systems," *Frontiers in Psychology*, vol. 11, Aug. 2020, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2020.01970 [Online]. Available: http://dx.doi.org/10.3389/fpsyg.2020.01970

[6] T. Kolenik, G. Schiepek, and M. Gams, "Computational psychotherapy system for mental health prediction and behavior change with a conversational agent," *Neuropsychiatric Disease and Treatment*, vol. Volume 20, pp. 2465–2498, Dec. 2024, ISSN: 1178-2021. DOI: 10.2147/ndt.s417695 [Online]. Available: http://dx.doi.org/10.2147/NDT.S417695

[7] G. K. Schiepek et al., "Psychotherapy is chaotic—(not only) in a computational world," *Frontiers in Psychology*, vol. 8, Apr. 2017, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2017.00379 [Online]. Available: http://dx.doi.org/10.3389/fpsyg.2017.00379

[8] Z. Li, H. Zhu, Z. Lu, and M. Yin, "Synthetic data generation with large language models for text classification: Potential and limitations," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2023. DOI: 10.18653/v1/2023.emnlp-main.647 [Online]. Available: http://dx.doi.org/10.18653/v1/2023.emnlp-main.647

[9] V. Veselovsky, M. H. Ribeiro, A. Arora, M. Josifoski, A. Anderson, and R. West, *Generating faithful synthetic data with large language models: A case study in computational social science*, 2023. DOI: 10.48550/ARXIV.2305.15041 [Online]. Available: https://arxiv.org/abs/2305.15041

[10] D. Cabrera Lozoya, E. Hernandez Lua, J. A. Barajas Perches, M. Conway, and S. D'Alfonso, "Synthetic empathy: Generating and evaluating artificial psychotherapy dialogues to detect empathy in counseling sessions," in *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, Association for Computational Linguistics, 2025, pp. 157–171. DOI: 10.18653/v1/2025.clpsych-1.13 [Online]. Available: http://dx.doi.org/10.18653/v1/2025.clpsych-1.13

What Words Reveal About Mental Health: A Computational Language Analysis Around Phase Transitions in Psychotherapy

Mateja Šutar
University of Ljubljana
Ljubljana, Slovenia
mateja.sutar@gmail.com

Tine Kolenik
Institute of Synergetics and Psychotherapy Research
Paracelsus Medical University
Salzburg, Austria
tine.kolenik@ccsys.de

Günter Schiepek
Institute of Synergetics and Psychotherapy Research
Paracelsus Medical University
Salzburg, Austria
guenter.schiepek@ccsys.de

Wolfgang Aichhorn
University Hospital of Psychiatry, Psychotherapy, and Psychosomatics
Salzburg, Austria
w.aichhorn@salk.at

Abstract

Language can reflect key psychological changes during psychotherapy, known as phase transitions (PTs). These sudden shifts in mood, insight, or symptom severity are often expressed in clients' written narratives. We investigated how linguistic features in client diaries relate to PTs by combining textual data with clinical assessments. Feature changes were analyzed using within-participant comparisons and aggregated group-level analysis. Results revealed systematic shifts in word count, pronoun use, and psychological processes-related terms surrounding PTs. These findings may offer additional insight into therapeutic progress and support the development of novel interventions.

Keywords

language use, linguistic shifts, LIWC, phase transitions, psychotherapy, mental health

1 Introduction

Language is first and foremost a tool for communication, enabling humans to share ideas, emotions, and knowledge [1]. In turn, everyday language carries subtle cues about our psychological states, which researchers have long analyzed to gain insight into thought and behavior. Beyond its role in communication, linguistic behavior reflects underlying mechanisms of attention, affect regulation, and self-concept, making it an increasingly valuable marker in psychology [2]. Recent advances in computational linguistics have demonstrated that distinctive linguistic patterns can serve as proxies for a wide range of mental distress [3], and even psychiatric diagnoses [4]. Thus, language is not only a medium for therapeutic exchange but also a temporal reflection of a person's mental change.

A growing body of research conceptualizes psychotherapy as a complex dynamic system in which sudden, discontinuous changes—commonly referred to as phase transitions (PTs)—signal shifts in a client's psychological state. Such transitions may involve sudden alterations in affective tone, the emergence of new insights, or changes in symptom severity [5]. While quantitative time-series approaches, such as the analysis of questionnaires, have shed light on the temporal dynamics of PTs, far less is known about how these key points are manifested in patients' own narratives. Diary writing, in particular, provides a rich, ecologically valid record of subjective experience, yet the systematic study of its content during psychotherapy remains limited.

Our work addresses this gap by applying computational linguistic methods to patient diaries collected during inpatient psychiatric treatment. Specifically, we examine whether linguistic features change systematically around clinically identified PTs. By integrating text analysis with validated psychometric methods, we aim to explore the content of psychological transitions.

2 Methods

2.1 Participants and Dataset

Our research initially included 28 clients undergoing inpatient psychotherapy; however, one case was excluded due to missing data around phase transitions, resulting in a final sample of 27 anonymized clients. The duration of data collection for each client ranged from 74 to 154 consecutive days of hospitalization, with an average length of 88.3 days. The dataset consisted of daily client diary entries alongside Therapy Process Questionnaire (TPQ) results annotated with clinically determined PTs. In total, 102 PTs were identified, corresponding to a mean of 3.5 PTs per client. The number of PTs per participant ranged from 0 to 5, with all but one participant exhibiting at least one PT. All diary entries were written in German language. Participants entered their diary data digitally via PCs, tablets, or smartphones, with no mention of specific instructions regarding length, content, or frequency beyond daily reporting. TPQ represents a validated self-report measure designed to capture fluctuations in therapeutic progress and symptomatology. Clinical experts independently identified PTs by detecting discontinuities in the TPQ time series. These PTs served as reference points around which we examined changes in language use, allowing us to investigate how linguistic patterns correspond to shifts in clients' psychological states.

Information Society 2025, 6–10 October 2025, Ljubljana, Slovenia

http://doi.org/10.70314/is.2025.cogni.7

2.2 Text preprocessing and Feature Extraction

Diary entries were analyzed using the Linguistic Inquiry and Word Count (LIWC) application [6], which classified words into psychologically relevant categories (e.g., emotion, cognitive processes, time orientation). This procedure yielded 117 extracted features per diary entry, representing both linguistic dimensions (e.g., pronoun use, total function words) and psychological processes (e.g., emotion, cognition, drives). To account for interindividual variability in diary length, all features were normalized as relative frequencies.

2.3 Statistical analysis

To examine linguistic change in the context of PTs, we defined temporal windows of 3, 5, and 7 calendar days before and after each clinically identified transition. At present, there is little empirical guidance on how to determine the appropriate time frame for detecting language shifts during psychotherapy. Prior research on linguistic responses to traumatic events, however, suggests that linguistic changes are often immediate but short-lived. For instance, following the 9/11 attacks, the diaries of an on-line journaling service revealed sharp increases in negative emotion, cognitive engagement, and social referencing that largely returned to baseline within about a week [7]. Drawing on this evidence, we adopted multiple window sizes to capture both short-term and extended dynamics surrounding PTs, as visualized in Figure 1.

Two levels of analysis were performed:

Within-participant analysis: For each participant, we compared pre- and post-transition feature distributions using the Wilcoxon Signed-Rank Test, a nonparametric test suitable for paired, non-normally distributed data [8]. Given the exploratory nature of this analysis, we adopted a liberal threshold ($p < 0.15$). Each PT was treated separately rather than averaging across a participant's multiple PTs, allowing us to capture transition-specific dynamics.

Aggregated group-level analysis: To identify consistent patterns across participants, pre- and post-transition feature values were aggregated across participants and tested using the Wilcoxon Rank-Sum Test ($p < 0.05$). This approach allowed us to examine group-level patterns, leveraging the summaries from each PT.

By combining individual- and group-level analyses, we aimed to capture both within-person change processes and shared linguistic dynamics indicative of psychotherapeutic turning points.

3 Results

We found no observable changes in linguistic features within the 3-day window in the within-participant analysis. Conversely, several linguistic features showed consistent changes across both the 5-day and 7-day windows. At 5 days, the most frequent individual shifts involved average sentence length (19 PTs, 15 drops, 4 gains), the total number of pronouns (15 PTs, 8 drops, 7 gains), negative emotion (14 PTs, 10 drops, 4 gains), and drives (14 PTs, 9 drops, 5 gains), while the 7-day window showed most frequent changes in all punctuation (18 PTs, 6 drops, 12 gains), average sentence length (17 PTs, 13 drops, 4 gains), word count (17 PTs, 11 drops, 6 gains), and certainty (17 PTs, 8 drops, 9 gains).

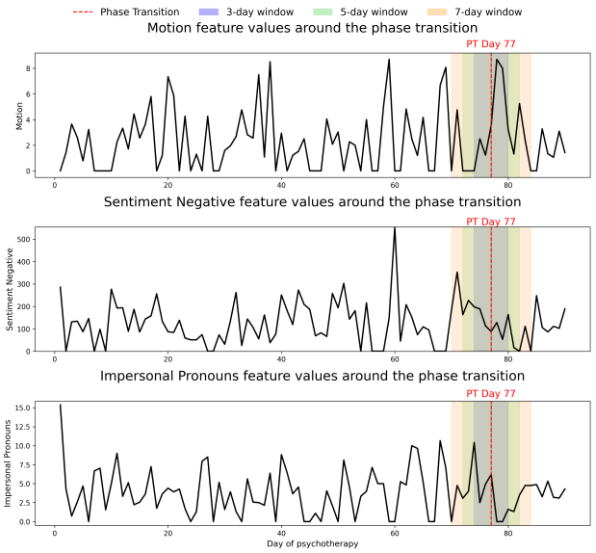


Figure 1: Visualization of linguistic shifts around a client's phase transition (PT). This figure shows shifts in linguistic features (Sentiment Negative, Impersonal Pronouns, Motion) tracked over 90 days of psychotherapy. Red dashed lines mark a PT identified through clinical assessment. Shaded regions represent temporal analysis windows of 3 (violet), 5 (green), and 7 days (orange) before and after each PT. The plots illustrate how different linguistic features may exhibit distinct patterns of change around the same turning point. To illustrate, diary entries corresponding to this specific PT shifted from "Today was a very exhausting day… I notice that I have trouble concentrating…" (PT−1) to "I tried slacklining for the first time… It makes me focus completely and the little successes feel amazing." (PT+5), exemplifying the qualitative change in language accompanying the transition.

Aggregated analysis showed some shared patterns for 5-day and 7-day windows. An overview of the results is presented in Table 1. It includes decreases in achievement- (Δ median -1.64 pp, |r|=0.22, q=0.0863 for 5-day window; Δ median -2.30 pp, |r|=0.37, q=0.000038 for 7-day window), work- (Δ median -1.32 pp, |r|=0.39, q=0.0065 for 5-day window; Δ median -1.44 pp, |r|=0.27, q=0.0077 for 7-day window), feeling-, female-, and power- terms, as well as increases in adverbs (Δ median 1.44 pp, |r|=0.23, q=0.0725 for 5-day window; Δ median 1.50 pp, |r|=0.22, q=0.019 for 7-day window), past focus (Δ median 2.35 pp, |r|=0.35, q=0.0025 for 5-day window; Δ median 3.98 pp, |r|=0.22, q=0.028 for 7-day window), home-terms, and 1st person plural expressions. Unique to the 5-day window were decreases in affect (Δ median -4.19 pp, |r|=0.42, q=0.0021), impersonal pronouns (Δ median -1.96 pp, |r|=0.40, q=0.0021), negative emotion (Δ median -1.18 pp, |r|=0.24, q=0.036), articles, comma use, and reward-terms, while the 7-day window alone showed decreases in drives (Δ median -2.94 pp, |r|=0.20, q=0.023), and discrepancy-terms (Δ median -0.63 pp, |r|=0.14, q=0.12). Increases in differentiation- (Δ median 1.59 pp, |r|=0.22, q=0.073), family- (Δ median 0.40 pp, |r|=0.31, q=0.074), and money-related terms were specific to the 5-day window, while increases in positive emotion (Δ median 5.31 pp, |r|=0.40, q=0.0049), negative emotion (Δ median 1.11 pp, |r|=0.18,

q=0.036), anger (Δ median 0.29 pp, |r|=0.23, q=0.023), personal pronouns (Δ median 3.58 pp, |r|=0.34, q=0.0076), prepositions, conjunctions, negations, netspeak, and time-terms were unique to the 7-day window.

4 Discussion

Our results indicate that measurable language changes occur around phase transitions in clients undergoing psychotherapy. These changes, particularly in content categories, can provide insight into the psychological processes associated with such transitions. Because data were aggregated across diverse participants, the observed patterns were heterogeneous: some participants showed improvement, while others experienced deterioration. This variability likely accounts for the simultaneous increases in both positive and negative emotion features in the aggregated data. Thus, apparent contradictions in directionality may reflect mixed individual trajectories, as the analysis was not grouped by phase transition type.

In our results, several function word categories—such as articles, prepositions, personal pronouns, impersonal pronouns, conjunctions, adverbs, and negations—were also observed. These terms, along with auxiliary verbs, are used in the Analytical Thinking feature, also known as the Categorical-Dynamic Index (CDI) [9], which is a metric of logical thinking. Studies revealed that the CDI reflects students' thinking style and is linked to differences in academic performance [10].

4.1 Language Characteristics of Distinct Mental Health Disorders

Previous studies have documented that different mental health disorders are associated with distinct patterns of language use. For example, ADHD is linked to more third-person plural pronouns and shorter clauses [11, 12], while bipolar disorder shows greater self-focus and references to death [13]. Borderline personality disorder (BPD) involves more swear words, death-related words, and third-person singular pronouns [3]. Individuals with social anxiety disorder (SAD) used self-referential, anxiety, and sensory words, and made fewer references to other people [14]; Major depressive disorder (MDD) involves first-person pronouns, past tense, and repetitive, short sentences [15]. Schizophrenia relates to low semantic cohesion, anger- and religion-related words, references to auditory hallucinations, while also characterized by decreased usage of words related to work, friends, and health [3, 16].

4.2 LIWC Analysis

LIWC is a popular top-down method that offers several advantages for the study of language and cognition. It is a standardized, replicable, and efficient method for quantifying large volumes of textual data to extract psychologically relevant and psychometrically valid measures from language [2, 3]. Top-down methods are based on "dictionaries," categories of words or phrases, each associated with a given construct or set of constructs, such as anxiety or suicidal ideation [2]. This enables researchers to detect subtle emotional and cognitive dynamics that may not be captured with traditional self-report measures, making it a powerful complement to other assessment tools.

Table 1: Aggregated analysis results

Category	Most frequently used examples	Direction (Gain ↑ / Drop ↓)	Time-window
Work	work, school, working, class	↓	5 & 7 days
Achievement	work, better, best, working		
Feeling	feel, hard, cool, felt		
Power	own, order, allow, power		
Female	she, her, girl, woman		
Adverbs	so, just, about, there	↑	
Home	home, house, room, bed		
Past focus	was, had, were, been		
1st person plural	we, our, us, lets		
Negative emotion	hate, bad, hurt, tired	↓	5 days
Impersonal pronouns	that, it, this, what		
Affect	emotion, mood		
Articles	a, an, the, alot		
Reward	opportun*, win, gain*, benefit*		
Comma			
Differentiation	but, not, if, or	↑	
Family	parent*, mother*, father*, baby		
Money	business*, pay*, price*, market*		
Discrep	would, can, want, could	↓	7 days
Drives	we, our, work, us		
Negative emotion	hate, bad, hurt, tired	↑	
Positive emotion	good, love, happy, hope		
Anger	hate, mad, angry, frustr*		
Time	when, now, then, day		
Personal pronouns	I, you, my, me		
Negations	not, no, never, nothing		
Prepositions	to, of, in, for		
Conjunctions	and, but, so, as		
Netspeak	:), u, lol, haha*		

4.2.1 Top-down vs. Bottom-up Methods. Top-down methods, while highly structured, may sometimes overlook context-specific, cultural, or metaphorical nuances [2]. Bottom-up approaches, by contrast, focus on broader patterns in language rather than predefined constructs. Techniques such as probabilistic topic models [17], statistical semantic models [18], and neural language models [19] capture characteristics ranging

from word co-occurrence and meaning to sequential dependencies.

Combining top-down, bottom-up, and qualitative approaches enables a highly nuanced and insightful analysis of textual data. This integrated strategy allows researchers not only to quantify specific psychological constructs but also to examine emergent patterns, contextual nuances, and complex semantic structures, providing a comprehensive understanding of language use and its psychological implications [2].

4.3 Limitations

Interpretation of our findings is limited by the absence of information about clients' diagnoses and annotations regarding the nature of phase transitions, indicating whether the transition represents improvement or worsening of symptoms. Other limitations include heterogeneity of participants, contextual limitations of LIWC, and the absence of fine-grained temporal resolution.

5 Conclusion

Our research suggests that language shifts hold potential as indicators of psychological change. Understanding these patterns may provide clinicians with more sensitive indicators of therapeutic progress, offering potential guidance for interventions, and improving the precision of treatment monitoring in inpatient psychiatric care.

6 Future Work

Future research could implement transformer-based neural network architectures (e.g., BERT, RoBERTa) to cluster participants according to symptom trajectories, such as improvement or deterioration. Analyses could then be conducted to examine differences in linguistic shifts across clusters. Where available, results from neural language models could be compared with clinical annotations to evaluate prediction accuracy. Future studies should aim to link these linguistic patterns more directly to specific mental states, ultimately supporting the development of clinically relevant interventions and applications.

Acknowledgments

This research was supported by Paracelsus Medical University, which also provided access to the clinical dataset utilized in this study. The language of this paper was revised with the assistance of ChatGPT-5.

References

[1] Evelina Fedorenko, Steven T. Piantadosi, and Edward A. F. Gibson. Language is primarily a tool for communication rather than thought. *Nature* 630, 8017 (Jul. 2024), 575–586. DOI: https://doi.org/10.1038/s41586-024-07522-w

[2] [2] Brendan Kennedy, Ashwini Ashokkumar, Ryan L. Boyd, and Morteza Dehghani. 2022. Text analysis for psychology: Methods, principles, and practices. In Handbook of Language Analysis in Psychology. Morteza Dehghani and Ryan L. Boyd (Eds.). The Guilford Press, New York, NY.

[3] Minna Lyons, Nazli D. Aksayli, and Gayle Brewer. Mental distress and language use: Linguistic analysis of discussion forum posts. Comput. Hum. Behav. 87 (Oct. 2018), 207–211. DOI:https://doi.org/10.1016/j.chb.2018.05.035

[4] Marco Spruit, Stephanie Verkleij, Kees de Schepper, and Floortje Scheepers. Exploring language markers of mental health in psychiatric stories. Appl. Sci. 12, 4 (Feb. 2022), 1–17. DOI:https://doi.org/10.3390/app12042179

[5] Günter K. Schiepek, Kathrin Viol, Wolfgang Aichhorn, Marc-Thorsten Hütt, Katharina Sungler, David Pincus, and Helmut J. Schöller. Psychotherapy is chaotic—(not only) in a computational world. Front. Psychol. 8 (Apr. 2017), 379. DOI:https://doi.org/10.3389/fpsyg.2017.00379

[6] Ryan L. Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W. Pennebaker. The development and psychometric properties of LIWC-22. University of Texas at Austin, Austin, TX. https://www.liwc.app

[7] Michael A. Cohn, Matthias R. Mehl, and James W. Pennebaker. Linguistic markers of psychological change surrounding September 11, 2001. Psychol. Sci. 15, 10 (Oct. 2004), 687–693. DOI:https://doi.org/10.1111/j.0956-7976.2004.00741.x

[8] Bernard Rosner, Robert J. Glynn, and Mei-Ling T. Lee. The Wilcoxon signed rank test for paired comparisons of clustered data. Biometrics 62, 1 (Mar. 2006), 185–192. DOI:https://doi.org/10.1111/j.1541-0420.2005.00389.x

[9] Boban Simonovic, Katia Correa Vione, Edward Stupple, and Alice Doherty. It is not what you think it is how you think: A critical thinking intervention enhances argumentation, analytic thinking and metacognitive sensitivity. Think. Skills Creat. 49 (Jun. 2023), 101362. DOI:https://doi.org/10.1016/j.tsc.2023.101362

[10] James W. Pennebaker, Cindy K. Chung, Joey Frazee, Gary M. Lavergne, and David I. Beaver. When small words foretell academic success: The case of college admissions essays. PLoS One 9, 12 (Dec. 2014), e115844. DOI:https://doi.org/10.1371/journal.pone.0115844

[11] Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. 2015. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, CLPsych 2015. June 5, 2015, Denver. Association for Computational Linguistics, Denver, CO, 1–10. DOI:https://doi.org/10.3115/v1/W15-1201

[12] Kyungil Kim, Seongjik Lee, and Changhwan Lee. College students with ADHD traits and their language styles. J. Atten. Disord. 19, 8 (Aug. 2015), 687–693. DOI:https://doi.org/10.1177/1087054712452343

[13] Marie Forgeard. Linguistic styles of eminent writers suffering from unipolar and bipolar mood disorder. Creat. Res. J. 20, 1 (Feb. 2008), 81–92. DOI:https://doi.org/10.1080/10400410701842094

[14] Barrett Anderson, Philippe R. Goldin, Keiko Kurita, and James J. Gross. Self-representation in social anxiety disorder: Linguistic analysis of autobiographical narratives. Behav. Res. Ther. 46, 10 (Oct. 2008), 1119–1125. DOI:https://doi.org/10.1016/j.brat.2008.07.001

[15] Raluca N. Trifu, Bogdan Nemeş, Carolina Bodea-Hategan, and Doina Cozman. Linguistic indicators of language in major depressive disorder (MDD): An evidence-based research. J. Evid.-Based Psychother. 17, 1 (Mar. 2017), 105–128.

[16] Michael L. Birnbaum, Sindhu K. Ernala, A. F. Rizvi, Elizabeth Arenare, Anna Van Meter, M. De Choudhury, and J. M. Kane. Detecting relapse in youth with psychotic disorders utilizing patient-generated and patient-contributed digital data from Facebook. NPJ Schizophr. 5, 1 (Dec. 2019), 17. DOI:https://doi.org/10.1038/s41537-019-0085-9

[17] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. J. Mach. Learn. Res. 3 (Mar. 2003), 993–1022.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In Proceedings of the 27th International Conference on Neural Information Processing Systems – Volume 2 (NIPS'13), December 5 - 10, 2013, Lake Tahoe Nevada. Curran Associates Inc., Red Hook, NY, USA, 3111–3119.

[19] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. J. Mach. Learn. Res. 3 (Mar. 2003), 1137–1155.

Measuring Therapist–Client Synchrony to Forecast Change Dynamics: EMA-based Protocol Pilot

Matej Vajda*
matej.vajda@mail.sfu.ac.at
Sigmund Freud University Vienna - Ljubljana branch
Ljubljana, Slovenia

Tine Kolenik*
Paracelsus Medical University
Salzburg, Austria
tine.kolenik@ccsys.de

Tatjana Rožič
Sigmund Freud University Vienna - Ljubljana branch
Ljubljana, Slovenia
tatjana.rozic@sfu-ljubljana.si

Nuša Kovačević Tojnko
Outpatient Mental Health Clinic Pamina
Maribor, Slovenia
nusa@pamina.si

Gašper Slapničar
Jozef Stefan Institute
Ljubljana, Slovenia
gasper.slapnicar@ijs.si

Miran Možina
Sigmund Freud University Vienna - Ljubljana branch
Ljubljana, Slovenia
miramozinaslo@gmail.com

Günter Schiepek
Paracelsus Medical University
Salzburg, Austria
guenter.schiepek@ccsys.de

Wolfgang Aichhorn
Paracelsus Medical University
Salzburg, Austria
w.aichhorn@salk.at

Abstract

We examine the feasibility and utility of a therapist–client monitoring protocol based on Ecological Momentary Assessment (EMA), designed to detect synchrony and forecast change dynamics in routine psychotherapy. Using the Synergetic Navigation System (SNS), we combined daily client reports with brief pre-/post-session questionnaires from therapists and clients. N=7 (3 therapists, 4 clients) participated over 4–9 weeks, completing daily TPQ-SA surveys and pre/post EMPIS-Q ratings; end-of-study evaluations assessed feasibility and user experience. Usability and perceived data safety were rated highly, while perceived usefulness was mixed. Clients often experienced EMA as obligatory and of limited immediate value; therapists noted missing alliance items and requested side-by-side access to clients' post-session responses. Notification glitches and limited uptake of feedback interviews further reduced engagement. Findings indicate that daily and session-based monitoring is feasible, but its value depends on workflow integration, a stronger relational focus, and reliable implementation. The very small sample and reliance on self-report limit generalizability. Future work will run a larger feasibility trial, refine questionnaires (including alliance items and paired therapist–client views), and pilot multimodal synchrony measures (session audio/video and physiology) toward scalable process–outcome monitoring.

Keywords

psychotherapy, change dynamics, synergetic navigation system, protocol, pilot

*Both authors contributed equally to this text.

Information Society 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.cogni.10

1 Introduction

Mental disorders contribute substantially to the global burden of disease. Recent estimates suggest that ~1 in 8 people were living with a mental disorder in 2019, with sustained growth in disability-adjusted life years through 2021, underscoring the need for scalable, higher-resolution care processes [15]. Psychotherapy remains a cornerstone of treatment; beyond specific techniques, robust evidence points to common therapeutic factors, especially the working alliance—being consistently linked to outcomes [25, 6]. Feedback-oriented psychotherapy (routine outcome monitoring and in-session process feedback) attempts to surface change signals early enough for course corrections, though effect sizes vary by context and tool [4, 14]. Parallel advances in mobile sensing and AI (e.g., digital phenotyping, multimodal modeling) enable dense, real-world measurement and interpretation of change processes at the individual (idiographic) level [3, 9].

The broader study we are building toward leverages these trends via a digital-twin approach that fuses session audio/video and physiology with ecological momentary assessments (EMA) to detect interpersonal synchrony, forecast tipping points, and inform just-in-time guidance to therapists [8].

This paper reports a focused pre-study, an EMA-based protocol pilot in three therapists and four clients—primarily to validate instruments, apps, and workflows, and to de-risk the methodological core for the larger study.

2 Related Work

High-frequency, routine monitoring of psychotherapy using the Synergetic Navigation System (SNS) and the Therapy Process Questionnaire (TPQ) has been piloted in several small studies. Concept and feasibility papers report that equidistant, daily self-ratings can be integrated into clinical settings with good compliance, especially when coupled to regular feedback sessions. Case-based work in suicide prevention, for example, showed near-perfect adherence across a 90-day period and emphasized that structured feedback, rather than the mere act of monitoring, appears critical for sustained engagement. More recent feasibility studies using personalized, daily process items in outpatient populations similarly found high perceived usefulness, alongside

caution about burden over longer durations (ie, signs of response fatigue). Together, these pre-studies suggest feasibility in routine care, but also highlight the importance of how data are fed back to patients and clinicians–so they can adjust focus or goals before progress falters or therapy discontinues [19, 23, 5, 12].

A second stream of pre-studies has tested EMA and feedback as an adjunct to therapy. In depressed outpatients, randomized trials of experience-sampling with weekly, personalized feedback reported that the add-on was both feasible and associated with symptom improvements versus controls; follow-on protocols have focused on pragmatic implementation and personalization. Broader EMA reviews in mental health consistently conclude that adherence is acceptable when sampling is purposeful, notifications are reliable, and feedback is built in—while also warning that implementation details (timing, prompt load, and integration into care) strongly determine perceived value [10, 1, 13].

Finally, dyadic, session-by-session monitoring has been moving toward "dual-perspective" designs that track both sides of the interaction (e.g., alliance after each session) and complement end-of-treatment measures with qualitative interviews. These studies foreground the clinical utility of post-session reflections and alliance tracking—elements our pilot also probes via therapist pre/post 5-factor ratings and client post-session reports alongside daily EMA [16].

3 Research Objectives

Overall aim of the larger program: To model psychotherapeutic change as a nonlinear, dyadic system and to forecast clinically actionable dynamics (e.g., ruptures, sudden gains) by integrating daily EMA with session-level synchrony and standardized questionnaires, grounded in Schiepek's five-factor model of change (EMPIS: **e**motions, **m**otivation to change, therapeutic **p**rogress/success, **i**nsight, problem **s**everity) [18]. To correlate observable behavioural and physiological signals, alongside textual data (transcripts, diaries), to aforementioned change dynamics, using state-of-the-art multi-modal deep learning approaches.

Specific objectives of this pre-study: i) Feasibility and fidelity. Verify day-to-day adherence, app stability, and data completeness for therapist/client EMA and session questionnaires (therapist pre/post "planned vs. enacted" five-factor ratings; client post-session intervention perceptions); ii) User experience and ethics. Collect end-of-month evaluations from both groups to surface burden, privacy, and workflow issues to remediate before scale-up (and align with transparency and equity safeguards) [3].

4 Materials and Methods

4.1 Ecological Momentary Assessment (EMA)

We used EMA to obtain equidistant, high-frequency measurements of clients' therapy-process states during everyday life. Limiting assessment to in-session ratings risks irregular or low-frequency sampling; in contrast, brief daily EMA increases ecological validity and yields time series suitable for detecting nonlinear change features and feeding back clinically meaningful signals [20]. In this pilot, clients completed one short smartphone survey per day for approximately one month, targeting core state variables and change-relevant markers aligned with the five-factor framework. This design supports forecasting/early-warning analyses and provides material for collaborative reflection in subsequent sessions [20].

4.2 Synergetic Navigation System (SNS)

Data collection and real-time monitoring were implemented with the SNS, a secure, web-based platform that schedules questionnaires at arbitrary intervals, supports Likert-type and V/A inputs across devices, and visualizes raw time-series for therapist and client use. It includes built-in analyses enabling process-oriented feedback and individualized decision-making [20, 21]. In this study we deployed the shortened TPQ for ambulatory use (TPQ-SA) for clients' daily reports and custom five-factor pre-/post-session forms for therapists and clients; accounts were protected via HTTPS with anonymized usernames and passwords, and outputs were available for optional feedback discussions [20, 21].

4.3 EMPIS questionnaire (EMPIS-Q)

The therapist pre-/post-session instruments were developed via a theory-driven item-generation and expert-consensus workflow grounded in Schiepek's five-factor change model EMPIS [18]. Concretely, three domain experts independently drafted candidate items to operationalize each factor for a pre-session "planned influence" pass and a post-session "realized influence/valence" pass. We then conducted iterative expert panel review to judge content relevance, clarity, and redundancy, reconciling wording by consensus—an approach consistent with standard content-validity procedures (e.g., expert-judge review, or modified Delphi-style consensus). Drafts were piloted internally to check interpretability and response burden, and response formats were simplified to Likert-type scales to fit the session workflow. Because the study operated bilingually, the final Slovenian versions were translated and back-translated to secure conceptual equivalence before deployment. This sequence aligns with recommended steps in scale development (theory-driven item generation → expert review for content validity → small-scale pretest) and with cross-cultural adaptation guidelines [2, 11].

4.4 Single-item Outcome Measure

EPO-1 is a single-item instrument, evaluating the responders' current emotional and psychological well-being [7]. It is used dimensionally with a visual analog scale (0: "Very poorly; I can barely manage to deal with things" to 100: "Very well; I have no important complaints").

4.5 Therapy Process Questionnaire – Short Ambulatory Use (TPQ-SA)

Daily measurements using the TPQ-SA [21, 17] (shortened version with 24 items for ambulatory use) yield time series data of psychotherapies that allow for capturing and identifying diversity and complexity of cases, as well as critical instabilities and nonstationarities. Unpredictability and complexity of change processes thus make close monitoring important.

4.6 Evaluation Questionnaire

Upon completion of data collection, tailored evaluation questionnaires were distributed to therapists and clients to gather feedback on study participation. Sections covered general research information (use of personal data, voluntariness, support availability, and need for additional information or training; for clients, also how the therapist presented the research objectives), therapist evaluation of pre- and post-session questionnaires (clarity, relevance of the five client variables, perceived influence

on session conduct, question quantity, and post-session usefulness; with space for comments), client evaluation of post-session and daily questionnaires (clarity, contribution to session comprehension, relevance, question quantity, response difficulty, utility, completion time, app prompt suitability, reference period, mode of completion, feedback interview experience if relevant, and comments), user experience (app/website usability, input preferences, timing, technical issues), and demographics (gender, age range). A separate free-text field allowed participants to provide additional comments to the research team.

4.7 Participants

The sample comprised two groups: three therapists, who were selected through convenience sampling based on prior knowledge and experience with SNS, and their four clients, who were identified through snowball sampling from therapists' current caseloads, with inclusion criteria of an established therapeutic alliance and therapist-assessed likelihood of consent to participate (see Table 1). Some authors also served as therapists in this study. Participation was voluntary and was not compensated. All the participants signed an informed consent form.

4.8 Feedback Interviews

In feedback interviews, therapist and client review visualisations of collected questionnaire data, with the therapist following rather than interpreting [22].

4.9 Procedure

EMPIS-Q was adapted for pre- and post-session use, translated into Slovenian, reviewed by four authors, and back-translated for accuracy. The Slovenian translation of the TPQ-SA was employed. All measures were implemented in the SNS. The visual analogue scale was replaced with Likert scales to assess perceived influence and valence, and a free-text field was added for optional comments. The EPO-1 was later included in the post-session client questionnaire.

Therapists familiar with SNS were invited (four approached; three consented). The research team did not rehearse client presentations or review questionnaires with them, and conducting feedback interviews was recommended (not mandatory). Therapists recruited clients, explained the study, obtained written informed consent, introduced SNS and questionnaires, and forwarded installation instructions and login credentials provided by the research team, which also activated client accounts in SNS. During data collection, clients received daily smartphone notifications to complete the TPQ-SA.

At study completion, EMPIS-Qs were deactivated; however, therapist–client dyads could continue using TPQ-SA voluntarily (two did). Therapists thanked clients, and the research team expressed appreciation to therapists. Clients and therapists each completed separate evaluation questionnaires.

5 Results

5.1 Evaluation Questionnaire - Therapists

Three therapists completed the evaluation (see Table 2). Perceived data safety was high (M=4.6/5). The pre-session instrument was rated comprehensive and relevant (both M=4.0) but had mixed impact on session conduct (ratings 2–4), reflecting unpredictable session topics and overlap among categories. The post-session instrument scored higher on comprehensiveness/relevance (M=4.6) with moderate perceived usefulness (M=4.0). Therapists asked to add items on alliance, emotions, and session atmosphere; they also requested side-by-side access to clients' post-session response. App usability was high (M=4.6); preferences included optional free-text and push notifications. One minor display issue was noted (line breaks for long Likert labels).

5.2 Evaluation Questionnaire - Clients

Four clients completed the evaluation (see Table 3). Presentation clarity was moderate (purpose $M = 3.5$, procedure $M = 3.75$), while perceived data safety was very high ($M = 5$). The daily TPQ-SA was seen as moderately comprehensive ($M = 4.25$) and moderately difficult ($M = 3.75$) but of limited immediate usefulness ($M = 2.75$), with comments about obligation and notification timing; typical completion time was 3–4 minutes. The post-session questionnaire was rated comprehensive ($M = 4.5$) and moderately helpful ($M = 4.0$), with slightly lower perceived relevance ($M = 3.75$); most clients found the item count appropriate. App ease of use was high ($M = 4.75$), though clients noted irregular/missing prompts and requested notification timing control; one initial login issue was reported. One feedback interview was conducted and described as yielding no major insights.

6 Discussion

This preliminary study aimed to evaluate the feasibility and effectiveness of a session-by-session monitoring system for both therapists and clients, in parallel to daily client measurements. The feedback gathered provides valuable insights into the strengths and areas for improvement of this methodology, which is intended to inform a larger-scale process-outcome study focusing on predicting therapeutic change.

Our findings echo prior pre-studies in three ways. First, feasibility with caveats: like earlier pilots, therapists and clients rated usability highly, yet clients sometimes experienced daily EMA as an obligation and usefulness dipped without structured feedback; prior work shows adherence and perceived value rise when regular feedback interviews are part of the protocol, something underused in our pilot (one interview only). Second, content focus: therapists' request to add alliance/emotion-of-the-session items mirrors dyadic monitoring protocols that track alliance every session; incorporating these in our post-session set should increase clinical relevance. Third, implementation details matter: notification glitches and timing issues we observed are the same levers highlighted in EMA literature as determinants of engagement. In short, our results are consistent with earlier pilots—daily monitoring is workable and accepted, but its utility depends on closing the loop (feedback), tuning item sets to the relational process, and getting the micro-UX right [12, 5, 16, 13, 24].

7 Limitations

This pilot has several limitations. The sample was very small (three therapists, four clients) and recruited by convenience/snowball methods, with role overlap (some authors as therapists), limiting generalizability and introducing possible expectancy and social–desirability biases. The observation window was short with few sessions per client, and feedback interviews were rarely used (one dyad), so acceptability and utility may be underestimated or mischaracterized. All primary measures were self-report; some post-session entries were delayed, possibly increasing recall bias but also possibly affording additional reflective processing and thus more considered responses. The session-related

Table 1: Participants and Data Collection Overview

Therapists			Clients					
	Gender	Age Range		Gender	Age Range	Feedback Interviews	No. of Sessions	Daily TPQ Duration
T1	M	35-45	C1	M	35-45	1	5	9 weeks
			C2	F	25-35	0	4	9 weeks
T2	F	45-55	C3	F	45-55	0	5	4 weeks
T3	F	35-45	C4	F	25-35	0	4	5 weeks
			4			1	18	

Table 2: Therapist Evaluation Questionnaire Results

Domain	M / Rating	Insights / Feedback	Illustrative Quote(s)
General Experience	Safety M = 4.6 (1–5)	Need for clearer guidance	"clearer instructions for presenting to clients' (T1)'; "support in explaining technical aspects" (T2); "rehearsing client presentations" (T3)
Pre-session Questionnaire	Comprehensiveness M = 4; Relevance M = 4; Influence rated 4 and 2	Difficult to predict topics; predefined aspects often intertwined; categories not natural but influenced focus; suggestion to rephrase items (wishes vs. intentions); frustration at not knowing client experience	"I was mostly guessing…clients came up with topics that changed the session course." (T1); "I rarely think about the aspects in such a structured way…that's why it influenced the session." (T2)
Post-session Questionnaire	Comprehensiveness & Relevance M = 4.6; Usefulness M = 4.0	Number of items appropriate; easier to complete than pre-session; items clear but limited new insight; missed items on alliance, emotions, session atmosphere; difficulty categorizing events; interested in client post-session reports	"relational aspect was missing" (T2); "I missed questions for [...] reflection, e.g. regarding alliance or feelings." (T1); "Sometimes it was difficult to determine exactly in which area something happened for the client." (T1)
User Experience (App)	Usability M = 4.6	All used mobile app; Likert scales suitable for focus; suggestion to add free-text fields; post-session timing added time pressure; preference for push notifications; minor technical issue	"It was an extra commitment between an already tight time-window between sessions." (T1); "... notifications [...] would remind me to fill in the questionnaire..." (T2)

Note. Ratings are on a 1–5 Likert scale (1 = low/poor, 5 = high/excellent). M denotes the mean across $N = 3$ therapists. Quotes translated from Slovene.

Table 3: Client Evaluation Questionnaire Results

Domain	M / Rating	Insights / Feedback	Illustrative Quote(s)
General Experience	Purpose clarity M = 3.5; Procedure clarity M = 3.75; Safety M = 5	Clients felt very safe; C4 unsure about right to withdraw; all knew whom to contact; mixed responses on support/presentation prior to study	—
TPQ	Comprehensiveness M = 4.25; Difficulty M = 3.75; Usefulness M = 2.75	Often experienced as obligation; sometimes forgotten; completed in 3–4 minutes; varied opinions on timing; some items too general/redundant; valued specificity of emotion items; mixed views on item count; one feedback interview, no major insights	"challenging because I often forgot and solved things in hindsight… [...] felt some pressure to complete it." (C1); "understandable, simple, but [...] a kind of obligation." (C3); "I answered all the items in a section the same"; "there weren't any groundbreaking insights." (C1)
Post-session Questionnaire	Comprehensiveness M = 4.5; Helpfulness M = 4.0; Relevance M = 3.75	Number of items appropriate (3) or excessive (1); items clear but limited usefulness for reflection	"I wouldn't say that the items were particularly useful in reflecting on the therapy itself. But they were clear." (C1)
User Experience	App Ease of use M = 4.75	Technical issues: missing or irregular prompts; initial login difficulty (C3); desire for more control over notifications	"notification… appeared exactly the other way around as it should… It would be better if I had some control over [it]." (C1); "The possibility to set the notification time." (C4)

Note. Ratings are on a 1–5 Likert scale (1 = low/poor, 5 = high/excellent). M denotes the mean across $N = 4$ clients. Quotes translated from Slovene.

instruments underwent content-focused development only and were not psychometrically validated; bilingual translation/back-translation may still leave subtle construct drift. Platform issues (notification irregularities, a display bug for long labels) may have affected adherence. Finally, the study did not include the planned multimodal synchrony streams (audio/video/physiology) or session-level alliance items, constraining insight into barriers additional data collection methods might introduce.

8 Conclusion and Future Work

This pre-study primarily assessed feasibility and barriers rather than effects. Daily EMA plus brief session questionnaires proved implementable and acceptable, but value depended on workflow fit and feedback loops. Key friction points were methodological (very small convenience sample with role overlap; self-report only; delayed post-session entries), procedural (rare use of feedback interviews; pre-session "planned influence" sometimes felt guess-like amid emergent themes; lack of explicit alliance/emotion-of-session coverage), and technical (notification irregularities; minor display issues). These constraints shaped engagement and data quality as much as the instruments themselves, underscoring that successful monitoring is a service-design problem—stable micro-UX, clear rationale, and structured feedback, not merely a measurement problem.

Future work will be a larger trial focused on de-risking these barriers. It will prioritize pragmatic endpoints (adherence, timeliness, missingness, prompt reliability, time-to-completion, usability, protocol fidelity) and data-quality safeguards (harmonized scales, timestamp checks to quantify recall lag, basic psychometrics for EMPIS-Q). Finally, it will pilot the planned multimodal streams (session A/V and physiology) strictly for feasibility (consent rates, capture success, clinician burden) before testing prognostic utility in subsequent outcome-focused studies.

Funding

This work was partly funded by a Sigmund Freud University Vienna internal Initial Funding project grant (January–May 2025).

Acknowledgements

The authors would like to thank the participating clients for their time and effort.

References

[1] Jojanneke A. Bastiaansen, Maaike Meurs, Renee Stelwagen, Lex Wunderink, Robert A. Schoevers, Marieke Wichers, and Albertine J. Oldehinkel. 2018. Self-monitoring and personalized feedback based on the experiencing sampling method as a tool to boost depression treatment: a protocol of a pragmatic randomized controlled trial (zelf-i). *BMC Psychiatry*, 18, 1, (Sept. 2018). doi:https://doi.org/10.1186/s12888-018-1847-z.

[2] Godfred O. Boateng, Torsten B. Neilands, Edward A. Frongillo, Hugo R. Melgar-Quiñonez, and Sera L. Young. 2018. Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Frontiers in Public Health*, 6, 149, (June 2018). doi:https://doi.org/10.3389/fpubh.2018.00149.

[3] Pasquale Bufano, Marco Laurino, Sara Said, Alessandro Tognetti, and Danilo Menicucci. 2023. Digital phenotyping for monitoring mental disorders: systematic review. *J Med Internet Res*, 25, (Dec. 2023), e46778. doi:10.2196/46778.

[4] Kim de Jong, Judith M. Conijn, Roisin A.V. Gallagher, Alexandra S. Reshetnikova, Marya Heij, and Miranda C. Lutz. 2021. Using progress feedback to improve outcomes and reduce drop-out, treatment duration, and deterioration: a multilevel meta-analysis. *Clinical Psychology Review*, 85, (Apr. 2021). doi:https://doi.org/10.1016/j.cpr.2021.102002.

[5] Clemens Fartacek, Günter Schiepek, Sabine Kunrath, Reinhold Fartacek, and Martin Plöderl. 2016. Real-time monitoring of non-linear suicidal dynamics: methodology and a demonstrative case report. *Frontiers in Psychology*, Volume 7 - 2016. doi:10.3389/fpsyg.2016.00130.

[6] Christoph Flückiger, A C Del Re, Bruce E Wampold, and Adam O Horvath. 2018. The alliance in adult psychotherapy: a meta-analytic synthesis. en. *Psychotherapy (Chic.)*, 55, 4, (Dec. 2018), 316–340.

[7] Miguel M. Gonçalves et al. 2024. Developing a european psychotherapy consortium (epoc): towards adopting a single-item self-report outcome measure across european countries. *Clinical Psychology in Europe*, 6, 3, (Sept. 2024), 1–15. doi:10.32872/cpe.13827.

[8] Evangelia Katsoulakis et al. 2024. Digital twins for health: a scoping review. *npj Digital Medicine*, 7, 1, (Mar. 2024), 1–11. doi:https://doi.org/10.1038/s41746-024-01073-0.

[9] Tine Kolenik. 2022. Methods in digital mental health: smartphone-based assessment and intervention for stress, anxiety, and depression. In *Integrating Artificial Intelligence and IoT for Advanced Health Informatics: AI in the Healthcare Sector*. Carmela Comito, Agostino Forestiero, and Ester Zumpano, editors. Springer International Publishing, Cham, 105–128. ISBN: 978-3-030-91181-2. doi:10.1007/978-3-030-91181-2_7.

[10] Ingrid Kramer et al. 2014. A therapeutic application of the experience sampling method in the treatment of depression: a randomized controlled trial. *World Psychiatry*, 13, 1, (Feb. 2014), 68–77. doi:https://doi.org/10.1002/wps.20090.

[11] Mary R. Lynn. 1986. Determination and quantification of content validity. *Nursing Research*, 35, 6, (Nov. 1986), 382–386. doi:https://doi.org/10.1097/00006199-198611000-00017.

[12] Rosa Michaelis, Friedrich Edelhäuser, Yvonne Hülsner, Eugen Trinka, and Günter Schiepek. 2022. Personalized high-frequency monitoring of a process-oriented psychotherapeutic approach to seizure disorders: treatment utilization and participants' feedback. *Psychotherapy*, 59, 4, (Feb. 2022), 629–640. doi:https://doi.org/10.1037/pst0000430.

[13] Inez Myin-Germeys, Zuzana Kasanova, Thomas Vaessen, Hugo Vachon, Olivia Kirtley, Wolfgang Viechtbauer, and Ulrich Reininghaus. 2018. Experience sampling methodology in mental health research: new insights and technical developments. *World Psychiatry*, 17, 2, (May 2018), 123–132. doi:https://doi.org/10.1002/wps.20513.

[14] Ole Karkov Østergård, Hilde Randa, and Esben Hougaard. 2018. The effect of using the partners for change outcome management system as feedback tool in psychotherapy—a systematic review and meta-analysis. *Psychotherapy Research*, 30, 2, (Sept. 2018), 1–18. doi:https://doi.org/10.1080/10503307.2018.1517949.

[15] The Lancet Psychiatry. 2024. Global burden of disease 2021: mental health messages. *The Lancet Psychiatry*, 11, 8, (Aug. 2024), 573. doi:10.1016/S2215-0366(24)00222-0.

[16] Yvonne Schaffler, Andrea Jesser, Elke Humer, Katja Haider, Christoph Pieh, Thomas Probst, and Brigitte Schigl. 2024. Process and outcome of outpatient psychotherapies under clinically representative conditions in austria: protocol and feasibility of an ongoing study. *Frontiers in psychiatry*, 15, (Mar. 2024). doi:https://doi.org/10.3389/fpsyt.2024.1264039.

[17] Günter Schiepek. 2022. Prozess- und outcome-evaluation mithilfe des synergetischen navigationssystems (sns). German. *Psychotherapie-Wissenschaft*, 12, 1, 43–56. „Der TPB umfasst für die ambulante Therapie 33 Items (Kurzfassung: 24 Items).". https://www.psychotherapie-wissenschaft.info/article/view/3969.

[18] Günter Schiepek, Benjamin Aas, and Kathrin Viol. 2016. The mathematics of psychotherapy: a nonlinear model of change dynamics. *Nonlinear dynamics, psychology, and life sciences*, 20, 3, (July 2016), 369–99. https://pubmed.ncbi.nlm.nih.gov/27262423/.

[19] Günter Schiepek, Benjamin Aas, and Kathrin Viol. 2016. The mathematics of psychotherapy: a nonlinear model of change dynamics. *Nonlinear dynamics, psychology, and life sciences*, 20 3, 369–99. https://api.semanticscholar.org/CorpusID:40177925.

[20] Günter Schiepek, Wolfgang Aichhorn, Martin Gruber, Guido Strunk, Egon Bachler, and Benjamin Aas. 2016. Real-time monitoring of psychotherapeutic processes: concept and compliance. *Frontiers in Psychology*. doi:10.3389/fpsyg.2016.00604.

[21] Günter Schiepek, Wolfgang Aichhorn, and Guido Strunk. 2012. Der therapieprozessbogen (tpb)–faktorenstruktur und psychometrische daten. *Zeitschrift für Psychosomatische Medizin und Psychotherapie*, 58, 3, 257–266.

[22] Günter Schiepek, Heiko Eckert, Benjamin Aas, Sebastian Wallot, and Anna Wallot. 2016. *Integrative psychotherapy: A feedback-driven dynamic systems approach*. Hogrefe Publishing GmbH.

[23] Günter Schiepek, Barbara Stöger-Schmidinger, Helmut Kronberger, Wolfgang Aichhorn, Leonhard Kratzer, Peter Heinz, Kathrin Viol, Anna Lichtwarck-Aschoff, and Helmut Schöller. 2019. The therapy process questionnaire - factor analysis and psychometric properties of a multidimensional self-rating scale for high-frequency monitoring of psychotherapeutic processes. *Clinical Psychology Psychotherapy*, 26, 5, (July 2019), 586–602. doi:https://doi.org/10.1002/cpp.2384.

[24] Matej Vajda. 2024. Barriers and facilitators to the introduction of feedback-informed treatment in organisations: a review of research. *Kairos–Slovenian Journal of Psychotherapy*, 18, 3-4.

[25] Bruce E Wampold. 2015. How important are the common factors in psychotherapy? an update. en. *World Psychiatry*, 14, 3, (Oct. 2015), 270–277.

Towards a Possible Solution of Chalmers' Hard Problem and to Definitions of Life and Consciousness

Marko Vitas†
Independent Researcher
Laze pri Borovnici 38,
Borovnica,
1353 Slovenia
vitas.marko83@gmail.com

Abstract/Povzetek

There is no consensus about what cognition and its emergent form, consciousness, are. Yet this expanded abstract proposes a new definition of consciousness. As many researchers, philosophers and other thinkers believe that life means cognising, this new definition of consciousness stems from a generalisation of the existing Vitas & Dobovišek definition of life which postulates that *Life is a far from equilibrium self-maintaining chemical system capable of processing, transforming, and accumulating information acquired from the environment*. The new definition includes the thermodynamical aspect as a far from equilibrium system and considers the flow of information from the environment to a conscious system. The new definition of consciousness is formulated in a minimal manner; simultaneously, it is general enough to cover all emergent forms of cognition, *e.g.* thinking and rationality. The newly formulated definition states that *Consciousness is an emergent property of a far from equilibrium system of quantum particles sustained by an autopoietic system and capable of processing, transforming, and accumulating information acquired from the environment*. The newly proposed definition of consciousness may be of interest to cognitive and computer sciences – and even to the development of artificial intelligence. I propose a possible another alternative generalisation by introducing quantum particles to the Vitas & Dobovišek definition of life which refining it into a broader concept: *Life is a far from equilibrium self-maintaining system of quantum particles capable of processing, transforming, and accumulating information acquired from the environment.* A question might be posed here, whether we are not therefore encompassing other complex forms of matter which cannot be considered as life. It is here worth mentioning that some authors, for instance, consider dusty plasmas from the thermosphere – although they are not self-sustaining – as a fourth state of matter and fourth domain of life, something in between non-living and living matter. Newly formulated definition of consciousness presents a possible solution to Chalmers' hard problem of consciousness. By including quantum particles which do not have classical trajectories, in my definition of consciousness, which is apparently an emergent property of cognition, the solution to the Chalmers' hard problem may be found in the introduction of additional multiple dimensions. Organisms (including individual cells) are those who interpret; the interpretation process or semiosis (in the sense of C. Peirce) is the process of life. Digital coding might relate to the reduction of dimensions, and it is highly context-dependent, like digital coding of analogue protein three-dimensional structures in the unidimensional linear, genetic sequences or vice versa expanding of dimensions after interpretation, translation of linear digital genetic sequences into three-dimensional analogue protein structures. At this point, it is worth mentioning that interpretants should have the same dimension as analogue structures, providing additional information. Each biopolymer is an emergent molecule. Evolution gives rise to emergence. Undoubtedly, there is an emergence of three-dimensional structures from linear unidimensional digital sequences. Likewise, consciousness is an interpretant of the signals coming from the environment. Adding extra dimensions for the interpretant might shed new light on problems connected with consciousness, including Chalmers' hard problem. Yet perhaps a question worth posing at this point is whether we are not living in some sort of hyper-digital world coding for a hyper-analogue world. Could this view present a possible solution to Chalmers' hard problem of consciousness?

Keywords/Ključne besede

Definition of Consciousness, Definition of Life, Origins of Life, Chalmers Hard Problem, Cognition, Far from Equilibrium

Acknowledgments/Zahvala

Sincere thanks to Andrei Igamberdiev for reading the manuscript and suggesting valuable clues and constructive comments to enhance its quality. The author is also grateful to David H. Wolpert, Jan Karbowski, Pamela Lyon and Andrej Dobovišsek for insightful comments regarding issues presented in this Abstract. Thanks also to Arto Annila for reading my articles and everyone who contributed to successful launch the newly proposed definition of consciousness .

Information Society 2025, 6–10 October 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.cogni.1

References/Literatura

Vitas, M. & Dobovišek, A. (2017) On a quest of reverse translation. Found Chem 19, 139–155. https://doi.org/10.1007/s10698-016-9260-5

Vitas, M. & Dobovišek, A. (2019) Towards a General Definition of Life. Origins of Life and Evolution of Biospheres 49, 77–88. https://doi.org/10.1007/s11084-019-09578-5

Vitas, M. (2025) Towards a Possible Definition of Consciousness. BioSystems 254, 105526. https://doi.org/10.1016/j.biosystems.2025.105526

Analiza kognitivnih zmogljivosti LLM: Strateško načrtovanje z uporabo testa Tower of London

Evaluating LLM Cognitive Capabilities: A strategic Planning Analysis Using the Tower of London Test

Katarina Žužek
Kognitivna znanost
Univerza v Novem mestu
Fakulteta za ekonomijo in informatiko
Slovenija

Matjaž Gams
Oddelek za inteligentne sisteme
Institut "Jožef Stefan"
Jamova cesta 39, 1000 Ljubljana
Slovenija

Povzetek

Prispevek raziskuje zmožnosti strateškega načrtovanja velikih jezikovnih modelov (LLM) z uporabo besedilne različice testa Tower of London (ToL). Preizkušenih je bilo pet najzanimivejših modelov v času testiranja, in sicer: DeepSeek V3, Grok 3, Gemini 2.0 Flash, Qwen 235B-A22B in Mistral 12B, na nalogah različnih zahtevnosti. Uspešnost modelov je bila tesno povezana z njihovo arhitekturo in velikostjo, pri čemer so se pri nalogah z visoko kognitivno zahtevnostjo pojavile jasne omejitve. Rezultati poudarjajo potencial LLM-jev za približek kognitivnim procesom človeka, obenem pa opozarjajo na potrebo po optimizaciji učnih podatkov in razvoju naprednejših arhitektur za izboljšanje strateškega načrtovanja. Raziskava prispeva k povezovanju kognitivne znanosti in umetne inteligence ter odpira nove možnosti za uporabo standardiziranih psiholoških testov pri ocenjevanju LLM-jev.

Ključne besede

umetna inteligenca, veliki jezikovni modeli, načrtovanje, Tower of London, kognitivna znanost

Abstract

This paper investigates the strategic planning capabilities of large language models (LLM) using a text-based adaptation of the Tower of London (ToL) test. Five of the most relevant at the time of testing models were evaluated: DeepSeek V3, Grok-3, Gemini 2.0 Flash, Qwen 235B-A22B, and Mistral 12B, on tasks of varying complexity. Performance was closely tied to model architecture and size, with clear limitations emerging in highly cognitive tasks. The findings highlight LLMs potential to approximate human cognitive processes while emphasizing the need for optimized training data and advanced architectures to enhance planning capabilities. This research bridges cognitive science and artificial intelligence, opening new avenues for using standardized psychological tests to evaluate LLMs.

Information Society 2025, 6-10 October 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.cogni.2

Keywords

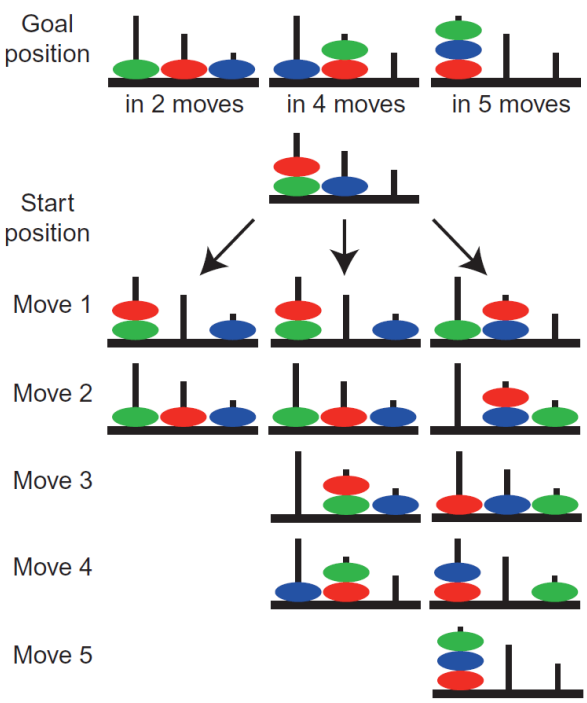
artificial intelligence, large language models, planning, tower of london, cognitive science

1 Uvod

V zadnjih desetletjih je področje umetne inteligence doživelo izjemen razvoj. Od klasičnih simbolnih pristopov, ki so skušali posnemati človeško logiko z eksplicitnimi pravili, smo prešli k adaptivnim modelom, ki temeljijo na nevronskih arhitekturah in se učijo iz obsežnih podatkovnih zbirk. Sodobni LLM-ji, temelječi na arhitekturi »Transformer«, so v zadnjem desetletju bistveno spremenili področje obdelave naravnega jezika [2, 3]. Njihova zmožnost generiranja koherentnih in semantično bogatih besedil je vzbudila zanimanje za njihove kognitivne zmogljivosti, ki presegajo zgolj jezikovno produkcijo in posegajo na področja, kot sta logično sklepanje in strateško načrtovanje. Kljub velikim izboljšavam na jezikovnem področju so LLM-ji še vedno omejeni pri nalogah abstraktnega razumevanja, vzročno-posledičnega sklepanja in dolgoročnega načrtovanja. Te omejitve so še posebej izrazite pri kompleksnih problemih, ki zahtevajo ohranjanje informacij v delovnem spominu in prilagodljivo mišljenje [4]. Podobno opozarjajo tudi Binz idr. [11], da LLM-ji kljub napredku pogosto odpovedo pri nalogah, ki zahtevajo večstopenjsko strateško načrtovanje in posploševanje v nepoznanih situacijah.

Test »Tower of London«, znan tudi kot »Hanojski stolpi« ali »Londonski stolpi«, je uveljavljeno orodje v kognitivni znanosti, ki omogoča natančno ocenjevanje prostorskega načrtovanja in logičnega sklepanja z nalogami preurejanja krogcev oz. kroglic na treh palicah [2, 9]. V zadnjih preglednih raziskavah je bilo poudarjeno, da bo za doseganje višjih kognitivnih sposobnosti ključno povezovanje nevronskih mrež s simbolnimi mehanizmi razumevanja [12]. Na Sliki 1 je predstavljeno zaporedje potez pri enostavni nalogi te igre.

Ta raziskava ocenjuje zmožnosti LLM-jev za strateško načrtovanje. Uporabili smo metodološko prilagojeno besedilno različico testa Tower of London, ki je zasnovana za merjenje izvršilnih funkcij in strateškega načrtovanja. Glavni cilj raziskave je ugotoviti, kako različni LLM-ji delujejo pri reševanju nalog različnih stopenj zahtevnosti in kako njihova arhitektura vpliva na njihovo uspešnost pri reševanju nalog.



Slika 1: Reševanje naloge v testu Tower of London.

2 Teoretični okvir in metodologija

Strateško načrtovanje je ključna izvršilna funkcija, ki vključuje zaporedje dejanj za dosego zastavljenega cilja. Omogoča organizacijo vedenja, oblikovanje strategij ter predvidevanje prihodnjih stanj, kar je bistveno za učinkovito prilagajanje kompleksnim okoljem. Test ocenjuje zmožnosti prostorskega načrtovanja z manipulacijo kroglic na palicah, pri čemer morajo udeleženci najti optimalno pot od začetne do ciljne konfiguracije. Za potrebe te raziskave smo test prilagodili v besedilno obliko, ki omogoča standardizirano interakcijo z LLM-ji.

2.1 Načrtovanje kot kognitivna funkcija in Tower of London

Strateško načrtovanje je ključna izvršilna funkcija, ki vključuje zaporedje dejanj za dosego zastavljenega cilja. Omogoča organizacijo vedenja, oblikovanje strategij ter predvidevanje prihodnjih stanj, kar je bistveno za učinkovito prilagajanje kompleksnim okoljem. Test ToL, ki ga je razvil Shallice [10], ocenjuje te zmožnosti z manipulacijo kroglic na treh palicah. Udeleženci morajo najti optimalno pot od začetne do ciljne konfiguracije z minimalnim številom potez ob upoštevanju pravil. Glavna zahteva testa je rešitev naloge z minimalnim številom potez, ob upoštevanju preprostih pravil, kot so premikanje ene kroglice na potezo in omejitve kapacitete palic. Uspešnost pri reševanju testa je povezana s funkcijami delovnega spomina, prilagodljivostjo razmišljanja in sposobnostjo zaviranja impulzivnih odločitev [9].

2.2 Prilagoditev testa ToL za LLM-je

Prilagoditev testa je bila ključna za omogočanje standardizirane interakcije z modeli, ki so primarno zasnovani za obdelavo in generiranje besedilnih podatkov. Vsaka naloga vsebuje natančen opis začetne in ciljne konfiguracije ter pravil za premikanje. Na primer, ena od nalog s petimi premiki je bila oblikovana takole: "Reši test ToL na plošči s tremi navpičnimi palicami, kjer so barvne kroglice razporejene takole: Palica 1: [rdeča (zgoraj), zelena (spodaj)], Palica 2: [modra (samostojna)], Palica 3: [prazna]. Ciljna konfiguracija: Palica 1: [modra (zgoraj), zelena (vmes), rdeča (spodaj)], Palica 2: [(prazna)], Palica 3: [(prazna)]. Navedite zaporedje potez za premik kroglic, ki doseže cilj z minimalnim številom korakov, pri čemer upoštevajte pravila: premikanje samo zgornje ali samostojne kroglice, premik ene kroglice na potezo, upoštevanje omejitev kapacitete palic (Palica 1: največ tri kroglice, Palica 2: največ dve, Palica 3: največ ena) in pravil gravitacije (kroglice padajo od zgoraj navzdol)." Zmožnost razumevanja navodil, ustvarjanja pravilnega zaporedja potez in doseganje ciljne konfiguracije z minimalnim številom potez je služila kot primarni kazalnik uspešnosti pri ocenjevanju strateškega načrtovanja [6] [9].

2.3 Izbira modelov in postopek testiranja

Raziskovalni vzorec je obsegal pet sodobnih, arhitekturno raznolikih LLM-jev: DeepSeek V3, Grok-3, Gemini 2.0 Flash, Qwen 235B-A22B in Mistral 12B. Modela DeepSeek V3 in Qwen 235B-A22B, znana po naprednih arhitekturah (npr. mešanica strokovnjakov) in obsežnih učnih podatkih, sta predstavljala zgornji razred zmogljivosti. Grok-3 in Gemini 2.0 Flash sta bila izbrana zaradi visokih hitrosti obdelave in specifičnih optimizacij, medtem ko je bil Mistral 12B vključen kot primer manjšega, a učinkovitega modela. Podoben eksperimentalni pristop so uporabili Xu in sod., ki so ocenjevali izvršilne funkcije umetne inteligence [13] z uporabo standardiziranih kognitivnih nalog različnih zahtevnosti.

Testiranje je bilo izvedeno z uporabo sedmih nalog testa ToL, katerih zahtevnost je bila določena s številom minimalnih potez (od 2 do 7). Vsaka naloga je od modelov zahtevala ustvarjanje zaporedja potez za dosego ciljne konfiguracije kroglic na palicah ob upoštevanju jasno opredeljenih pravil, ki posnemajo originalni test ToL [6]. Od modelov se je pričakovalo, da bodo ustvarili zaporedje potez, ki vodi do pravilne rešitve z minimalnim številom korakov.

Podatki so bili zbrani aprila 2025 v izoliranih sejah brez dodatnega konteksta, da bi zagotovili enotne pogoje za vse modele. Uspešnost je bila ocenjena z dvema meriloma: pravilnostjo in optimalnostjo. Pravilnost pomeni, ali je bila dosežena ciljna konfiguracija. Optimalnost smo merili glede na to, ali je model nalogo rešil z najmanjšim možnim številom potez, kar kaže na njegovo sposobnost strateškega načrtovanja. Ta merilo je bil določeno binarno: vrednost 1 je bila dodeljena le, če je model uporabil natanko toliko potez, kot je bilo teoretično minimalno za dano nalogo; v nasprotnem primeru (tudi pri eni sami dodatni potezi) je bila vrednost 0. Rešitev je bila ocenjena kot uspešno le, če je bila hkrati pravila in optimalna. To pomeni, da je model moral doseči ciljno konfiguracijo z natančno določenim minimalnim številom potez. Vsako odstopanje je bilo razumljeno kot neuspeh pri celotni nalogi. Rezultati so bili zabeleženi strukturirani Excel tabeli in statistično analizirani s programom SPSS z uporabo t-testa za en vzorec (za preverjanje uspešnosti nad 50 %) in Kruskal-Wallisovega testa za primerjavo uspešnosti med modeli [7].

2.4 Analiza podatkov

Zbrane podatke o uspešnosti posameznih LLM-jev pri nalogah različnih zahtevnosti smo analizirali z uporabo statističnega orodja SPSS. Osredotočili smo se na ocenjevanje uspešnosti glede pravilnost ter optimalnost rešitev. Statistična analiza je omogočila primerjavo uspešnosti med modeli in ugotavljanje statistično pomembnih razlik glede na njihovo arhitekturo in zahtevnost nalog.

Rezultati kažejo, da naraščajoča zahtevnost nalog negativno vpliva na uspešnost vseh modelov ($p < 0,01$). Grok-3 je dosegel najvišjo povprečno uspešnost (80,95 %, 17/21), sledita DeepSeek V3 (66,67 %, 14/21) in Qwen 235B-A22B (61,90 %, 13/21). Gemini 2.0 Flash (42,86 %, 9/21) in Mistral 12B (33,33 %, 7/21) sta pokazala nižjo uspešnost, zlasti pri nalogah, ki zahtevajo več kot štiri poteze. Noben model ni uspešno rešil naloge s sedmimi potezami, kar razkriva omejitve pri obvladovanju visoke kognitivne zahtevnosti.

Kruskal-Wallisov test je potrdil statistično pomembne razlike med modeli ($H = 22,03$, $df = 4$, $p < 0,001$). Qwen 235B-A22B in DeepSeek V3 sta izkazala večjo odpornost pri zahtevnih nalogah, kar lahko pripišemo naprednim arhitekturam, kot je mešanica modelov in obsežnim učnim podatkom [5]. T-test za en vzorec je pokazal, da povprečna uspešnost modelov (57,14 %, SD = 0,49) presega referenčni prag 50 % ($t(104) = 11,92$, $p < 0,05$), kar kaže na zmožnost sistematičnega reševanja nalog, čeprav zaostajajo za človeško uspešnostjo (70–80 %) pri zahtevnih nalogah [1].

3 Rezultati

Empirična evalvacija je potrdila statistično pomembno korelacijo med arhitekturno zahtevnostjo modelov in njihovo uspešnostjo pri strateškem načrtovanju. Rezultati so pokazali, da sta modela Qwen 235B-A22B in DeepSeek V3 dosegla najvišjo povprečno uspešnost pri nalogah različnih zahtevnosti, kar podpira tezo, da večji modeli z naprednimi arhitekturami učinkoviteje obvladujejo kompleksne naloge. Pri nalogah, ki so zahtevale 4 do 6 potez, je bila njihova uspešnost visoka, vendar se je pri najzahtevnejši nalogi s sedmimi znatno zmanjšala, kar razkriva omejitve trenutnih arhitektur.

Manjši modeli, kot sta Gemini 2.0 Flash in Mistral 12B, so bili učinkoviti pri enostavnejših nalogah (2-3 poteze), vendar so hitro dosegli svoje meje pri večji zahtevnosti.

4 Analiza in diskusija

Rezultati raziskave poudarjajo dvojnost zmogljivosti LLM-jev. Po eni strani modeli, kot sta DeepSeek V3 in Qwen 235B-A22B, kažejo izjemen potencial za reševanje kompleksnih kognitivnih nalog, ki presegajo zgolj jezikovno obdelavo. Uspešnost pri nalogah srednje zahtevnosti testa ToL nakazuje, da so njihove arhitekture, podprte z obsežnimi učnimi podatki, omogočajo določeno stopnjo logičnega sklepanja in zmožnostjo strateškega načrtovanja, ki je v nekaterih primerih primerljiva s človeškimi sposobnostmi. Kljub temu ti rezultati predstavljajo le statističen približek človeškim rešitvam v specifičnih okoliščinah in ne pomenijo simulacije kognitivnih procesov. To postane še posebej očitno pri kompleksnejših nalogah, ki presegajo zmožnosti modelov, saj le-ti delujejo na podlagi verjetnostnih korelacij in ne razumevanja. Po drugi strani raziskava razkriva značilne omejitve teh modelov. Vsi modeli so odpovedali pri najzahtevnejši nalogi ToL s sedmimi potezami, kar izkazuje pomanjkanje sposobnosti za abstraktno, večstopenjsko razmišljanje in dolgoročno načrtovanje. To je verjetno posledica njihove statistične narave, ki temelji na prepoznavanju vzorcev. Modeli se morda soočajo z "zastajanjem" v lokalnih optimumih in imajo omejene sposobnosti obdelave informacij v delovnem spominu, kar omejuje reševanje kompleksnih problemov.

Uporaba besedilnega formata za test ToL, čeprav omogoča standardizirano interakcijo z modeli, predstavlja določeno omejitev. Proces razumevanja in reševanja nalog v besedilni obliki se lahko bistveno razlikuje od vizualno-prostorskega pristopa, uporabljenega pri testiranju ljudi. Prihodnje raziskave bi se lahko osredotočile na hibridne pristope, ki bi združili jezikovne

modele z vizualnimi modeli, da bi preverili njihove zmožnosti reševanja problemov v multimodalnem kontekstu [8].

Poleg tega bi bilo zanimivo raziskati, kako se LLM-ji odzivajo na naloge z večjo stopnjo nejasnosti ali nedoločenosti, kar bi lahko vključilo metode iz področja teorije odločanja ali Bayesovskih mrež. Takšne naloge zahtevajo sposobnost ocenjevanja verjetnosti in izbiranja med alternativami ob nepopolnih informacijah, kar je pomembna lastnost napredne kognicije. S tem bi lahko razširili okvir za vrednotenje umetne inteligence z vidika adaptivnosti in robustnosti v realnih, nestandardnih situacijah [12]. V ta namen novejše raziskave predlagajo razvoj kognitivnih arhitektur z zmožnostjo notranje reprezentacije ciljev in večstopenjskega razmišljanja [14], kar bi lahko omogočilo učinkovitejše strateško načrtovanje v modelih.

Poleg testa ToL obstajajo številni drugi standardizirani kognitivni testi, kot so Wisconsin Card Sorting Test (WCST), Raven Progressive Matrices ali Stroopov test, ki bi jih bilo mogoče prilagoditi za evalvacijo LLM-jev. Ti testi vključujejo različne kognitivne domene, od fleksibilnega razmišljanja do sklepanja in inhibicije, in bi omogočili bolj celostno oceno umetne inteligence v primerjavi s človeškimi udeleženci.

Nadaljevanje raziskav v tej smeri bi lahko podprlo razvoj hibridnih modelov, ki vključujejo tako nevronske kot simbolne komponente, kar je v skladu s trenutnimi smernicami za razvoj razložljive in robustne umetne inteligence [11]. Poleg tega bi bilo smiselno vključiti teste analognega sklepanja, kot predlagata Ghosh in Holyoak [15], saj so ti pokazatelji višjega reda kognicije v modelih in so dobro uveljavljeni v psihologiji.

Za nadaljnje raziskave priporočamo razširitev vzorca, vključitev dodatnih vrst nalog (npr. logično-matematične, ustvarjalne) in poglobljeno analizo napak modelov. Pomembni so tudi etični vidiki, kot so zasebnost podatkov, pravičnost in okoljski vpliv treniranja modelov, so pomembni tudi v kontekstu raziskave ToL. Uporaba obsežnih podatkov za treniranje modelov lahko vključuje občutljive informacije, npr. podatke iz kognitivnih študij o človeških udeležencih, pristranskost v evalvacijskih podatkih lahko izkrivi rezultate pri simulaciji kognitivnih nalog, visoka poraba energije pri testiranju modelov pa vpliva na okolje. Ti vidiki zahtevajo skrbno obravnavo in dodatno pozornost [8].

5 Zaključek

Raziskava potrjuje pomemben potencial LLM-jev za simulacijo kognitivnih procesov, vendar so njihove zmožnosti strateškega načrtovanja še vedno omejene. To je še posebej očitno pri visokonivojskem abstraktnem sklepanju in reševanje nestandardnih problemov, ki zahtevajo večstopenjsko logiko. Za nadaljnji napredek bo ključen razvoj hibridnih pristopov, ki bi združili statistično moč globokega učenja z bolj simbolnimi in logičnimi pristopi k sklepanju.

Takšna integracija bi omogočila razvoj resnično robustne in razložljive umetne inteligence, ki bi lahko postala učinkovit partner pri reševanju kompleksnih problemov. Raziskava tako predstavlja most med kognitivno znanostjo in umetno inteligenco ter odpira vrata za nadaljnjo uporabo standardiziranih kognitivnih testov pri vrednotenju naprednih zmogljivosti umetne inteligence. Prihodnje študije bi morale raziskati tudi, kako se učni podatki in arhitekturne inovacije, kot so mehanizmi

delovnega spomina, odražajo v sposobnostih strateškega načrtovanja [8].

Raziskava, izvedena aprila 2025, je delno zastarela zaradi hitrega napredka na področju LLM-jev. Po njej so se pojavili novejši sistemi, kot so GPT-5, Claude 4, Grok-4 in Gemini 2.5, ki dosegajo boljše rezultate pri nalogah abstraktnega razmišljanja in kognitivnih izzivih, podobnih testu ToL. Ti dosežki nakazujejo potrebo po nadaljnjih raziskavah, da bi bolje razumeli zmogljivosti in omejitve sodobnih modelov umetne inteligence.

Literatura

[1] Boccia, M. idr. 2017. Test Tower of London (ToL) v Italiji: Standardizacija testa ToL v italijanski populaciji. Neurological Sciences, 38(7), 1263–1270. DOI: https://doi.org/10.1007/s10072-017-2957-y.

[2] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S. V., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N. S., Chen, A. S., Creel, K. A., Davis, J., Demszky, D., Donahue, C. in Liang, P. (2021). On the opportunities and risks of foundation models. https://doi.org/10.48550/arXiv.2108.07258.

[3] Brown, T. B., Mann, B., Ryder, N. in Amodei, D. (2020). Language models are few-shot learners. https://doi.org/10.48550/arXiv.2005.14165

[4] Chollet, F. (2019). On the measure of intelligence. https://doi.org/10.48550/arXiv.1911.01547.

[5] DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F. in Pan, Z. (2024). Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges. https://doi.org/10.48550/arXiv.2409.02387.

[6] Fimbel, E., Lauzon, S. in Rainville, C. (2009). Performance of humans vs. exploration algorithms on the Tower of London test. PLoS ONE, 4 (7),e7263.https://pdfs.semanticscholar.org/29f9/e4c7671f7bfd20a487ef9f913bdac53536a8.pdf.

[7] Harsa, P., Břeňová, M., Bezdicek, O. in Michalec, J. (2022). Tower of London test - short version. Neurologia i Neurochirurgia polska, 56(3), 243–250. https://doi.org/10.5603/PJNNS.a2022.0037.

[8] Hernández-Orallo, J. 2017. Vrednotenje v umetni inteligenci: Od usmerjenosti v naloge k merjenju zmožnosti. Artificial Intelligence Review, 48(3), 397–447. DOI: https://doi.org/10.1007/s10462-016-9505-7.

[9] Kaller, C.P., Unterrainer, J.M. in Stahl, C. (2012). Assessing planning ability with the Tower of London task: Psychometric properties of a structurally balanced problem set. Psychological assessment, 24 (1), 46-53. https://doi.org/10.1037/a0025174.

[10] Shallice, T. (1982). Specific impairments of planning. Philosophical Transactions of the Royal Society B, 298(1089), 199–209. https://doi.org/10.1098/rstb.19

[11] Binz, M., & Schulz, E. (2024). Evaluating Planning and Reasoning in Language Models. *Nature Machine Intelligence*. DOI: 10.1038/s42256-024-00896-1

[12] Lake, B. M., Ullman, T. D., & Tenenbaum, J. B. (2024). Symbolic reasoning in the age of deep learning. *Annual Review of Psychology*. DOI: 10.1146/annurev-psych-030322-020111

[13] Xu, Y., et al. (2025). Assessing Executive Function in AI Systems Using Cognitive Benchmarks. *Cognitive Computation*, 17(1). DOI: 10.1007/s12559-025-10200-6

[14] Creswell, A., Shanahan, M., & Kaski, S. (2025). Cognitive Architectures for Multistep Reasoning in LLMs. *Journal of Artificial General Intelligence*. DOI: 10.2478/jagi-2025-0003

[15] Ghosh, A., & Holyoak, K. J. (2025). Analogical Reasoning in Large Language Models: Limits and Potentials. *Cognitive Science*, 49(2). DOI: 10.1111/cogs.13301

Indeks avtorjev / Author index