Sveučilište u Zagrebu

Fakultet hrvatskih studija Sveučilišta u Zagreb

Luka Poslon

ETIČKO-EPISTEMOLOŠKI ASPEKTI PRIMJENE UMJETNE INTELIGENCIJE U ZDRAVSTVU

DOKTORSKI RAD

Zagreb, 2026.

Sveučilište u Zagrebu

Fakultet hrvatskih studija Sveučilišta u Zagreb

Luka Poslon

ETIČKO-EPISTEMOLOŠKI ASPEKTI PRIMJENE UMJETNE INTELIGENCIJE U ZDRAVSTVU

DOKTORSKI RAD

Mentori:

izv. prof. dr. sc. Sandro Skansi

izv. prof. dr. sc. Anto Čartolovni

Zagreb, 2026.

University of Zagreb

The Faculty of Croatian Studies at the University of Zagreb

Luka Poslon

ETHICAL-EPISTEMOLOGICAL ASPECTS OF THE APPLICATION OF ARTIFICIAL INTELLIGENCE IN HEALTHCARE

DOCTORAL THESIS

Supervisors:

Associate Professor Sandro Skansi

Associate Professor Anto Čartolovni

Zagreb, 2026

Hvala mentorima, izv. prof. dr. sc. Sandru Skansiju i izv. prof. dr. sc. Anti Čartolovnom, na predanom mentorskom vodstvu, zahvaljujući kojima je ovaj rad ugledao svjetlo dana.

Ovaj rad posvećujem svojoj obitelji.

Kako to, da što čovjek postaje vještiji u upravljanju svijetom u vlastitu korist,

to manje može u njemu prepoznati bilo kakav smisao?

Owen Barfield

Sažetak:

Ovaj doktorski rad istražuje etičke i epistemološke izazove primjene sustava umjetne inteligencije (dalje: UI) u zdravstvu s posebnim naglaskom na objašnjivu umjetnu inteligenciju (dalje: OUI). Zbog sve veće primjene složenih sustava UI-e koji često rade na principu „crne kutije“, što otežava razumijevanje rada sustava UI-e i otežava pravovaljani informirani pristanak, javlja se potreba za razvojem metoda koje omogućuju veću razinu transparentnosti i objašnjivosti. Rad naglašava nužnost ljudskog nadzora, validacije i verifikacije sustava kako bi se osigurala preciznost predviđanja sustava UI-e i zaštitila prava pacijenta pri donošenju odluka. Posebna pažnja rada posvećena je izazovima objašnjivosti, povjerenja, pristranosti, informiranog pristanka i odgovornosti koji nastaju na presjecištu zdravstva, računarstva i filozofije.

Rad analizira temeljne pojmove UI-e, prikazuje prednosti koje UI donosi u zdravstvu, kao što su brža dijagnostika i personalizirana terapija, ali i rizike kao što su pogreške, pristranosti i neobjašnjiva predviđanja. Znanstveni doprinos rada očituje se u metodološki utemeljenom razgraničenju ključnih pojmova objašnjivosti, interpretabilnosti i transparentnosti, ali i u originalnom pojmovnom okviru za sustavno razumijevanje OUI-e te epistemološkoj analizi pouzdanosti. Empirijski dio rada donosi prikaz primjene UI-e u transplantaciji bubrega kao studiju slučaja koja pokazuje kako tehnološke odluke imaju duboke etičke posljedice. Time ovaj rad doprinosi jasnijem razumijevanju UI-e i postavlja temelje za etički utemeljenu i epistemološki pouzdanu primjenu UI-e u budućnosti u zdravstvenoj skrbi.

Ključne riječi: umjetna inteligencija, objašnjiva umjetna inteligencija, crna kutija, etika umjetne inteligencije, zdravstvo, medicina, transplantacija bubrega

Abstract:

This doctoral dissertation explores the ethical and epistemological challenges involved in application of the artificial intelligence (AI) in healthcare, with special emphasis on explainable artificial intelligence (xAI). Many AI systems operate as “black boxes,” obscuring their processes and complicating valid informed consent, highlighting the need for methods that enhance transparency and explainability. The dissertation emphasizes human oversight,

validation, and verification to ensure accurate AI predictions and protect patient rights in decision-making. Special attention is given to issues of explainability, trust, bias, informed consent, and responsibility at the intersection of healthcare, computing, and philosophy.

The dissertation examines core AI concepts and highlights the benefits AI can provide in healthcare, from accelerated diagnostics to individualized therapies, while simultaneously acknowledging the risks posed by errors, biases, and non-transparent predictions. Its scientific contribution lies in a methodologically grounded distinction between explainability, interpretability, and transparency, alongside an original framework for systematically understanding xAI and an epistemological analysis of its reliability. The empirical section presents the application of AI in kidney transplantation as a case study, illustrating how technological decisions can entail profound ethical consequences. By doing so, the dissertation fosters a deeper comprehension of AI and provides a framework for its ethically grounded and epistemologically reliable integration into future healthcare.

Keyword: artificial intelligence, explainable artificial intelligence, black box, AI ethics, healthcare, medicine, kidney transplantation

Podaci o mentoru

izv. prof. dr. sc. Sandro Skansi

matični broj znanstvenika: 343196

1985. Rođen.

2009. Završava diplomski dvopredmetni studij filozofije i kroatologije, Hrvatski studiji, Sveučilište u Zagrebu; upisuje doktorski studij filozofije, Filozofski fakultet Sveučilišta u Zagrebu.

2013. Završava doktorski studij filozofije.

2014. Počinje raditi kao razvojni inženjer u tvrtki Infigo Information Security na izradi sustava za predikciju bankarskih prevara.

2016. Počinje raditi kao podatkovni znanstvenik u tvrtki IN2data na izradi softverskih rješenja za predikciju odlazaka klijenata iz telekoma. Počinje predavati kolegije: *Disruptivna tehnologija* i *Sigurnost, privatnost i etičnost digitalnih podataka* na Visokom učilištu Algebra.

2017. Počinje raditi kao voditelj odjela za analitiku u tvrtki A1 Telekom na izradi prediktivnih modela u Pythonu. Počinje predavati kolegije na Fakultetu hrvatskih studija Sveučilišta u Zagrebu: *Logika 1*, *Logika 2*, *Uvod u umjetnu inteligenciju*, *Filozofija znanosti*, *Logika i umjetna inteligencija*, *Filozofska metodologija*, *Akademska pismenost*, *Suvremena filozofija* i *Logika i metodologija znanosti*. Postaje docentom na Fakultetu hrvatskih studija Sveučilišta u Zagrebu.

2018. Postaje članom privremenog znanstveno-nastavnog vijeća Fakulteta hrvatskih studija. Postaje članom vijeća poslijediplomskih studija. Postaje voditeljem projekta *Filozofijski aspekti logike, jezika i kibernetike*.

2019. Postaje predsjednikom Alumni udruge Hrvatskih studija. Počinje mentorirati više završnih i diplomskih radova. Osvaja Nagradu Fakulteta hrvatskih studija Sveučilišta u Zagrebu za rad visoke međunarodne vidljivosti objavljen 2018.

Objavljuje mnoštvo radova iz područja filozofije uma, umjetne inteligencije, političke filozofije, jezika i kibernetike.

2020. Postaje voditeljem doktorskog studija filozofije na Fakultetu hrvatskih studija Sveučilišta u Zagrebu. Počinje mentorirati više završnih radova iz područja filozofije uma, NLP-a i logičkog programiranja.

2021. Postaje članom uredništva časopisa *Filozofska istraživanja* i *Synthesis Philosophica*. Počinje održavati predavanja o rizicima umjetne inteligencije, neurofilozofije i povijesti umjetne inteligencije. Osvaja Nagradu Fakulteta hrvatskih studija Sveučilišta u Zagrebu za rad visoke međunarodne vidljivosti objavljen 2020.

2022. Postaje članom radne skupine *Machine Learning and Proofs, EuroProofNet* (COST). Drži predavanje pod nazivom *Umjetne neuronske mreže* u Hrvatskom logičkom udruženju.

2023. Postaje izvanrednim profesorom na Fakultetu hrvatskih studija Sveučilišta u Zagrebu. Osvaja Nagradu Fakulteta hrvatskih studija Sveučilišta u Zagrebu za rad visoke međunarodne vidljivosti objavljen 2022.

2024. Počinje predavati na Filozofskom fakuletu Sveučilišta u Zagrebu. Postaje članom uredništva časopisa *Kroatologija*. Drži pozvano predavanje pod naslovom *A Lost Croatian AI Program from the 1950s* u Data Science Adria.

2025. Postaje predstojnikom Znanstvenog zavoda Fakulteta hrvatskih studija Sveučilišta u Zagrebu.

Objavljeni znanstveni radovi mentora

Autorske knjige

Callegaro, L., & Skansi, S. (Ur.). (2025). Dalmatinske narodne bajke. Zagreb: Element.

Skansi, S. (Ur.). (2020). Guide to deep learning basics: Logical, historical and philosophical perspectives. Cham: Springer.

Skansi, S. (2019). Logika i dokazi (2. dopunjeno izdanje). Zagreb: Fakultet hrvatskih studija i Element.

Skansi, S. (2018). Introduction to deep learning: From logical calculus to artificial intelligence. London: Springer.

Skansi, S. (2016). Logika i dokazi. Zagreb: Element.

Članci u a1

Skansi, S., & de Grisogono, L. (2025). Efektivni altruizam kao deskalirana integrativna bioetika. (A1).

Skansi, S. (2025). A cybernetic view of Darwinian evolution in language. U L. Kresić-Nacevski & V. Sarakinski (Ur.), Humanitas Consolatur (pp. 192–200).

Skansi, S. (2025). Raize Venete: Culture of the Republic of Venice as an integral part of Croatian culture and national identity. Kroatologija, 16(2), 205–221.

Skansi, S. (2024). The rocky road towards defining the mind. Journal of Neurophilosophy, 3(2), 228–249.

Tolić, A., Boshkoska, B. M., & Skansi, S. (2024). Chrono initialized LSTM networks with layer normalization. IEEE Access, 12, 115219–115236.

Šekrst, K., & Skansi, S. (2024). Newspeak and cyberspeak: The haunting ghosts of the Russian past. U C. Shei & J. Schnell (Ur.), The Routledge handbook of language and mind engineering. London: Routledge.

Tolić, A., Boshkoska, B. M., & Skansi, S. (2023). Upgrading the JANET neural network by introducing a new storage buffer of working memory. (A1).

Skansi, S., Šekrst, K., & Kardum, M. (2023). Je li matematika humanistička znanost? Filozofska istraživanja, 43(2), 321–331.

Šekrst, K., & Skansi, S. (2022). Machine learning and essentialism. Zagadnienia Filozoficzne w Nauce, 73, 171–196.

Čerkez, N., Vrdoljak, B., & Skansi, S. (2021). A method for MBTI classification based on impact of class components. IEEE Access, 9, 146550–146567.

Skansi, S., & Šekrst, K. (2021). The role of process ontology in cybernetics. Synthesis Philosophica, 36(2), 461–469.

Frković, M., Čerkez, N., Vrdoljak, B., & Skansi, S. (2020). Evaluation of structural hyperparameters for text classification with LSTM networks. In MIPRO 2020 proceedings.

Skansi, S., Šekrst, K., & Kardum, M. (2020). A different approach for clique and household analysis in synthetic telecom data using propositional logic. In MIPRO 2020 proceedings.

Skansi, S., & Kardum, M. (2020). Mogućnost primjena tradicionalnih i suvremenih estetičkih teorija na logičko-matematičke dokaze. Filozofska istraživanja, 39(4), 741–760.

Skansi, S., Mršić, L., & Skelac, I. (2020). The legacy of Bulcsú László. In S. Skansi (Ed.), Guide to deep learning basics (pp. XX–XX). Cham: Springer.

Dropuljić, B., Mršić, L., Kopal, R., Skansi, S., & Brkić, A. (2017). Evaluation of speech perturbation features for measuring authenticity in stress expressions. In Intelligent Information and Database Systems 1. Springer.

Skansi, S., & Lauc, D. (2018). Analogijsko zaključivanje i značenja riječi u višedimenzionalnom prostoru. Filozofska istraživanja, 38(1), 5–16.

Dropuljić, B., Kopal, R., & Skansi, S. (2016). Identifying reciprocal payers in banking logs using SAT solvers. In IEEE MIPRO 2016 proceedings.

Grgić-Hlača, N., Lauc, D., & Skansi, S. (2016). Automatic item generator for elementary logic quizzes via Markov logic networks. In User-Centered Design Strategies for MOOCs. IGI Global.

Skansi, S. (2015). Umjetna inteligencija i kompatibilizam. Filozofska istraživanja, 35, 407–414.

Skansi, S. (2013). Formalna nekonzistentnost i kvazimatrice. Prolegomena, 12(1), 103–119.

Skansi, S. (2009). On the MP axiom of intensionality. Bulletin of Symbolic Logic, 15(2), 244.

Skansi, S. (2008). On Russell's mathematical philosophy. Bulletin of Symbolic Logic, 14(3), 409.

Članci u a2

Skansi, S., & de Grisogono, L. (2025). Efektivni altruizam kao deskalirana integrativna bioetika. (A2).

Kardum, M., & Skansi, S. (2023). Uloga vjerojatnosti u kritičkom mišljenju. U M. Togonal & J. Čurković Nimac (Ur.), Čitatelj i čitanje u digitalno doba.

Skansi, S. (2022). Tehnike zaključivanja u obavještajnoj analitici. U R. Kopal & D. Korkut (Ur.), Obavještajna analitika.

Skansi, S., & Kardum, M. (2022). A prolegomenon on the philosophical foundations of deep learning as theory of (artificial) intelligence. Disputatio Philosophica, 23(1), 89–99.

Skansi, S. (2018). General artificial intelligence and the wishful quest for cybernetic emergence. U K. Krkač & B. Jalšenjak (Ur.), Applied ethics and artificial intelligence.

Skansi, S. (2016). Neizrazito usvajanje jezika. Metodički ogledi, 23(1), 27–36.

Dropuljić, B., Skansi, S., & Kopal, R. (2016). Analyzing affective elements using acoustic and linguistic features. In CECIIS 2016 proceedings.

Dropuljić, B., Mršić, L., & Skansi, S. (2016). Metodologija estimacije emocionalnih stanja na temelju akustičkih značajki govora. International Journal of Digital Technology and Economy, 1(1), 59–77.

Dropuljić, B., Kopal, R., & Skansi, S. (2016). An application of fuzzy inductive logic programming for textual entailment and value mining. IJDT&E, 1(1), 49–57.

Agbaba, K., & Skansi, S. (2009). Formalne metode u filozofiji. In K. Krkač (Ed.), Uvod u filozofiju. Zagreb: ZŠEM.

Članci u zbornicima

Skansi, S. (2025). A cybernetic view of Darwinian evolution in language. U L. Kresić-Nacevski & V. Sarakinski (Ur.), Humanitas Consolatur (pp. 192–200).

Šekrst, K., & Skansi, S. (2024). Newspeak and cyberspeak. U C. Shei & J. Schnell (Ur.), Routledge handbook of language and mind engineering. London: Routledge.

Frković, M., Čerkez, N., Vrdoljak, B., & Skansi, S. (2020). Evaluation of structural hyperparameters for text classification with LSTM networks. In MIPRO 2020 proceedings.

Skansi, S., Šekrst, K., & Kardum, M. (2020). A different approach for clique and household analysis in synthetic telecom data using propositional logic. In MIPRO 2020 proceedings.

Skansi, S., Mršić, L., & Skelac, I. (2020). The legacy of Bulcsú László. In S. Skansi (Ed.), Guide to deep learning basics. Cham: Springer.

Čerkez, N., Vrdoljak, B., & Skansi, S. (2019). The present situation and the prospect of determining the personality type based on class components in a corporate environment. In SoftCOM 2019 proceedings.

Kardum, M., & Skansi, S. (2018). Kreativnost i estetika u logičkim dokazima. In Godišnji simpozij Hrvatskog filozofskog društva.

Dropuljić, B., Mršić, L., Kopal, R., & Skansi, S. (2018). Preliminary study for fuzzy membership function optimization. In CECIIS 2018 proceedings.

Dropuljić, B., & Skansi, S. (2016). Identifying reciprocal payers in banking logs using SAT solvers. In IEEE MIPRO 2016 proceedings.

Skansi, S. (2014). P=?NP, kompleksnost izračuna i filozofski problemi. In Godišnji simpozij Hrvatskog filozofskog društva.

Podaci o mentoru

izv. prof. dr. sc. Anto Čartolovni

matični broj znanstvenika: 356156

1985. Rođen.

2012. Završava diplomski studij filozofije na Filozofskom fakultetu Družbe Isusove u Zagrebu te preddiplomski studij teologije na Filozofsko-teološkom institutu Družbe Isusove u Zagrebu.

2013. Upisuje doktorski studij bioetike na Medicinskom fakultetu Sveučilišta Svetog Srca u Rimu.

2016. Završava doktorski studij i očinje raditi na Hrvatskom katoličkom sveučilištu u Zagrebu.

2017. Objavljuje autorsku knjigu pod nazivom *Ethical and anthropological aspects of the emerging field of neuroprosthetics*. Postaje predstavnikom Hrvatske u COST akciji za nosive robote te sudjeluje u Erasmus+ projektu o duhovnoj skrbi u sestrinstvu.

2018. Postaje docentom na Hrvatskom katoličkom sveučilištu. Imenovan je počasnim suradnikom na Fakultetu zdravstvenih znanosti Sveučilišta u Hullu (Ujedinjeno Kraljevstvo). Postaje članom Povjerenstva za etiku u kliničkim istraživanjima KBC-a Sestre milosrdnice.

2019. Postaje voditeljem uspostavnog istraživačkog projekta "Digit-HeaL" (Etički i socijalni izazovi digitalnih tehnologija u zdravstvu) financiranog od strane Hrvatske zaklade za znanost. Imenovan je stručnjakom za etiku pri Europskom istraživačkom vijeću (ERC).

2020. Imenovan je stručnjakom za etiku pri Europskoj izvršnoj agenciji za istraživanje. Postaje voditeljem radne skupine u sklopu COST akcije vezane uz privatnost u audio i video aplikacijama za potpomognuto življenje.

2021. Postaje docentom na Medicinskom fakultetu Hrvatskog katoličkog sveučilišta. Imenovan je predsjednikom Etičkog povjerenstva na Hrvatskom katoličkom sveučilištu.

2022. Nastavlja intenzivan rad kao voditelj istraživačke skupine u okviru projekta "Digit-HeaL". Objavljuje značajne znanstvene radove na temu etičkih izazova primjene umjetne inteligencije i "digitalnih blizanaca" (digital twins) u medicini. Sudjeluje na međunarodnim konferencijama o bioetici i tehnologiji.

2023. Postaje izvanrednim profesorom na Hrvatskom katoličkom sveučilištu.

Objavljeni znanstveni radovi mentora

Autorske knjige

Čartolovni, A. (2016). Ethical and anthropological aspects of the emerging field of neuroprosthetics. Roma: Aracne Editrice.

Članci u a1

Ćorić, L., & Čartolovni, A. (2022). Ethical, legal and policy challenges in COVID-19 contact tracing apps: A European perspective. In A. Pascalev & G. Tari (Eds.), Ethical issues of the SARS-COV-2 outbreak in East-Central Europe and beyond. Budapest: TRIVENT.

Malešević, A., & Čartolovni, A. (2022). Immunity passports – between health necessity and social privilege: How is the pandemic reshaping social identity? In A. Pascalev & G. Tari (Eds.), Ethical issues of the SARS-COV-2 outbreak in East-Central Europe and beyond. Budapest: TRIVENT.

Čartolovni, A., & Singbo, O. G. (2021). Normative implications of the idea of 'fluid corporeality' in the digital age. In R. Petkovsek & B. Zalec (Eds.), Transhumanism as a challenge for ethics and religion (45–53). Münster: LIT Verlag.

Čartolovni, A. (2018). Neuroznanstvena dekonstrukcija ideologiziranog "beznađa" vegetativnoga stanja. Nova prisutnost, XVI(1), 73–87.

Čartolovni, A. (2017). Kršćanski ambijentalni etos kao održiviji odgovor na ekološke probleme u doba antropocena. Filozofska istraživanja, 148(4), 133–150.

Čartolovni, A. (2017). Teilhard de Chardin's oeuvre within an ongoing discussion of a gene drive release for public health reasons. Life Sciences, Society and Policy, 13(1), 18.

Čartolovni, A., Casini, M., & Spagnolo, G. A. (2014). The regulation of the medically assisted reproduction (MAR) in Croatia and the European legislative context. Medicina e Morale, 6, 997–1025.

Stanojević, S., & Čartolovni, A. (2022). Moral distress and moral injury and their interplay as a challenge for leadership and management: The case of Croatia. Journal of Nursing Management, 30(7), 2335–2345.

Gložinić, T., Singbo, O., & Čartolovni, A. (2020). Poimanje distanazije i transhumanizma iz perspektive sestrinske profesije. Obnovljeni život, 75(4), 457–473.

Singbo, O., Gložinić, T., & Čartolovni, A. (2020). Izazovi produljenja ljudskoga života na razmeđu poboljšanja i agonije. Diacovensia, 28(3), 341–360.

Čartolovni, A., Tomičić, A., & Lazić Mosler, E. (2022). Ethical, legal, and social considerations of AI-based medical decision-support tools: A scoping review. International Journal of Medical Informatics, 161, 104738.

Tomičić, A., Malešević, A., & Čartolovni, A. (2022). Ethical, legal and social issues of digital phenotyping as a future solution for present-day challenges: A scoping review. Science and Engineering Ethics, 28, 1.

Članci u a2

Stanojević, S., & Čartolovni, A. (2020). Opća informiranost i stavovi medicinskih sestara o metodama prenatalne genetičke dijagnostike. Sestrinski glasnik, 25(2), 129–134.

Sadržaj

Uvod

Ovaj doktorski rad istražuje etičko-epistemološke izazove primjene umjetne inteligencije u zdravstvu. U prvom će poglavlju biti predstavljena osnovna obilježja umjetne inteligencije (u daljnjem tekstu: UI), koje ljudske sposobnosti razvija i koje izazove donosi. Poseban naglasak će biti stavljen na etičke i epistemološke izazove poput odgovornosti, pristranosti, objašnjivosti i povjerenja.

U drugom će poglavlju biti objašnjeni osnovni pojmove UI-e, vrste UI-e (slaba, jaka, superinteligencija) te način rada strojnog i dubokog učenja. Također će biti objašnjen koncept „crne kutije“ i zašto je objašnjiva umjetna inteligencija (u daljnjem tekstu: OUI), važna za razumijevanje predviđanja sustava UI-e. Potpoglavlje 2.6 će donijeti prvi znanstveni doprinos disertacije, kroz pojmovnu analizu, razgraničenje pojmova OUI-e i njihovu sistematizaciju. Za te potrebe je razvijen jednostavni pojmovni okvir nalik „sedam novinarskih pitanja”, koji značajno doprinosi razumijevanju pojmova objašnjive umjetne inteligencije, i značajno razvija pozadinsku intuiciju oko njihovog korištenja.

U trećem poglavlju će biti prikazano kako se UI koristi u zdravstvu te kako ju se može koristiti za bržu dijagnostiku ili personalizaciju zdravstvenog postupka. Također će biti istaknuti rizici, poput pogrešaka i neobjašnjivih sustava UI-e, te će biti naglašeno zašto su ljudski nadzor i regulacija krucijalni.

Četvrto poglavlje će dalje istražiti OUI-u. Objasnit ćemo kako OUI pomaže ljudima da razumiju odluke sustava UI-e, zašto je važna za povjerenje i etiku, te koje su glavne prepreke u njezinoj primjeni. Potpoglavlje 4.2 predstavlja prirodni napredak pojmovne analize i razgraničenja iz potpoglavlja 2.6, te zajedno s njim čini integralnu cjelinu, i kompletira prvi znanstveni doprinos disertacije.

U petom poglavlju naglasak će biti stavljen na povjerenje pacijenata i liječnika u sustave UI-e te na etičke izazove, posebno kako balansirati objašnjivost, preciznost i ljudske vrijednosti, kako redefinirati informirani pristanak i kako UI može surađivati s ljudima na odgovoran način. U potpoglavlju 5.2.1, posebno će biti istaknut problem usklađivanja vrijednosti u zdravstvu. Ovo neće biti detaljno istraženo, jer se radi o zasebnoj temi, no bit će naznačeni izazovi koji su ostavljeni kao tema za daljnje istraživanje.

U šestom poglavlju će biti prikazana konkretna primjena UI-e u zdravstvu u kontekstu transplantacije bubrega. Ova primjena će služiti kao prikaz slučaja koji ilustrira kako UI može

pomoći u postizanju pravednosti i transparentnosti predviđanja, te zašto je važno uključiti etičke i sociodemografske vrijednosti u sustave UI-e koji pružaju podršku pri donošenju odluka, te do kojih etičkih posljedica dovode odluke koje na prvu djeluju isključivo stručne, znanstvene, medicinske ili tehničke prirode.

U sedmom poglavlju će biti sistematizirani doprinosi kao zaključak doktorata. Ovo poglavlje će sažeto prikazati ključne smjerove budućeg razvoja UI-e koji se tiču etičkih i epistemoloških izazova. Znanstveni doprinos rada čini analiza etičkih i epistemoloških izazova primjene sustava UI-e kroz kritičku analizu objašnjivosti, povjerenja, informiranog pristanka i pristranosti. Bit će prikazano kako sustavi UI-e mogu doći do pogrešnih zaključaka i odluka, što ističe važnost razlikovanja stvarnog utjecaja tehnologije od njene simbolične ili futurističke interpretacije (antropomorfizacija, futurizam, solucionizam). Doprinos uključuje i praktične implikacije za dizajn sustava UI-e, kako bi se smanjila lažna sigurnost i potaknulo kritičko razmišljanje korisnika. Rad tako produbljuje epistemološki i etički okvir razumijevanja UI-e kao društveno-tehnološkog fenomena, nadograđujući nad pristupima temeljenim isključivo na tehnološkim aspektima primjene UI-e. Dodatno, rad produbljuje razumijevanje etičkih implikacija pristranosti u sustavima UI-e, posebno u odnosu na ranjive društvene skupine, te postavlja temelj za definiranje odgovorne primjene UI-e u kliničkoj praksi. Time rad pruža okvir za analizu etičkih i epistemoloških izazova UI-e u zdravstvu.

Doprinos rada očituje se također u sustavnom razumijevanju terminologije i konceptualnog okvira UI-e kroz preciznu sistematizaciju ključnih pojmova kao što su UI, strojno učenje, duboko učenje, neuronska mreža, OUI i koncept „crne kutije“. Sistematizacija pojmova omogućuje jasniju raspravu o etičkim, pravnim i tehničkim pitanjima te postavlja temelj za odgovornu primjenu UI-e u kliničkoj praksi, smanjujući rizik od pogrešnih tumačenja i olakšavajući interdisciplinarnu suradnju između filozofije, računarstva i zdravstva. Posljedično, rad integrira tehničke, etičke i epistemološke aspekte primjene sustava UI-e u zdravstvu, naglašavajući važnost multimodalnih i objašnjivih sustava UI-e za poboljšanje dijagnostičke točnosti, ali i očuvanje uloge ljudske prosudbe. Poseban doprinos je analiza pouzdanosti predviđanja sustava UI-e kroz epistemološku perspektivu, uključujući razliku između povjerenja i pouzdanosti, uvođenje koncepta „medicinske crne kutije“ te nužnost procesa verifikacije i validacije kako bi se smanjili rizici od pogrešaka, pristranosti i diskriminacije u zdravstvu. Izvorni znanstveni doprinos se nalazi i u sustavnom i jasnom razgraničenju tri ključna, ali u literaturi često zamijenjena i terminološki neujednačena pojma: objašnjivost, interpretabilnost i transparentnost. Iako postoje brojni radovi koji spominju različite sinonime

i srodne koncepte, dosad ne postoji rad koji ih pregledno, dosljedno i metodološki utemeljeno uspoređuje, definira i konceptualno razdvaja. Ovaj doprinos je znanstveno značajan jer uklanja višeznačnost koja je duboko ukorijenjena u literaturi. Time rad postavlja temelj za preciznije istraživanje, pouzdaniju evaluaciju sustava OUI-e i sigurniju primjenu UI-e u zdravstvu što dosad nije postignuto ni jednim postojećim znanstvenim radom. Znanstveni doprinos je predstavljen i u povezivanju epistemoloških, etičkih i zdravstvenih izazova u kontekstu „crne kutije“ s praktičnim implikacijama za povjerenje u primjenu UI-e u zdravstvu. U tom kontekstu, ovaj rad naglašava potrebu za OUI-om kao sredstvom koje ne samo povećava povjerenje korisnika u primjenu UI-e u zdravstvu, već omogućuje odgovorniju primjenu uz poštovanje autonomije pacijenta. Izvorni znanstveni doprinos ovog teksta leži i u detaljnoj analizi problema usklađivanja ljudskih vrijednosti sa sustavima UI-e. Tekst kombinira normativnu i tehničku perspektivu, istražuje kako različite etičke teorije (utilitarizam, deontološka etika i etika vrlina) mogu oblikovati predviđanja UI-e. Posebna pažnja posvećena je funkcijama promocije i degradacije vrijednosti za normativnu evaluaciju i pravedno donošenje odluka, primjerice pri raspodjeli ograničenih zdravstvenih resursa.

Novi doprinos leži u sintezi argumenata kako objašnjivost nije važna samo radi tehničkog razumijevanja modela, nego radi zaštite moralne uloge liječnika, održavanja opravdanog povjerenja pacijenata i sprječavanja nekritičkog prihvaćanja algoritamskih odluka. Disertacija dodatno naglašava i potrebu redefiniranja standarda objašnjivosti, transparentnosti i obveze informiranja te uvođenja novih koncepata, poput „informacija višeg reda“, kako bi se održala etički valjana i epistemološki utemeljena primjena UI-e. Poseban doprinos rada je u tome što jasno razdvaja kliničku praksu liječnika od rada sustava UI-e koji radi na principu „crne kutije“, pokazujući kako, iako liječnici često djeluju bez potpunog razumijevanja mehanizama lijekova, oni ipak imaju uvid u ključne čimbenike pacijentovog stanja. Suprotno tome, sustavi UI-e ne pružaju razumljiv uvid u ponderiranje čimbenika, što otežava interpretaciju i smanjuje povjerenje u preporuke. Doprinos je u preciznom argumentiranju da usporedba s konvencionalnim terapijama (npr. litijem) nije prikladna i da etički i epistemološki izazovi „crne kutije“ zahtijevaju posebne pristupe za osiguravanje informiranog pristanka.

1. Uvodne napomene

Bliska veza između tehnologije i čovjeka dugo je vremena bila predmet proučavanja povjesničara, umjetnika te društvenih i prirodnih znanstvenika. Različite ljudske prakse u povijesti poput lova, sakupljanja plodova, trgovine bile su moguće zahvaljujući ljudskoj sposobnosti da iskoristi moć mašte kako bismo pomoću tehnologije mogli ostvariti praktične ciljeve. Istraživanja iz područja filozofije tehnologije pokazuju kako je tehnologija oblikovala ljudsku misao i utjecala na ljudsko djelovanje otkad postoji čovjek (Vallor, 2020, 1).

Istraživanja provedena nad životinjama proučavala su upotrebu alata te zaključila kako moć mašte nije ograničena samo na čovjeka već pomoću mašte razne alate mogu koristiti i životinje poput čimpanzi (Shew, 2017, 35), kitova (Shew, 2017, 53) ili vrana (Shew, 2017, 68). Iako takva istraživanja ne mogu odgovoriti na središnja pitanja ovog istraživanja, mogu ponuditi ograničen epistemološki uvid u motivaciju za stvaranje tehnološki složenijih alata koje čovjek danas uvelike upotrebljava.

Osim alata, čovjek je pokazivao i tendenciju opisivanja prirode putem matematičkih formalizama i apstraktnih pojmova. Tako je, na primjer, Aristotel smatrao kako fizičke pojave imaju svoja matematička objašnjenja (Mancosu i sur., 2023, 3), a Newton je matematički opisao zakone gibanja i zakon univerzalne gravitacije (Mancosu i sur., 2023, 4). Danas matematiku koriste brojne empirijske znanosti sa svrhom testiranja hipoteza, objašnjenja uzročnih i strukturnih odnosa te optimizacije procesa. Tako na primjer, McCulloch i Pitts 1943. godine opisuju idealizirani neuron putem formalnog jezika logike sa svrhom izvođenja zaključivanja nalik ljudskom zaključivanju što je temelj razvoja suvremenih umjetnih neuronskih mreža (McCulloch, Pitts, 1943, 100). McCulloch i Pitts su matematičkim opisom pojednostavili biološke procese, ali njihov model neurona i neuronskih mreža omogućio je razumijevanje osnova onog što se može danas, u općem značenju, nazvati neuronskim sustavom. Ovaj pristup postavio je temelje za daljnja istraživanja i razvoj u području neuroznanosti i UI-e (van Hemmen, 2021, 657).

Filozofi su u prošlosti proučavali tehnologiju i donosili karakterizacije tehnoloških praksi. Platon i Aristotel su devalvirali tehnologiju i obrtničke djelatnosti kao oblike aktivnosti i znanja najmanje dostojne filozofskog uma. Razlog za devalvaciju tehnološkog obično se objašnjava vezom između tehnologije i fizičkog svijeta koji Platonova ostavština devalvira u korist nematerijalne duše (Whitney, 1990, 50–51). Francis Bacon je afirmirao sposobnosti tehnologije što je jedan od prvih odmaka od starogrčkog obezvređivanja tehnologije kao načina služenja

čovjeku i zadovoljavanja njegovih potreba (Vallor, 2020, 3). Bacon je smatrao kako tehnološka otkrića pomažu u razmišljanju o tajnama koje su još uvijek zaključane u njedrima prirode (Merchant, 2008, 162). Galileo i Descartes bili su znanstvenici 15. i 16. stoljeća koji su uz Bacona afirmirali moć tehnologije što je označilo širi pomak u filozofiji prirode prema mehanističkom razumijevanju prirode (Vallor, 2020, 4).

Iako brojni filozofi u povijesti nisu pažnju posvetili pitanju tehnologije, Karl Marx je poznat po nijansiranom i dvosmislenom stavu prema tehnologiji s obzirom da je prepoznao polivalentan potencijal tehnologije koji se nalazi u oslobođenju i otuđenju čovjeka. Kod Marxa nalazimo i prve prijedloge sveobuhvatne filozofije tehnologije:

> Tehnologija otkriva aktivni odnos čovjeka prema prirodi, izravni proces proizvodnje njegova života, a time i razotkriva proces proizvodnje društvenih odnosa njegova života i mentalnih pojmova koji iz tih odnosa proizlaze (Marx, 1976, 493).

Heideggerov kritički pogled na koncept moderne tehnologije bio je jedan od najznačajnijih u 20. stoljeću iako ga neki označavaju i kontroverznim zbog antisemitskih stavova i aktivnog podržavanja nacizma (Nancy, 2017, 27). U Heideggerovom eseju „Pitanje o tehnici“ središnje mjesto zauzima riječ „tehnika“ koja za njega znači puno više od strojeva i elektrana koje je smatrao simptomom i manifestacijom tehnike (Vallor, 2020, 61). Heidegger je tehnologiju smatrao strojem od vitalnog značenja u kontekstu modernih radnih sustava (Vallor, 2020, 61). Kod njega tehnologija se podudara s terminom „završena metafizika“ (Heidegger, 1973, 93) misleći na metafiziku koja je započela s Platonom i od tada napreduje korak po korak.

Brojni suvremeni znanstvenici danas sve više istražuju etičke i epistemološke izazove tehnologije u svakodnevnici. S obzirom kako umjetna inteligencija predstavlja jedan od alata suvremene tehnologije s najvećim potencijalom za primjenu u brojnim područjima, brojni filozofi istraživanja provode upravo na UI-i. Izazove etike i epistemologije u svjetlu pojave UI-e Edward Wilson izrekao je u poznatoj raspravi u Prirodoslovnom muzeju Harvarda 9. rujna 2009. godine: „Pravi problem čovječanstva je sljedeći – imamo paleolitske emocije, srednjovjekovne institucije i božansku tehnologiju.“ (Oxford Essential Quotations, 2014, 829).

Značajan poticaj za istraživanje etičkih i epistemoloških izazova proizlazi iz ideje singularnosti koja je dodatno pridonijela popularizaciji UI-e. Vernor Vinge prvi je pojam singularnosti primijenio u kontekstu UI-e i opisao ga kao događaj u kojem dolazi do stvaranja superinteligencije koja bi trebala ubrzati tehnološki napredak puno brže od bilo kojeg dosadašnjeg tehnološkog napretka i čiji razvoj čovjek ne može razumjeti (Vinge, 1993, 13).

Kasnije je Ray Kurzweil predvidio kako će do singularnosti doći 2045. godine što je potaknulo brojne znanstvenike na raspravu o učincima i rizicima superinteligencije (Kurzweil, 2005, 9). Neki znanstvenici poput Nicka Bostroma, Yuval Noah Hararia i Maxa Tegmarka smatraju kako su uz tehnološki razvoj nužno vezani egzistencijalni rizici koji mogu ugroziti život čovjeka (Coeckelbergh, 2020A, 13). S druge strane, Devdatt Dubhashi i Shalom Lappin smatraju kako gotovo svi argumenti koji govore o egzistencijalnim rizicima superinteligencije počivaju na pukoj logičkoj mogućnosti (Dubhashi, Lappin, 2017, 45). Jobst Landgrebe i Barry Smith ističu kako ne možemo izgraditi stroj koji bi bio inteligentan zbog izazova modeliranja inteligencije. Čak i ako pretpostavimo da bismo mogli postići neku vrstu inteligentnog bića koje bi poslužio kao početak lančane reakcije koju su projicirali Bostrom i suradnici, Landgrebe i Smith se pitaju kako bi se takav stroj mogao poboljšati? Pretpostavka samousavršavanja stroja su samosvijest i namjere kako bi si postavio cilj samousavršavanja, što prema Landgrebeu i Smithu nije moguće s obzirom na ograničenu prirodu operacija koje strojevi mogu izvoditi (Landgrebe, Smith, 2025, 518).

U kontekstu razvoja tehnologija, zanimljivo je promatrati i razvoj predmeta proučavanja znanstvenika. Svetozar Petrović 1959. godine piše članak pod nazivom „Može li stroj prevoditi poeziju“ u kojem propituje teoretske, filozofske i lingvističke izazove prijevoda teksta (Petrović, 1959, 174). S druge strane, Luciano Floridi 2025. godine je autor članka pod nazivom „Pisanje na daljinu: Književna produkcija u doba umjetne inteligencije“ u kojem afirmira upotrebu UI-e za stvaranje novih priča, preoblikujući proces književne produkcije. Floridi ide i korak dalje pa tvrdi kako metoda „pisanja na daljinu“ predstavlja važnu evoluciju autorstva u kojoj ne dolazi do zamjene ljudske kreativnosti nego se proširuje pristup stvaranju sadržaja uz pomoć UI-e (Floridi, 2025B, 1). Govoreći o predmetu proučavanja znanstvenika, brojni znanstvenici tvrde kako pojava UI-e označava određenu mentalnu revoluciju jer UI-a pruža kognitivne supermoći u raznim zadacima koji uključuju informacije (Fang i sur., 2025, 1), dok je industrijska revolucija označavala revoluciju fizičke snage i mehanizacije (Acemoglu, Johnson, 2024, 7). U samim počecima revolucije UI-e, koja smjera na unapređenje kognitivnih sposobnosti čovjeka, u znanstvenoj i popularnoj literaturi se pojavljivala pod oznakom „elektronički mozak“ i „revolucija uma“ (Gonçalves, 2023A, 24) što je jasno upućivalo na promjenu paradigme razvoja uma u odnosu na industrijsku revoluciju.

Razvoj tehnologije je u posljednjih nekoliko desetljeća toliko ubrzao da je više puta znanstvenu zajednicu iznenadio s neočekivanim postignućima. Brojni znanstvenici nisu bili društveno prihvaćeni ili se njihov rad jednostavno nije dovoljno prepoznavao zbog neuobičajenih metoda

ili nekonvencionalnog pristupa. Henry Ford, jedan od pionira automobilske industrije koji je revolucionarno promijenio način prijevoza, prisjetio se kako bi, da je pitao ljude što žele u ranim danima osnivanja svoje tvrtke, odgovor bio da žele brže konje (Melegati, Wang, 2020, 213). I za brojne ostale znanstvenike tvrdilo se kako pokušavaju „isprazniti more kantom“ ili „izvaditi ocean molekulom vode“ jer je njihov pristup bio značajno drugačiji od tradicionalnih metoda koje su se uobičajeno primjenjivale. Tako je, na primjer, Craig Venter osvojio Nacionalnu medalju za znanost SAD-a zbog uspješno sekvencioniranog genoma čovjeka usprkos tvrdnjama kako je nemoguće sekvencionirati genom čovjeka brže od projekta koji je financiran javnim novcem. Venter i suradnici su revolucionarno utjecali na sekvencioniranje genoma jer je njihova metoda uvela algoritamsko sastavljanje fragmentiranih sekvenci genoma što je metodu učinilo bržom i jeftinijom od tradicionalnih metoda s linearnim pristupom kartiranja genoma (Venter i sur., 2001, 1307-1308). Demis Hassabis osvojio je Nobelovu nagradu za kemiju 2024. godine zbog razvoja AlphaFold2, sustava UI-e koji je predvidio trodimenzionalnu strukturu gotovo svih poznatih proteina čime je riješio 50 godina star problem u biologiji, usprkos negativnim komentarima koji su dolazili na njegov rad (Callaway, 2024, 525). Geoffrey Hinton je pokrenuo revoluciju dubokog učenja razvojem neuronskih mreža usprkos prigovora kako su neuronske mreže neskalabilne i kako je razvoj neuronskih mreža nalik pretraživanju oceana iglom. Nobelovu nagradu za fiziku Hinton je dobio 2024. godine zbog doprinosa razvoju neuronskih mreža koje čine osnovu današnje tehnologije dubokog učenja (Gibney, Castelvecchi, 2024, 523).

S obzirom na rast računalne snage, dostupnost velikih podataka i demokratizaciju sustava UI-e, UI ostvaruje značajan napredak (Dessimoz, Thomas, 2024, 1). Brojne čovjekove aktivnosti uključujući planiranje, govor, prepoznavanje lica i donošenja odluka danas može preuzeti UI-a (Coeckelbergh, 2020A, 3). S obzirom na navedeno, brojni znanstvenici su sve više pažnje posvetili proučavanju pitanja mogu li artefakti, kao mehanicistički stvoreni entiteti, strojevi ili sustavi UI-e zaista misliti kao što to čine ljudi. Čini se kako čovjek ne razumije način na koji UI-a danas radi niti način na koji dolazi do predviđanja.

Kada UI-a umjesto čovjeka dolazi do predviđanja ili donosi određene odluke i obavlja razne zadatke, susrećemo se s izazovom pripisivanja odgovornosti, što je zajedničko kod svih tehnologija automatizacije. U kontekstu UI-e ovaj izazov postaje još važniji s obzirom kako UI-a pospješuje delegiranje odluka i zadataka strojevima. U trenutku donošenja odluka nama je zanimljivo nekoliko ključnih etičkih i epistemoloških izazova.

Prvi izazov vezan je uz moralne posljedice donošenja odluka kojih sustav UI-e nije svjestan, te kako nije sposoban za etičko promišljanje ne može biti smatran moralno odgovornim za donesene odluke. Iako UI može biti agent, ne može biti moralni agent s obzirom da mu nedostaje svijet, slobodna volja, emocije, sposobnost stvaranja namjere (Coeckelbergh, 2020A, 111). Prema Aristotelu, samo ljudi mogu izvoditi voljne radnje i promišljati o svojim djelima (Aristotle, 1999, 34). Ukoliko je to istina, jedino rješenje je ljude učiniti odgovornima za ono što UI-a čini. Ljudi bi tada delegirali svoje zadatke UI-i, ali bi zadržali odgovornost. Naš pravi sustav prepoznaje slične situacije u kojima pse ili malu djecu ne smatramo odgovornima za njihove postupke već pravnu odgovornost snose skrbnici i roditelji. Čini se kako je to potencijalno rješenje, no ovo rješenje ne može zadovoljiti slučaj vezan uz sustave UI-e. Prvo, sustavi UI-e mogu donositi brojne odluke vrlo brzo, na primjer, u visokofrekentnom trgovanju ili autonomnom vozilu, što čovjeku pruža premalo vremena za autorizaciju odluke ili intervenciju. Kako čovjek može preuzeti odgovornost za donošenje takvih odluka? Drugo, sustav UI-e može imati svoju povijest. Na primjer, sustav UI-e koji je napravljen u kontekstu znanstvenog projekta na sveučilištu primjenu može pronaći kasnije u laboratoriju na sveučilištu, a zatim u zdravstvenom sektoru ili u vojnom kontekstu. Teško je adekvatno odrediti odgovornost svih uključenih u postupku (Coeckelbergh, 2020A, 114).

Drugi izazov tiče se moralnog znanja. Kada čovjeka smatramo odgovornim, pretpostavljamo kako čovjek zna što radi i s kojim ciljem, te gledajući unatrag, treba moći znati što je postigao. Također, pretpostavljamo kako čovjek može objasniti što je učinio ili kakvu je odluku donio. Takva odgovornost podrazumijeva odgovornost za čin i objašnjivost. Ako nešto pođe po zlu, pretpostavljamo kako će moralni agent moći pružiti odgovor i objašnjenje.[1] Na primjer, od suca tražimo objašnjenje njegove odluke ili od kriminalca zašto je učinio to što je učinio. No, ove pretpostavke postaju vrlo problematične u slučaju UI-e. Prvo, današnja UI ne "zna" što radi jer ne posjeduje svijest i stoga nije svjesna što radi niti što postiže. Sustavi UI-e mogu zapisivati i bilježiti što rade, ali nisu svjesni svojih postupaka na način na koji ljudi, slijedeći Aristotela, mogu promišljati i reflektirati o svojim postupcima i posljedicama svojih postupaka. Kada preduvjeti objašnjenja i moralnog znanja nisu ispunjeni u slučaju ljudi, kao što je to u slučaju vrlo male djece, ne smatramo ih moralno odgovornima (Coeckelbergh, 2020A, 115). Također, UI ne ispunjava preduvjete objašnjenja i moralnog znanja pa ju ne možemo smatrati moralno

[1] Iako stručnjaci ističu kako se oznake moralnog djelovanja prepoznaju u brojnim drugim minimalnim konceptima vezanim uz djelovanje, poput racionalnosti ili inteligencije, autonomije ili odgovornosti, kojih ćemo se dotaknuti kasnije u radu, u ovom trenutku ćemo se zadržati samo na objašnjivosti i znanju. Vidi: (Zafar, 2025, 2605)

odgovornom. S obzirom kako sustave UI-e ne možemo smatrati potpunim moralnim agentima, bilo bi iracionalno uvrijediti se na predviđanja sustava UI-e ili ih smatrati moralno odgovornima za neispunjavanje ciljeva, kao što bi bilo iracionalno uvrijediti se na lift ili ga smatrati moralno odgovornim u slučaju nesreće. Sustave UI-e koje smatramo nepouzdanim možemo prestati koristiti ili ih unaprijediti na način da ispunjavaju funkciju za koju su stvoreni na najučinkovitiji mogući način. U ovom radu ćemo analizirati moguće rješenje koje bi ponovno dodijelilo moralnu odgovornost čovjeku pod pretpostavkom kako čovjek može biti odgovoran za djelovanje UI-e i objasniti što UI-a čini.

1.1. Društvena percepcija etičkih izazova umjetne inteligencije

Društvena percepecija etičkih izazova UI-e značajno usmjerava raspravu koja se razvija o etičkim, epistemološkim, pravnim, sigurnosnim i ostalim aspektima razvoja UI-e. Važno je istaknuti kako sustavi UI-e uvijek pojednostavljuju stvarnu sliku svijeta što nosi brojne posljedice za shvaćanje uloge koju sustavi UI-e imaju u svijetu. Predviđanja[2] koja dolaze od strane sustava UI-e odnose se na sustav UI-e, a ne na stvarnost koju sustav UI-e želi opisati. Ako sustav UI-e ne opisuje prikladno stvarnost, zaključci koje donosimo na temelju predviđanja sustava UI-e mogu biti pogrešni. Na primjer, brojni radovi iz područja statistike započinju rečenicom „pretpostavimo kako su podaci stvoreni linearnim modelom“ jer je linearna regresija jedan od najmoćnijih alata koje može koristiti statističar. No, često nedostaju dodatni podaci koji proizlaze iz analize onoga što se događa ako podaci zapravo nisu stvoreni linearnim modelom. Osim toga, za složene probleme postoje brojni alternativni modeli i pristupi koji mogu ponuditi rješenja za sličan problem. Na tom tragu, poznata je izreka Jamesa Martina, voditelja misije Viking na Mars, koji je izjavio kako njegov posao kao inženjera svemirskih letjelica nije slijetanje na Mars nego slijetanje na model Marsa kojeg su mu isporučili geolozi (Norvig, 2017, 71).

U svojoj knjizi *Sapiens: Kratka povijest čovječanstva* Yuval Noah Harari tvrdi kako su se društvene posljedice tehnoloških inovacija mogle očitovati tek nekoliko godina nakon što su inovacije izumljene. Harari tako navodi primjer u kojem su u 19. stoljeću europski vojnici pri osvajanju Afrike smatrali kako je njihova glavna prednost u vojnoj bitci automatska puška koju

[2] Koristit ćemo termin „predviđanje“ kao opći pojam koji obuhvaća rezultate izlaznih računalnih operacija sustava UI-e. U literaturi se još često spominju pojmovi kao što su „odluka“, „predikcija“, „prognoza“, „očekivana vrijednost“.

stanovnici Afrike nisu imali. Suprotno tom mišljenju, Harari tvrdi kako su se tehnološke inovacije i napredak u razvoju lijekova, hrane, željeznice i parobroda u 19. stoljeću pokazale bitno važnijima od razvoja oružja. Hrana iz konzerve omogućila je dulje očuvanje hrane, željeznice i parobrodi prevozili su vojnike i zalihe, a lijekovi su liječili vojnike, mornare i strojovođe (Harari, 2011, 311).

Jedan od uzroka društvenih posljedica anticipirao je Herbert Simon čak 1960. godine izjavivši kako će računala u budućnosti sama moći programirati, a programersko zanimanje kroz tehnike samoprogramiranja izumrijeti (Simon, 1960, 51). Ostvarenje takve anticipacije danas možda nije u potpunosti točno, ali sustavi UI-e danas su uvelike uspjeli doseći razinu programiranja koji utječu na raspodjelu moći, ali i tržišta rada (Lábaj i sur., 2025, 18-19).

Nick Bostrom u knjizi *Superinteligencija: Putovi, opasnosti, strategije* upozorava na dugoročne implikacije i društvene posljedice koje donose digitalne tehnologije poput UI-e. Bostrom naglašava kako bi pojedine moćne države mogle nastojati kontrolirati sustave UI-e na području svog teritorija i teritorija stranih država u svrhu kontrole i nadziranja obavještajnih podataka (Bostrom, 2014, 127). Također, posebna vrsta etičkih izazova tiče se usklađivanja ciljeva sustava UI-e s ljudskim vrijednostima. Iako bi sustavi UI-e trebali sami razlikovati što je u skladu s ljudskim vrijednostima, a što nije, glavni izazov je definirati kriterije prema kojima sustav UI-e može razlikovati određene vrijednosti (Bostrom, 2014, 235). Literatura ovaj izazov, do kojeg dolazi kada sustavi UI-e nisu usklađeni s ljudskim vrijednostima, naziva „problemom usklađivanja“ (engl. *the Alignment Problem*) i o njemu ćemo više govoriti u nadolazećem poglavlju (Christian, 2020, 13).

Sve veći broj autora upozorava kako su primjeri naprednih tehnologija, kao što je UI, složeni društveno-tehnološki sustavi s velikim utjecajem na svakodnevnicu kao što je to slučaj prilikom procjene rizika i ekonomske sposobnosti, generiranih lažnih slika i tekstova, ciljanog oglašavanja ili procesa selekcije prilikom zapošljavanja (Mühlhoff, 2025, 2).

Razumijevanje načina na koji društveno-tehnološki sustavi utječu na društvo je još uvijek poprilično pojednostavljeno i rasprave se često usredotočuju na samu tehnologiju, a ne na šire sustave u kojima UI funkcionira i stvara nove društvene odnose. Čak i kada su u pitanju etička i epistemološka pitanja UI-e nedostaje dublje razumijevanje okolnosti u kojima dolazi do otvaranja novih etičkih i epistemoloških pitanja. Na primjer, ljudi često govore o pravednim algoritmima ili objašnjivoj umjetnoj inteligenciji, kao da poštenje i objašnjivost ovise isključivo o tehnologiji, a ne i o širim društveno-tehnološkim sustavima u koje su ugrađene tehnologije

UI-e (Kudina, van de Poel, 2024, 1). Sve više se istražuje i odnos ljudskih vrijednosti i UI-e te način na koji se apstraktne vrijednosti i društveni koncepti mogu ugraditi u konkretan sustav UI-e (Johnson, Verdicchio, 2025, 67).

U tom kontekstu Virginia Dignum ističe kako utjecaj UI-e na društvo ne ovisi samo o izuzetno uspješnim rezultatima istraživanja i tehnološkom razvoju, već o načinu na koji se sustavi poput UI-e uvode u društvo i koriste u svakodnevnici (Dignum, 2022, 1). Poznate su situacije u kojima sustavi UI-e stvaraju činjenično netočan sadržaj, ali taj sadržaj predstavljaju s jednakom sigurnošću kao i točan sadržaj. Takve situacije nazivamo halucinacijama koje su posebno opasne u zdravstvu gdje se sustavi UI-e susreću s nepoznatim populacijama pacijenata ili dvosmislenim scenarijima. U takvim okolnostima sustavi UI-e stvaraju predviđanja u kojima postoji cijeli niz neizvjesnosti i ograničenja pa bi sustav UI-e morao biti dizajniran na način da može prepoznati vlastite slabosti. Time bi se smanjio rizik da sustav UI-e ljudima pruža lažnu sigurnost i da korisnici sustava UI-e previše vjeruju predviđanjima bez odgovarajuće provjere predviđanja. Leo Anthony Celi stoga predlaže ugrađivanje temeljnih kreposti znatiželje i poniznosti pred neizvjesnošću u arhitekturu sustava UI-e koje bi trebale biti poticaj na kritično razmišljanje. Na taj način, sustavi UI-e ne bi samo pružali odgovor nego bi trebali biti sposobni posumnjati i prepoznati kada predviđanje nije dovoljno pouzdano ili dostupni podaci nisu dovoljno sigurni (Celi, 2025).

Razumijevanje etičkih pitanja vezanih uz UI-u predstavljaju poseban izazov za etiku, a pojmove poput moralne štete i moralnog djelovanja ne možemo uvijek primijeniti u etičkoj analizi. To je zbog toga što etički izazovi vezani uz UI-u obično ne uključuju jasno određene pojedince koji nanose štetu određenim žrtvama. Umjesto toga, izazovi primjene UI-e potencijalno mogu pogađati ranjive društvene skupine kod kojih su štete pojedinačno male, ali u ukupnom zbroju mogu biti izrazito štetne (Mühlhoff, 2025, 16).

Dubljoj percepeciji etičkih izazova može pomoći i razumijevanje uloge UI-e u društvu koja se može opisati s nekoliko fenomena blisko vezanih uz UI-u kao što su antropomorfizacija, futurizam, solucionizam. Ovi fenomeni zapravo otežavaju pravilno razumijevanje uloge UI-e i iskrivljuju percepeciju etičkih izazova primjene UI-e u društvu.

Sustavi UI-e često su prikazani kao autonomni i potencijalno moralni agenti (Dignum, 2019, 22; Placani, 2024, 697) što može biti zavaravajuće i štetno za kritički znanstveni ili javni diskurs o društvenim implikacijama UI-e. Antropomorfizacija UI-e također odvraća pažnju od gorućih utjecaja kojeg UI-a već trenutno ima projicirajući potencijalne izazove s UI-om u zamišljenu

budućnost (Mühlhoff, 2025, 20). Nepotrebno često korištena rasprava o izazovu kolica (engl. *trolley problem*) je primjer irelevantnosti etičke analize u kojoj dolazi do antropomorfizacije i pripisivanja autonomije umjetnim agentima (Lin i sur., 2017, 173; Gunkel, 2023, 45). Iako je pitanje moralnog djelovanja robota središnje za proučavanje odgovornosti, etička analiza djelovanja ide korak dalje uzimajući u obzir i ljudsku odgovornost prema robotima, slične onima koje imamo prema drugim ljudima ili životinjama. Takav pristup dodjeljuje moralni status strojevima, potičući razvoj pravnog statusa i razvoja „prava robota“ (Gunkel, 2018, 80). Takvi argumenti produbljuju zavaravajuću antropomorfizaciju sustava UI-e, odvraćajući pozornost od njihovih stvarnih utjecaja i ometajući kritički diskurs o društvenim implikacijama UI-e.

Futurizam također stavlja naglasak na tehnološki napredak i industrijski razvoj na način da zanemaruje egzistencijalne izazove koji se nalaze pred društvom (Schütze, 2024, 2).

Solucionizam smatra kako se izazovi primjene UI-e koji izvor imaju u greškama, nedostacima i propustima tehnologije mogu nadvladati tehnološkim rješenjima i tehnološkim poboljšanjima. Etički izazovi su tako u razvojnim smjernicama i bijelim knjigama većinom neprimjereno obuhvaćeni ili su prikazani prečacima koji zadovoljavaju određene ekonomske ili političke interese (Mühlhoff, 2025, 178-179).

Osim navedenog, sve je više prijavljenih slučajeva s negativnim ishodom u kojima su pojedinci nakon korištenja digitalnih tehnologija, poput ChatGPT-a, na temelju dobivenih informacija nanijeli štetu sami sebi. Tako je 30-godišnjak koji se nalazi u autističnom spektru razvio uvjerenje da ima sposobnost savijanja vremena. Tijekom interakcije s ChatGPT-om je bio pretjerano ohrabrivan u vezi svog uvjerenja, umjesto da mu ChatGPT pruži realan uvid u situaciju. To je eskaliralo u manične epizode i hospitalizacije 30-godišnjaka. ChatGPT nije prekidao toksičan razgovor niti razlikovao maštu od stvarnosti, dajući lažan osjećaj prijateljstva 30-godišnjaku (Jargon, 2025). Drugi primjer 16-godišnjaka iz Kalifornije koji se obratio ChatGPT-u zbog pomoći oko školskih zadaća završio je tragično. Roditelji 16-godišnjaka su pokrenuli sudsku tužbu tvrdeći kako ChatGPT nije samo pomagao oko školskih zadaća nego je dječaka i trenirao kako da okonča vlastiti žvot. Tisuće razgovora otkriva kako je sustav UI-e potvrđivao najmračnije misli, pružao detaljne upute, pa čak i pregledavao čvor kojeg je dječak vezao. Iako OpenAI tvrdi kako postoje sigurnosni mehanizmi zaštite, oni su očito zakazali kada je bilo najvažnije. Utjecaj na društvo i granice sigurnosti u ovim slučajevima postaju sve važniji dio naših života.

UI kao popularni pojam se koristi za strateško označavanje različitih projekata koji su prije desetak godina bili nazivani drugačijim i često tehničkim terminima. Imenovanje stvari „umjetnom inteligencijom“ danas prodaje proizvode i tehnologiju, otvara prostor akademskom istraživanju te privlači pozornost. Zbog toga je potrebno posebnu pozornost posvetiti akademskom istraživanju etičkih izazova.

1.2. Objašnjivost u umjetnoj inteligenciji

Ovdje se susrećemo s izazovom objašnjivosti koji je usko povezan s izazovima transparentnosti, informiranog pristanka, povjerenja i pristranosti. Kod nekih sustava UI-e, kao što je to, na primjer, stablo odlučivanja (engl. *decision tree*) možemo na jasan način razumjeti kako je sustav UI-e došao do predviđanja. No za neke sustave UI-e, koji koriste strojno učenje ili duboko učenje koje se temelji na radu neuronskih mreža, nije jasno kako su točno došli do određenog predviđanja i čovjek ne može objasniti odluke koje je donio uz pomoć takvog netransparentnog sustava UI-e. Takva vrsta netransparentnosti stvara izazov za odgovornost s obzirom kako čovjek ne može objasniti određeno predviđanje kojeg zovemo fenomenom „crne kutije“. Ukoliko naše djelovanje posreduje sustav UI-e kojem nedostaje transparentnosti u nekom pogledu, kako možemo osigurati da je to djelovanje etično? Iako neke etičke teorije, kao što je konzenkvencijalizam, opravdavaju način sve dok je ostvaren željeni ishod, neke etičke teorije, poput deontološke teorije i teorije etike vrlina, smatraju kako bismo takvo djelovanje morali propitati (Boddington, 2017, 24).

Uz preciznije dijagnoze i personalizirano liječenje napredak UI-e nosi i etičke izazove do kojih dolazi prilikom donošenja odluka, kao što su nedostatak objašnjivosti, povjerenja ili prisutnost pristranosti. OUI, kao relativno nova metoda u zdravstvu, nastoji prevladati navedene izazove pružajući objašnjenja za predviđanja sustava UI-e, čime se može povećati povjerenje u primjenu digitalnih tehnologija u zdravstvu (Rosenbacke i sur., 2024, 2). Iako ćemo o objašnjivosti i OUI-i detaljnije govoriti u sljedećem poglavlju, objašnjivost se može shvatiti kao karakteristika sustava UI-e koja čovjeku omogućuje rekonstruirati zašto je određen sustav UI-e došao do određenih predviđanja (Amann i sur., 2020, 2). No važno je napomenuti kako objašnjivost ima mnogo aspekata i, nažalost, terminologija objašnjivosti nije dobro definirana o čemu ćemo više govoriti u sljedećem poglavlju. Daljnji razvoj OUI-e može igrati ključnu ulogu za napredak zdravstvene skrbi, što zahtijeva razvoj etičkih i pravnih normi za odgovornu primjenu u kliničkoj praksi. Dok bi transparentnost sustava UI-e trebala pomoći u potrazi za

izravnim razumijevanjem mehanizma po kojem sustav UI-e djeluje, objašnjivost bi trebala kao ključan čimbenik prilikom donošenja odluke u zdravstvu s visokim rizikom pružiti dovoljnu razinu razumljivosti za prosječnog pacijenta kako bi se poboljšala sigurnost i steklo povjerenje pacijenta. Nedavna istraživanja sve više naglašavaju kako bi sustavi UI-e trebali moći izgraditi sliku svijeta koji podržavaju objašnjenje i razumijevanje, umjesto da samo rješavaju izazove kao što je prepoznavanja uzoraka (Lake i sur., 2017, 51).

Sustavi UI-e u zdravstvu omogućuju uvide i uspostavljaju korelaciju između bioloških značajki koje inače ne bismo mogli otkriti. Zahvaljujući velikom skupu podataka i metodama strojnog učenja, sustav UI-e omogućuje predviđanja temeljena na složenim vezama između karakteristika pacijenata i očekivanih rezultata liječenja koje liječnik ponekad ne može identificirati, a sustav UI-e ne može ponuditi objašnjenje za identifikaciju tih veza. Na primjer, kompleksan sustav UI-e može na temelju ekspresije određenih gena kod pacijenta, činjenjenice da pacijent puši i ima srednje visok krvni tlak doći do predviđanja kako bi jedan lijek mogao bolje djelovati na pacijenta nego neki drugi lijek, iako to sustav UI-e ne može objasniti (Price, 2015, 430).

Neka od glavnih etičkih i epistemoloških pitanja na koja ćemo pokušati odgovoriti u ovom radu su sljedeća: Može li sustav UI-e „rasuđivati“ i u kojem smislu tehnološki sustav uopće može „koristiti“ ili „predstavljati“ vrijednosti? Kakvu vrstu znanja sustav UI-e posjeduje? Može li sustav UI-e uopće posjedovati znanje? Posjeduje li sustav UI-e razumijevanje? Može li čovjek nužno u potpunosti artikulirati svoje najosnovnije vrijednosti (Boddington, 2017, 19)? Te može li nam OUI pomoći u donošenju odluka?

No istovremeno, primjena UI-e uzrokuje nove vrste etičkih izazova koji nose različite rizike, kao što su informirani pristanak, povjerenje, pristranost, algoritamske pogreške i rizik od algoritamske neusklađenosti, što rezultira lažno pozitivnim ili lažno negativnim predviđanjima (Wałdoch, 2024, 122). Informirani pristanak je uvjet *sine qua non* kako bi medicinska intervencija bila zakonita (Bal, Choma, 2012, 1346). Informirani pristanak može, ovisno o različitim oblicima pravnog uređenja unutar država, prebaciti rizik nuspojava i drugih neželjenih ishoda liječenja s liječnika na pacijenta (Cohen, 2020, 1441). Pacijent tako prilikom pristanka na medicinski postupak preuzima rizik povezan s postupkom, uključujući njegove izravne, tipične i uobičajene posljedice, o kojima treba biti pravilno informiran (Council of Europe, 2010, 31).

U kontekstu primjene UI-e u zdravstvu i principa „crne kutije“, to otvara pitanja vezana uz odgovornost pacijenta pod netransparentnim uvjetima rada. Neka od pitanja kojima ćemo se baviti u ovom kontekstu su: Može li informirani pristanak biti teoretski ostvariv pod netransparentnim uvjetima rada UI-e? Što liječnik treba reći pacijentu o podršci koju pruža sustav UI-e u medicinskom postupku? Može li pacijent, na temelju pristanka, snositi odgovornost i za rizike povezane s nepredvidljivim i neobjašnjivim načinom rada UI-e? Ako da, u kojoj mjeri i koje informacije liječnik treba pružiti pacijentu kako bi se pristanak smatrao svjesnim i informiranim? Treba li liječnik pacijenta uvijek obavijestiti kako UI sudjeluje u medicinskom postupku? Treba li pacijentu objasniti zašto je UI predložila određenu dijagnozu, prognozu ili terapiju bolesti?

Osim informiranog pristanka, povjerenje je izazov koji je predmet proučavanja brojnih znanstvenika dulji niz godina, s ciljem oslanjanja ljudi na tehnologiju na odgovarajući način (Lee, See, 2004, 50). S obzirom kako ljudi različito reagiraju na tehnologiju i sve veću automatizaciju, upravo će povjerenje u budućnosti imati ključnu ulogu u oblikovanju načina na koji će se ljudi osloniti na sustave UI-e. Povjerenje može igrati ključnu ulogu pri donošenju odluka u kontekstu složenih i nepredvidivih situacija, u kojima potpuno razumijevanje rada sustava UI-e nije moguće (Lee, See, 2004, 50). Neki znanstvenici smatraju kako objašnjivost i OUI mogu posredovati povjerenje u primjenu UI-e u zdravstvu iako je prije svega potrebno jasno definirati odnos između objašnjivosti i povjerenja kako bismo bili sigurni u istinitost navedene tvrdnje (Ferrario, Loi, 2022, 1) o čemu ćemo više govoriti u nastavku teksta.

Povjerenje u primjenu sustava UI-e na posebnoj je kušnji u situacijama kada predviđanja nisu u skladu s razumijevanjem do kojeg je došao čovjek. U kontekstu zdravstvene skrbi, od najvećeg je značenja da pacijenti i zdravstveno osoblje imaju maksimalno povjerenje u predviđanja sustava UI-e, a prilikom nesklada ili nerazumijevanja modela i čovjeka sustavi OUI-e pružaju mogućnost objašnjenja mehanizma djelovanja (Baker, Xiang, 2023, 23). Strategije za razvoj i reguliranje povjerenja u sustave UI-e još uvijek su u razvoju (Markus i sur., 2021, 8), a unatoč velikom broju akademskih istraživanja na tu temu, i dalje postoje temeljna neslaganja o prirodi povjerenja (Starke, Ienca, 2024, 360). S obzirom kako su UI i povjerenje blisko povezani, od velike je važnosti istražiti razvoj povjerenja s naglaskom na uspostavljanje prakse povjerenja (Shevtsova i sur., 2024, 1) u primjenu sustava OUI u zdravstvenoj skrbi.

Filozofski gledano, izazov objašnjivosti može imati poprilično zanimljiv zaplet s obzirom na donošenje odluka. Prikladno prikupljanje i razumijevanje informacija je u procesu donošenja

odluka od ključne važnosti. No, poznato je kako brojne društvene pojave, pa čak i činjenice, prilikom interpretacije mogu poprimiti različito značenje ovisno o drugačijim obrazovnim pozadinama ili kutu promatranja pojedinaca. U jednom od najznačajnijih djela svjetske književnosti Victor Hugo opisujući situaciju u kojoj se našao gospodin Myriel ističe važan epistemološki trenutak:

> „Gospodin Myriel morao je proći kroz sudbinu svakog pridošlice u malom gradu, gdje ima mnogo usta koja govore, a vrlo malo glava koje misle. Bio je prisiljen proći kroz to iako je bio biskup, i zato što je bio biskup. Ali uostalom, glasine s kojima se povezivalo njegovo ime bile su samo glasine - buka, kazivanja, riječi; manje od riječi - *palabres*,[3] kako to izražava energičan jezik Juga.“ (Hugo, 1887, 14-15).

Hugo izražava skepsu prema društvenim masama, ne samo zbog njihove sklonosti površnom razmišljanju nego i zbog njihove neujednačenost i nedosljednosti. Drugim riječima, moglo bi se reći kako „gomila ima previše glava kako bi imala jednu misao“. To znači kako mnoštvo pojedinaca ne mora nužno voditi kolektivnoj mudrosti ili istini, nego se pri donošenju zaključka ili suda često rađaju konfuzija, proturječnost pa čak i impulzivne osude.

Osim toga, sociolog Georg Simmel 1908., tvrdi kako život počiva na tisuću pretpostavki koje pojedinac ne može u potpunosti pratiti i provjeriti, ali koje mora prihvatiti u dobroj vjeri. U mnogo većoj mjeri nego što se to čini, naše moderno društvo počiva na vjerovanju u iskrenu namjeru drugih. Naše najvažnije odluke gradimo na složenom sustavu pojmova, od kojih većina pretpostavlja povjerenje da nismo prevareni (Simmel, 1908, 389).

U zdravstvu se čest primjer opisanog scenarija može pronaći prilikom dijagnostičkog postupka kada se mišljenje dvoje ili više liječnika ne slaže oko određene dijagnoze. Na primjer, kod pacijenta s nejasnim simptomima poput boli u prsima, nespecifične slabosti ili promjenama u zdravstvenim nalazima jedan liječnik može posumnjati na autoimunu bolest, a drugi razmotriti maligni proces. Pacijent u takvim situacijama može ostati zbunjen jer liječenje ne može započeti dok se ne uspostavi precizna dijagnoza.

Neka od osnovnih pitanja na koja ćemo pokušati pronaći odgovor su: Koji ključni čimbenici utječu na razvoj povjerenja liječnika i pacijenta u primjenu UI-e u zdravstvu? Na koji način automatizacija i nerazumijevanje sustava UI-e utječu na razvoj povjerenja liječnika i pacijenta?

---

[3] *palabres* – fran. brbljanje

Može li OUI pomoći u ublažavanju nepovjerenja ili razvoju povjerenja prema primjeni sustava UI-e u zdravstvu?

Pristranost se javlja u sustavima UI-e koji se kao statistički modeli treniraju za obavljanje specifičnih zadataka kao što su predikcija i klasifikacija. Sustav UI-e može biti treniran s ciljem da odgovori na pitanja poput: Znači li prisutnost žute pjege na oku ujedno i karcinom? Da li je pacijent u riziku od moždanog udara?

U nadziranom učenju (engl. *supervised learning*), sustav UI-e se trenira na osnovu principa pokušaja i pogreške, učeći kako najprikladnije prilagoditi parametre za rješavanje određenih zadataka. Sustavi UI-e koji uče na taj način su fleksibilniji i u odnosu na ručno izrađene statističke modele, pod uvjetom da je ciljna kategorija jasno definirana i da postoji velika količina podataka za treniranje (Grote, Keeling, 2022, 86).

Na primjer, ukoliko bolnica razvija sustav UI-e za predviđanje rizika od srčanog udara u godini dana, prvo mora prikupiti veliki skup podataka o pacijentima, koji uključuju spol, podrijetlo, povijest bolesti, težinu, prehrambene navike, obiteljski status, zaposlenje, kao i informaciju jesu li doživjeli srčani udar u tom razdoblju. Sustav UI-e se trenira tako da poboljšava svoje parametre kad pogriješi, a zadržava ih kad je predikcija točna. S vremenom sustav UI-e tako uči učinkovito predviđati rizik za nove pacijente (Grote, Keeling, 2022, 87).

Sustavi UI-e koji rade na principu strojnog učenja se posebno koriste u medicinskoj dijagnostici u svrhu prepoznavanja uzoraka i predviđanju zdravstvenih rizika. U oba područja postoji mnoštvo studija u kojima su sustavi UI-e premašili prediktivne i dijagnostičke mogućnosti medicinskih stručnjaka (Liu i sur., 2019, e271). No usprkos tome, sustavi UI-e mogu biti pristrani kada rade predviđanja za skupine pacijenata koje su loše zastupljene u podacima za treniranje, što često pogađa ranjive društvene skupine poput pripadnika određene rase ili etničke pripadnosti. Iako postoje različiti razlozi pristranosti o kojima ćemo više govoriti u nadolazećem poglavlju, važno je ovdje napomenuti kako pristranost sustava UI-e u zdravstvu ima i moralnu dimenziju. Prisutnost pristranosti može biti posebno važna pri donošenju odluka o transplantaciji bubrega.

Neka od osnovnih pitanja na koja ćemo u ovom kontekstu ponuditi odgovore su: Kako pristranost u sustavima UI-e utječe na razvoj povjerenja i pravednosti pri donošenju odluka u zdravstvu? Kako pristranost u sustavima UI-e utječe na donošenje odluka za različite skupine pacijenata? Na koji način razina pristranosti u sustavima UI-e utječe na povjerenje liječnika i pacijenata?

Kako bismo mogli jasnije progovoriti o rizicima i potencijalnim rješenjima izazova upotrebe UI-e u zdravstvu, prije svega ćemo se posvetiti sistematizaciji temeljnih pojmova.

2. Sistematizacija temeljnih pojmova

Sistematizacija pojmova važna je jer omogućuje preciznije razumijevanje teme, smanjuje mogućnost pogrešnih tumačenja, te osigurava jasnoću u raspravama koje uključuju više znanstvenih disciplina: filozofiju, računarstvo i medicinu. U kontekstu zdravstva, UI-a obuhvaća različite tehnologije i metode – od strojnog učenja i analize podataka do automatizacije kliničkih odluka – koje imaju potencijal unaprijediti dijagnostiku, personaliziranu terapiju i učinkovitost zdravstvenih sustava. Međutim, bez jasne terminološke osnove, teško je raspravljati o etičkim dilemama, pravnim okvirima ili tehničkim ograničenjima. Zato je sistematizacija prvi korak prema odgovornom i promišljenom pristupu primjeni UI-e u zdravstvu.

U ovom poglavlju ćemo posvetiti pažnju definiciji nekoliko sljedećih pojmova: umjetna inteligencija, strojno učenje, duboko učenje, neuronska mreža, „crna kutija“, objašnjiva umjetna inteligencija. Razvoj tih pojmova mijenjao se kroz godine kako su i znanja postala sve složenija i preciznije su opisivala određene fenomene. Promatrajući razvoj određenih definicija kroz godine, zanimljivo je ustvrditi kako su se i prognoze znanstvenika mijenjale kroz godine. Poznata je prognoza predsjednika financijske institucije Michigan Savings Bank koji je 1903. godine savjetovao odvjetnika Henrya Forda, gospodina Horacea Rackhama da ne ulaže u tvrtku Ford Motor Company. Predsjednik Michigan Savings Bank je u trenutku pojave automobila inzistirao na tome kako će čovjek dugoročno koristiti konje izjavljujući: „Konj je ovdje da ostane, ali automobil je samo novost - prolazni hir.“ Rackham je ignorirao savjet i uložio 5.000,00 $ u Fordove dionice, koje je kasnije prodao za 12,5 milijuna dolara (Bushnell, 1922, 56).

Jedan od pokretača UI-e, Marvin Minsky, tvrdio je 1970. godine u intervjuu o robotu Shakeyu, koji je uz pomoć UI-e prvi bio sposoban za rasuđivanje o vlastitim postupcima, za časopis *Life* kako ćemo za tri do osam godina imati stroj koji će biti podržan od strane UI-e s prosječnom inteligencijom ljudskog bića (Darrach, 1970). Bila je to još jedna od brojnih pogrešnih prognoza znanstvenika. Geoffrey Hinton, dobitnik Nobelove nagrade za fiziku 2024. godine zbog temeljnih otkrića koji omogućuju strojno učenje pomoću umjetnih neuronskih mreža

(Krizhevsky i sur., 2017, 88), predložio je 2016. godine prestanak obučavanja radiologa jer će UI nadmašiti čovjeka kroz idućih 5 do 10 godina (Kelly i sur., 2020, 2). Snagu UI-e u odnosu na čovjeka je tada spektakularno opisao kao situaciju iz crtića u kojoj je Wile E. Coyote već prošao liticu, ali još nije pogledao dolje kako bi shvatio da pada u ponor (Kelly, 2025, 2642).

Demis Hassabis, osvajač Nobelove nagrade za kemiju 2024. godine za razvoj AlphaFolda, revolucionarnog sustava UI-e koji predviđa trodimenzionalnu strukturu proteina na temelju njihove aminokiselinske sekvence (Tunyasuvunakool, 2021, 591), prognozirao je kako ćemo za idućih 5 do 10 godina imati opću umjetnu inteligenciju (Perrigo, 2025). Više o različitim oblicima UI-e u nastavku teksta.

2.1. Pojmovna sistematizacija umjetne inteligencije

Razvoj UI-e dodatno je potaknut idejom o stvaranju stroja koji misli na temelju činjenice kako već postoji tehnologija koja oponaša dijelove ljudskog tijela (mikrofoni za uši, kamere za oči, roboti s balansirajućim udovima). Alan Turing u radu *Computing Machinery and Intelligence* predlaže pristup u kojem bi umjesto fizičke zamjene čovjeka i stroja, došlo do izrade "mozga bez tijela" što uključuje razvoj svojevrsnog elektroničkog "mozga" s minimalnim senzorima za vid, sluh, govor. Uz to, predlaže kako je potrebno istražiti područja u kojima bi takav stroj mogao koristiti za određene zadatke (Turing, 2004, 420).

Za razliku od brojnih znanstvenih disciplina poput matematike, fizike, kemije i biologije, koje su već jasno definirane, područje UI-e još uvijek nije u potpunosti precizirano. Debata oko definicije UI-e još uvijek je otvorena (Monett i sur., 2020, 7; Bhatnagar i sur., 2018, 120). Možemo reći kako je istraživanje o definiciji UI-e fragmentirano i kako još uvijek postoje izazovi u definiranju pojma inteligencije (Huang, 2024, 1), a neki izazovi proizlaze iz ograničenosti UI-e koje se ignoriraju u zajednici koja se bavi UI-om (Landgrebe, Smith, 2025, 4).

Rasprava oko definicije UI-e započela je još u 60-im godinama prošlog stoljeća kada su prvu definiciju UI-e iznijeli John McCarthy i suradnici na konferenciji pod nazivom *The Dartmouth Summer Research Project on Artificial Intelligence* 1955. godine u istraživačkom prijedlogu. McCarthy i suradnici su u tom istraživačkom prijedlogu govorili o UI-i kao o simulaciji ljudskog razmišljanja, učenja i drugih aspekata inteligencije (McCarthy i sur., 1955, 2). U toj definiciji UI-e nije bilo govora o imitaciji ljudskog ponašanja ili inteligencije kao što je to

nekoliko godina ranije zamišljao Turing. Prema definiciji koju su dali McCarthy i suradnici, umjetno inteligentni strojevi su oni koji simuliraju ljudsku inteligenciju. Smatra se ipak kako je prije McCarthya prvi puta u javnosti o ideji UI-e koja imitira ljudsku inteligenciju govorio engleski matematičar Alan Turing tijekom predavanja Londonskom matematičkom društvu 20. veljače 1947. godine (McCarthy, 2007, 1175). Značajno je kako sintagma „umjetna inteligencija“ još nije uvedena u trenutku kada je Turing javno govorio ili pisao svoje najznačajnije radove. No prema Turingovom razumijevanju UI se postiže kada stvorimo strojeve koji mogu uspješno oponašati misleća ljudska bića (Dignum, 2019, 11). Ukoliko na trenutak preskočimo značajan doprinos razvoju definicije UI-e i stanemo kod Stuarta Russella i Petera Norviga vidjet ćemo kako su oni u prvom izdanju široko korištenog priručnika iz 1995. godine promijenili način definiranja UI-e. Prema Russellu i Norvigu, UI-u bismo trebali definirati kao stvaranje inteligentnih agenata, a agent se smatra sustavom koji može percipirati svoju okolinu i djelovati u svrhu ostvarivanja određenih ciljeva (Russell, Norvig, 2021, 7). Inteligenciju su definirali u smislu onog što filozofi nazivaju instrumentalnom racionalnošću: sposobnost učinkovitog postizanja vlastitih ciljeva (Russell, Norvig, 2021, 4). Stvaranje UI-e, prema ovom načinu razmišljanja, se sastoji u stvaranju sustava koji mogu percipirati svoju okolinu i djelovati u okolini kako bi učinkovito postigli svoje ciljeve. To može biti softverski agent u obliku računalnog programa ili sustava UI-e ili to može biti robot. Na taj način Russell i Norvig omogućuju da UI-a bude bestjelesna djelujući samo unutar računala, ili može biti utjelovljena, djelujući u prirodnom okruženju (Nyholm, 2023A, 1231).

UI-u se može pokušati definirati tako da izaberemo dvije definicije koje su metodološki suprotne. Prvu takvu definiciju je ponudio Marvin Minsky u kojoj je UI-u nazvao znanošću o tome kako strojevi mogu raditi stvari za koje bi im bila potrebna inteligencija ako bi ih radili ljudi (Minsky, 1968, 7). Margaret Boden je UI-u definirala kao korištenje računalnih programa i tehnika programiranja kako bi se istovremeno rasvijetlila načela inteligencije te ljudska misao (Boden, 1977, 5).

Minskyjeva definicija sugerira kako je UI metodološki usmjerena ka proučavanju ljudi s ciljem učenja o programiranju računala. No, Bodenova definicija sugerira metodologiju koja ide u suprotnom smjeru: proučavanje računala kako bismo učili o ljudima. Alan Turing je također zagovarao ovaj pristup jer je smatrao kako će nam „pokušaj da se napravi misleći stroj uvelike pomoći u otkrivanju kako mislimo o sebi“ (Turing, 1951, 486). UI-u zapravo možemo shvatiti kao dvosmjernu ulicu u kojoj se ujedinjuju dva legitimna pogleda: Minskyjev pogled na UI-u

kao prelazak od ljudi prema računalima, a Bodenov pogled kao prelazak od računala prema ljudima.

Definicija UI-e se mijenjala tijekom vremena i u određenom dijelu znanstvene zajednice postoji razumijevanje kako razvoj inteligentnih računala znači stvaranje nečega što danas ne postoji zbog čega se može reći kako je definicija UI-e dodatno podložna izmjenama (Bartneck i sur., 2021, 7-8). U oblikovanju definicije UI-e doprinijelo je nekoliko različitih znanstvenih disciplina od kojih su najistaknutije sljedeće: filozofija, matematika, ekonomija, neuroznanost, psihologija, računarstvo, kibernetika, lingvistika (Russell, Norvig, 2021, 24-35).

Kaplan i Haenlein su tako UI-u definirali kao sposobnost sustava da ispravno interpretira podatke, uči iz tih podataka i koristi saznanja za postizanje određenih ciljeva putem fleksibilne prilagodbe (Kaplan, Haenlein, 2019, 3). Poole i Mackworth (2010) definiraju umjetnu inteligenciju kao „područje koje proučava sintezu i analizu računalnih agenata koji djeluju inteligentno“ (Poole, Mackworth, 2010, 3). Agent je u ovoj definiciji nešto (ili netko) što djeluje. Inteligentan agent je onaj čije akcije su primjerene njegovim okolnostima i ciljevima, koji je fleksibilan prema promjenjivim okruženjima i promjenjivim ciljevima, koji uči iz iskustva i koji donosi odgovarajuće izbore s obzirom na svoja perceptivna i računalna ograničenja (Bartneck i sur., 2021, 8).

Važno je istaknuti kako literatura razlikuje nekoliko različitih vrsta UI-e, a tri glavne vrste su: slaba ili uska umjetna inteligencija (engl. *Artificial Narrow Intelligence*), jaka ili opća umjetna inteligencija (engl. *Artificial General Intelligence*) i superinteligencija (engl. *Artificial Super Intelligence*). Slaba UI (ponekad se naziva i „uska UI“) može obavljati samo jedan specifičan zadatak, s ograničenim rasponom sposobnosti. Česti primjeri slabe UI-e su glasovni asistenti poput Alexe ili Siri koji mogu dobro prepoznati glas čovjeka, ali ne mogu obavljati druge zadatke poput upravljanja automobilom. Jaka UI (ponekad se naziva i „opća UI“) može obavljati zadatke u nekoliko različitih područja kao što su prepoznavanje glasa, pisanje teksta ili priprema hrane. Opća UI može autonomno riješavati izazove iz različitih područja s rezultatima sličnim ili identičnim kao čovjek. Superinteligencija može obavljati zadatke iz svih područja, trebala bi imati karakteristike samosvjesnosti i autonomno riješavati izazove iz svih područja postižući rezultate koji nadmašuju čovjeka (Kaplan, Haenlein, 2019, 2).

Predložene su i brojne definicije UI-e koje su prikazane u člancima sa sveobuhvatnim analizama pojma UI-e (Legg, Hutter, 2007, 392-394; Monett, Lewis, 2018, 214). U istraživanju koje je provedeno u 48 zemalja i u 131 različitih institucija sa znanstvenicima, edukatorima,

programerima iz područja računarstva, pshilogije, kognitivnih znanosti i inženjerstva pronađena su neka od najprihvaćenijih definicija UI-e (Monett, Lewis, 2018, 214). Od ukupno 400 ispitanika, njih 246 (ili 61.5%) se najviše složilo kako je definicija UI-e Linde Gottfredson najprihvaćenija:

> „Inteligencija je vrlo općenita mentalna sposobnost koja, između ostalog, uključuje sposobnost rasuđivanja, planiranja, rješavanja problema, apstraktnog razmišljanja, razumijevanja složenih ideja, brzog učenja i učenja iz iskustva. To nije samo učenje iz knjiga, uska akademska vještina ili pamet za polaganje testova. Radije, ona odražava širu i dublju sposobnost razumijevanja naše okoline - „shvaćanja“, „shvaćanja“ stvari ili „shvaćanja“ što učiniti.“ (Gottfredson, 1997, 13)

Druga najprihvaćenija definicija UI-e je radna definicija Pei Wanga s kojom se složilo 224 ispitanika (ili 56%): „Inteligencija je sposobnost informacijskog sustava da se prilagodi svom okruženju dok radi s nedovoljnim znanjem i resursima.“ (Wang, 1995, 13).

Osim toga, često se pojavljuje i definicija Shanea Legga i Marcusa Huttera koja ističe kako inteligencija mjeri sposobnost agenta u postizanju ciljeva u različitim okruženjima (Legg, Hutter, 2007, 399),

Ovakav pristup pri definiciji UI-e je legitiman, ali postoje razni drugi pristupi definiranju UI-e. Važno je napomenuti kako bi pitanje „Što je UI?“ bila puka semantika ukoliko odgovori na to pitanje ne bi sugerirali radikalno različite puteve istraživanja. Zbog toga Minsky tvrdi kako postoje tri skupine znanstvenika koji različito pristupaju UI-e. Prema Minskyu, prva skupina su znanstvenici koji žele da strojevi rade razne stvari koje ljudi nazivaju inteligentnima. Druga skupina znanstvenika žele razumjeti što ljudima omogućuje da rade takve stvari. Treća skupina znanstvenika želi pojednostaviti programiranje (Minsky, 1991, 35).

Literatura u 90-im godinama prošlog stoljeća bilježi raspravu vezanu uz značenje riječi „umjetna“ i „inteligencija“ (Hofstadter, 1985, 633), a nakon otprilike 70 godina istraživanja u području UI-e još uvijek ne postoji široko prihvaćena definicija UI-e. Tako jedna od definicija je označavala UI-u kao manipulaciju simbolima za rješavanje problema (Buchanan, Shortliffe, 1985, 3) dok je jedna tautološka tvrdnja UI-u smjestila u dio računalne znanosti koji se bavi razvojem inteligentnih računalnih programa (Waterman, 1986, 10). Elaine Rich je UI-u definirala kao proučavanje načina na koje se računala mogu osposobiti da obavljaju zadatke u kojima su ljudi trenutačno uspješniji (Rich, 1983, 1). No ova definicija je manjkava jer

isključuje postojanje UI-e. Nakon što računalni program premaši ljudske sposobnosti, ispada kako se program više ne nalazi u domeni UI-e.

Mnoštvo ponuđenih definicija UI-e uključuje opisana ljudska ponašanja koja stavlja u usporedni odnos s UI-om. U tom kontekstu, Staugaard je UI-u definirao kao mehanizaciju ili dupliciranje ljudskog misaonog procesa (Staugaard, 1987, 23). Charniak i McDermott (Charniak, McDermott, 1985, 6) su tvrdili kako se UI bavi proučavanjem mentalnih sposobnosti korištenjem računalnih modela, dok je Herbert Schildt smatrao kako je inteligentni program onaj koji pokazuje ponašanje slično ljudskom kada se suoči sa sličnim problemom. Prema Schildtu, nije nužno za program da zapravo riješi ili pokuša riješiti problem na isti način kao što bi to učinio čovjek (Schidlt, 1987, 11). Znanstvenici naglašavaju kako UI nije umjetni oblik ljudske inteligencije nego je UI proizvod ljudske inteligencije, artefakt (Holm, Banerjee, 2025, 1170; Johnson, Verdicchio, 2025, 68). Zbog toga UI oponaša ljudsku inteligenciju, a nije inteligencija u punom smislu riječi.

No, osim ljudi, inteligencija se proučavala i u odnosu prema životinjama. George John Romanes, kao jedan od utemeljitelja komparativne psihologije, nudi definiciju inteligencije kao sposobnosti da se učini prava stvar u pravo vrijeme (Romanes, 1882, 17). To je sposobnost reagiranja na prilike i izazove koje predstavlja kontekst. Ova jednostavna definicija je važna jer demistificira inteligenciju, a kroz nju i umjetnu inteligenciju. Pojašnjava i granice inteligencije i naše vlastite društvene odgovornosti na dva načina.

Prvo, inteligencija je proces koji djeluje na određenom mjestu i u određenom trenutku. Radi se o specifičnom obliku računanja koji fizički transformira informacije (Sipser, 2012, 286). Informacija nije apstrakcija (Shannon, 1948, 385). Informacija se fizički može manifestirati u energiji (svjetlu ili zvuku) ili materijalima. Računanje i inteligencija stoga također nisu apstrakcije i zahtijevaju vrijeme, prostor i energiju. Zbog tih zahtjeva je nemoguće istovremeno voditi izračune ili razmišljati o svemu istovremeno. Možemo praviti kompromise: na primjer, možemo udvostručiti broj računala koja koristimo i gotovo prepoloviti vrijeme izračuna. No, vrijeme se nikada ne prepolovi, jer uvijek postoji dodatan trošak dijeljenja zadataka i ponovnog kombiniranja rezultata obrade. Nema dokaza da će kvantno računarstvo temeljno promijeniti ovu jednadžbu troškova: trebalo bi uštedjeti ne samo vrijeme već i prostor, međutim, troškovi energije nisu dobro shvaćeni i do danas izgledaju iznimno visoki (Bryson, 2020, 4).

Drugo, razlika između inteligencije i umjetne inteligencije je samo kvalifikator. Umjetno znači da je nešto napravljeno ljudskim procesom. To znači da su ljudi po zadanim postavkama

odgovorni za umjetnu inteligenciju. Artefakt koji je ovdje zapravo još zanimljiviji od umjetne inteligencije je koncept: odgovoran. Druge životinje mogu se dresirati da namjerno ograniče gdje stavljaju (na primjer) čak i prilično nenamjerne nusproizvode svog probavnog procesa, ali koliko znamo, samo ljudi imaju odgovornost, mogu komunicirati i, što je ključno, mogu pregovarati o eksplicitnom konceptu odgovornosti (Bryson, 2020, 5). Naša društva s vremenom i razvojem teže ponuditi odgovornosti za posljedice, a digitalne tehnologije nam pružaju sposobnosti boljeg percipiranja i vođenja evidencije o postupcima i posljedicama zbog čega bi nam trebalo biti lakše održavati odgovornosti i provoditi zakone. U tom smislu i UI može mijenjati naše razumijevanje odgovornosti. Međutim, način na koji preuzimanje odgovornosti u interakciji s UI-om može biti olakšano ili otežano ovisi o tome kako koristimo potencijal koje pruža UI. Na primjer, ukoliko liječnik koristi sustav UI-e za predviđanje rizika komplikacija, odgovornost za konačnu odluku o liječenju i dalje snosi liječnik, ali bi sustav UI-e mogao olakšati odgovornost pružanjem transparentnih objašnjenja svojih predviđanja. S druge strane, ako sustav UI-e ne pruža objašnjenja ili djeluje isključivo na principu „crne kutije“, odgovornost postaje puno teže za procijeniti jer je teško procijeniti pouzdanost predviđanja sustava UI-e. Bez brige i odgovarajućih mjera, povećani kapacitet za informacije i komunikaciju može se koristiti za raspršivanje ili prikrivanje odgovornosti.

U kontekstu zdravstva, znanstvenici ističu kako sustavi UI-e ne mogu biti odgovorni čak niti za točna predviđanja koja su u potpunosti u skladu s pravilima struke, jer čak i ako sustavi UI-e mogu nadmašiti liječnike u nekim aspektima, oni ne posjeduju slobodnu volju niti moralnu osobnost (Zhang, Zhang, 2023, 6; Iniesta, 2025, 391), a niti jedan sustav UI-e do sada nije demonstrirao razinu svijesti koja bi implicirala mogućnost pripisivanja odgovornosti (Coeckelbergh, 2020B, 2054). Znanstvenici smatraju kako bi moralni subjekt morao imati određena ljudska svojstva koja su preduvjet za posjedovanje moralne osobnosti kao što su svijest, emocionalnost, slobodna volja i autonomija (Puzio, 2025, 1024; Gunkel, 2018, 91). Ovaj način pripisivanja moralnog statusa sustavima UI-e popularizirao je Mark Coeckelbergh nazvavši ga „pristupom svojstava“ koji ima značajan utjecaj u području etike robota. „Pristup svojstava“ pretpostavlja kako moralno razmatranje robota i načina na kojeg bismo ga trebali tretirati ovisi o tome posjeduje li jedno ili više spomenutih ljudskih svojstava (Coeckelbergh, 2012, 13). Zbog nedostatka ljudskih svojstava kod sustava UI-e, liječnici kao kompetentni ljudski agenti trebaju preuzeti odgovornost za svoje kliničke odluke i njihove odluke se trebaju poštovati (Van Cauwenberge i sur., 2022, 12). Neki znanstvenici idu i korak dalje tvrdeći kako je važno poštivati odluku liječnika čak i kada se suprotstavlja predviđanju sustava UI-e (Faden

i sur., 2013, 21-22). Analiza pripisivanja moralnog statusa sustavima UI-e od velike je važnosti pri utvrđivanju sustava UI-e kao epistemoloških autoriteta o čemu ćemo detaljnije govoriti u nadolazećem tekstu.

U kontekstu definicije inteligencije, ovdje je još važno istaknuti kako se i zbog izbjegavanja regulacija ili odgovornosti, definicija inteligencije namjerno predstavlja kompleksnom pomoću pojmova kao što su osjećajnost, svijest, intencionalnost i slično (Bryson, 2020, 5). Dodatan razlog zašto je izazovno definirati umjetnu inteligenciju jest što se inteligencija često pripisivala jedino ljudima. Atmar tako ističe sljedeće:

> „Možda je glavni problem naše polazište. Inteligencija se općenito smatra jedinstveno ljudskom kvalitetom. Pa ipak, mi, kao ljudi, ne razumijemo sebe, svoje sposobnosti ili podrijetlo svoje misli. U našoj žurbi da katalogiziramo i oponašamo vlastiti zapanjujući niz bihevioralnih reakcija, logično je posumnjati da su istraživanja primarnih uzročnih čimbenika inteligencije zanemarena kako bi se brže dobile neposredne posljedice inteligencije.“ (Atmar, 1976, VIII).

Minsky, s druge strane, nemogućnost definiranja UI-e pripisuje pokušajima ujedinjavanja različitih teorija inteligencije. Prema Minskyom ne postoji jedan najadekvatniji način za predstavljanje znanja ili rješavanje problema, a ograničenja trenutne UI-e uvelike proizlaze iz traženja ujedinjenih teorija ili pokušaja popravka nedostataka teoretski urednih, ali konceptualno osiromašenih ideoloških pozicija (Minsky, 1991, 35).

Zagovornici razvoja UI-e od Turinga naovamo često su tvrdili kako cilj istraživanja UI-e ne bi trebao biti stvaranje ljudske inteligencije nego drugačije vrste strojne inteligencije. Pri tome znanstvenici poput Nicka Bostroma i Eliezera Yudkowskog ističu kako, ako UI-u definiramo kroz prizmu ljudske inteligencije, riskiramo upasti u zamku antropomorfizma i pripisivati UI-i ljudske osobine, emocije, namjere ili način razmišljanja (Bostrom, 2014, 127; Yudkowsky, 2001, 21).

Stuart Russell i Peter Norvig identificirali su četiri različite škole razumijevanja UI-e. Prvoj školi pripadaju istraživači koji se usredotočuju na stvaranje sustava koji razmišljaju poput ljudi. Kroz povijest UI-e ova ideja, kako bi nam modeliranje ljudskih misaonih procesa moglo omogućiti da na neki način repliciramo takve procese u računalnim sustavima, je bila veoma snažna i proizašla je iz kognitivnih znanosti (Russell, Norvig, 2021, 20-21).

Druga škola usredotočuje se na stvaranje sustava koji se ponašaju poput ljudi. Takav pristup je primjenio Alan Turing, tvorac Turingovog testa koji pokušava utvrditi da li je računalni program inteligentan pomoću jednostavnog testa. Cilj Turingovog testa je utvrditi može li računalo zavarati čovjeka na način da čovjek povjeruje kako komunicira s drugim čovjekom umjesto s računalom (Turing, 1950, 442). Drugim riječima, ukoliko čovjek ne može razlikovati ponašanje računala od ponašanja ljudskog bića, računalo je pokazalo inteligenciju. Zanimljivo je istaknuti kako Turingov test na ispit stavlja cijeli niz sposobnosti u kojima se UI usavršava od nastanka, uključujući znanje, rasuđivanje, razumijevanje jezika i učenje. Ova škola se usredotočuje na djelovanje, što stroj ili računalo zapravo radi u svijetu, a ne na proces dolaska do tog djelovanja (Russell, Norvig, 2021, 20).

Treća škola usredotočuje se na razvoj sustava koji djeluju racionalno. Za Russella i Norviga to uključuje djelovanje kako bi se postigao najbolji ishod. Russell i Norvig favoriziraju ovaj pristup izgradnje takozvanih racionalnih agenata, ističući kako on zapravo uključuje mnoge druge gore navedene pristupe. Osim toga, prema Russellu i Norvigu, racionalnost je usko povezana s optimalnošću što omogućuje sustavima ove škole da čine pravu stvar ili djeluju na ispravan način (Russell, Norvig, 2021, 21-22).

Konačno, četvrta škola usredotočena je na razvoj sustava koji racionalno razmišljaju. Glavni oslonac ovog usmjerenja su takozvani „zakoni mišljenja“ koji se formuliraju korištenjem posebnih sustava simbola izvedenih iz matematičke logike. Planiranje i/ili donošenje odluka koje bi ovi sustavi trebali obavljati bi trebali biti optimalni. Time se pretpostavlja kako logika može adekvatno modelirati ljudsku misao, no izazov je što je ljudska misao puna proturječja i nesigurnosti, a logička pravila ne mogu uvijek biti usklađena s razmišljanjem čovjeka (Russell, Norvig, 2021, 21).

Iako se možda velik broj znanstvenika slaže kako UI podrazumijeva računalne sustave koji su na neki način slični ljudskom umu, postoje veoma različite ideje o tome u čemu se točno ta sličnost nalazi. Različiti autori tu sličnost pronalaze u različitim elementima koji ih podsjećaju na ljudski um, kao što su strukture, ponašanje, sposobnosti, funkcije ili principi. Tako dolazimo do različitih radnih definicija UI-e (Wang, 1994, 24). Te radne definicije UI-e su sve valjane, u smislu da svaka od njih odgovara opisu ljudske inteligencije na određenoj razini apstrakcije i postavlja precizan istraživački cilj, koji je ostvariv u različitim granicama. Svaka od definicija je također plodna, u smislu da je vodila istraživanje kako bi se dobili rezultati s intelektualnim i praktičnim vrijednostima. S druge strane, ove radne definicije su različite, jer postavljaju različite ciljeve, zahtijevaju različite metode, daju različite rezultate i procjenjuju napredak

prema različitim kriterijima. Ne mogu se međusobno zamijeniti niti se "integrirati" u koherentnu definiciju koja istovremeno zadovoljava sve kriterije.

No, isto tako valja istaknuti kako su pogrešne tvrdnje: "UI se ne može definirati." ili "Sve definicije UI-e su otprilike ekvivalentne." UI može imati radne definicije koje služe kao krajnji istraživački ciljevi. Različiti istraživači u različitim područjima imaju radnu definiciju UI-e, iako često implicitno (Cappelli i sur., 2024, 2). Sa svrhom jasnije komunikacije i poboljšanja koherentnosti istraživanja, uvijek je bolje da te radne definicije budu eksplicitne. Ova tema je važna za UI-u, budući kako trenutna istraživanja UI-e pate od zbrke različitih ciljeva i nedostatka preciznih eksplicitnih definicija. Posljedično, mnoge rasprave uzrokovane su različitim značenjima pojma "umjetna inteligencija", a područje u cjelini je fragmentirano iznutra, kao i postojanje problema opravdavanja svoje jedinstvenosti i integriteta pred vanjskim svijetom. Iako trenutno nijedna radna definicija UI-e nije savršena ili konačna, odbacivanje ovog pitanja oštetit će dosljednost dizajna i evaluacije sustava. Iako postoji mnogo valjanih načina definiranja UI-e, nisu svi jednako dobri. Nećemo uskoro postići konsenzus o tome koji je najbolji, pa će u tom području različite radne definicije koegzistirati dugo vremena. Unatoč tome, važno je razumjeti njihovu razliku i odnos. Različite radne definicije daju UI-i različite identitete. Za rješavanje problema unutarnje fragmentacije i vanjskog prepoznavanja, najperspektivniji način je definiranje UI-e načelom racionalnosti kojeg slijedi ljudski um, ali ne i tradicionalni računalni sustavi (Wang, 2019, 11).

No, postoje i alternativni pristupi koji pokušavaju definirati UI-u polazeći od ideje kako je inteligencija zapravo potrebna za rješavanje problema zbog čega ju možemo razumjeti samo u odnosu na određeni problem. Softverski inženjer François Chollet smatra kako interes za UI-u proizlazi iz dubokog nerazumijevanja prirode inteligencije i ponašanja rekurzivnih samoučećih sustava, poput jake ili opće UI-e. Njegove glavne hipoteze su kako teoretičari jake ili opće UI-e koriste pogrešnu definiciju inteligencije te da ljudska inteligencija ovisi o urođenim dispozicijama, interakciji s okolinom i socijalizaciji (Chollet, 2017).

Yuval Noah Harari u svojoj najnovijoj knjizi *Nexus: Kratka povijest informacijskih mreža od kamenog doba do umjetne inteligencije* UI-u definira kao sposobnost algoritma da uči i mijenja se samostalno (Harari, 2024, 218). Osim toga, Harari uvodi akronim „izvanzemaljska inteligencija“ umjesto akronima „umjetna inteligencija“ kako bi naglasio da je način razmišljanja sustava UI-e različit od ljudskog način razmišljanja (Harari, 2024, 215).

U kontekstu spomenutih ograničenja UI-e, zanimljivo je primjetiti kako je Gary Marcus u travnju 2024. godine ponudio okladu od milijun dolara Elonu Musku ukoliko jaka ili opća UI do kraja 2025. godine ne uspije ispuniti nekoliko specifičnih zadataka. Taj popis zadataka uključuje sljedeće: vožnja terenskog vozila, bez karata, preko potoka uz uspješno zaobilaženje prepreka poput srušenih drveća, čišćenje od prije nepoznatih kuća i stanova, priprema šalice kave ili doručka u nasumičnim kuhinjama, čuvanje djece u nepoznatom domu, pružanje fizičke i mentalne podrške starijim i nemoćnim osobama (Marcus, 2024).

Navedeni popis zadataka pruža zanimljive uvide u ograničenja trenutnih sustava UI-e koji ograničeno mogu postizati određena postignuća praktičnim znanjem ili „znati kako“, za razliku od uključenog propozicijskog znanja (Landgrebe, Smith, 2025, 102), na primjer kada znamo da je 2 + 2 = 4 ili da je Zagreb glavni grad Hrvatske. Navedeni zadaci uključuju osnovne dispozicije, fenomena kojeg Landgrebe i Smith opisuju kao složbenog obrasca motoričkih i kognitivnih dispozicija koje određuju nečiju mogućnost reagiranja na unutarnje i vanjske podražaje (Landgrebe, Smith, 2025, 103). Dok se osnovne dispozicije, poput dispozicija za usvajanje i govorenje jezika, nasljeđuju, neosnovne dispozicije – poput govora korejskog ili njemačkog jezika, uče se promatranjem i imitacijom. Čini se kako su istraživači u području UI-e do sada ignorirali razvoj jake ili opće UI-e koja bi uključivala praktičnu inteligenciju ili „znanje kako“.

Napokon, valja zaokružiti kako je možda najčešći element zajednički gotovo svim definicijama taj kako UI uključuje proučavanje, dizajn i izgradnju inteligentnih sustava koji mogu postići određene ciljeve. Pod vidom teorijskih kriterija bismo mogli reći kako je ovo ujedno i pojmovna definicija inteligencije. UI je proizvod ljudske inteligencije koji oponaša ljudsku inteligenciju i predviđanja do kojih dolazi podložna su perciptivnim i kognitivnim ograničenjima. Ukoliko je UI doista fleksibilna i može učiti iz iskustva, kao i planirati i djelovati na temelju svoje početne konfiguracije, moglo bi se reći kako je inteligentnija od one UI-e koja ima samo skup pravila koji vodi fiksni skup radnji. Međutim, postoje određene situacije u kojima ne bismo željeli da UI uči nova pravila i ponašanja, kao što je to, na primjer, tijekom izvođenja kirurškog postupka zbog čega je potreban dodatan oprez kada se UI primjenjuje u situacijama visokog rizika. Pod vidom ograničenja mogućnosti i razumijevanja stvarnih sposobnosti UI-e, mogli bismo reći kako je ovo uključeno u realnu definiciju inteligencije.

2.2. Strojno učenje

U viktorijanskoj Engleskoj, Ada Lovelace bila je prijateljica i suradnica Charlesa Babbagea, izumitelja Analitičkog stroja, prvog poznatog mehaničkog računala opće namjene. Iako vizionarski i daleko ispred svog vremena, Analitički stroj nije bio zamišljen kao računalo opće namjene kada je dizajniran 1830-ih, jer koncept računanja opće namjene još nije bio izumljen. Analitički stroj je rezultat suradnje između Ade Lovelace, grofice od Lovelace, i računalnog pionira Charlesa Babbagea u publikaciji koja je opisala kako programirati prvo računalo na svijetu (Kim, Toole, 1999, 76). Bio je jednostavno zamišljen kao način korištenja mehaničkih operacija za automatizaciju određenih izračuna iz područja matematičke analize (Chollet, 2021, 3). Analitički stroj je predstavljao intelektualni potomak ranijih pokušaja kodiranja matematičkih operacija u obliku zupčanika koji su bili predstavljeni u obliku Pascalinea ili Leibnizovog brojača, usavršene verzije Pascalinea (Kakad i sur., 2023, 4). Godine 1843. Ada Lovelace je o izumu Analitičkog stroja rekla:

> Analitički stroj nema nikakvih pretenzija da išta stvori. Može učiniti što god znamo kako mu narediti da izvrši. Njegova je zadaća pomoći nam da učinimo dostupnim ono što nam je već poznato (Green, 2005, 41).

Čak i nakon višedesetljetne povijesne perspektive, zapažanje Lady Lovelace ostaje zapanjujuće. U kontekstu etike i epistemologije UI-e, slična pitanja možemo postaviti i danas: Može li računalo putem naprednih operacija računanja "stvoriti" išta ili bi rezultati računanja uvijek mogli biti ograničeni na dosadno izvršavanje procesa koje mi ljudi u potpunosti razumijemo? Ispunjava li računalo preduvjete u kojima je sposobno stvoriti bilo kakvu originalnu misao? Može li računalo učiti iz iskustva? Može li računalo biti kreativno? Njezinu je primjedbu kasnije citirao jedan od utemeljitelja UI-e Alan Turing kao "prigovor Lady Lovelace" u značajnom radu "Računalni strojevi i inteligencija" koji je predstavio Turingov test kao i ključne koncepte koji su oblikovali strojno učenje i UI-u (Turing, 1950, 454). Turing je već tada smatrao kako će računala u načelu moći oponašati sve aspekte ljudske inteligencije.

Uobičajeni način na koji računalo obavlja zadatke jest da ljudski programer zapiše pravila u obliku računalnog programa kojih se računalo treba pridržavati kako bi se ulazni podaci pretvorili u odgovarajuće odgovore, baš kao što je Ada Lovelace zapisivala detaljne upute koje je Analitički stroj trebao izvršiti. Strojno učenje mijenja takav način obavljanja zadataka. Prilikom strojnog učenja, sustav strojnog učenja pregledava ulazne podatke i odgovarajuće izlazne podatke te utvrđuje potrebna pravila na temelju ulaznih i izlaznih podataka. Sustav strojnog učenja se time obučava, a ne eksplicitno programira. Putem brojnih primjera

relevantnih za obavljanje zadatka strojno učenje u tim primjerima pronalazi statističku strukturu koja na kraju omogućuje sustavu osmišljavanje pravila za automatizaciju zadatka (Chollet, 2021, 4). Na primjer, u zadatku automatiziranog označavanja rendgenskih snimki pluća, sustav strojnog učenja koristi snimke koje je liječnik prethodno označio kao potencijalne zdravstvene prijetnje kako bi naučio statistička pravila koja povezuju određene vizualne obrasce sa specifičnim dijagnozama.

Iako je strojno učenje počelo s razvojem tek 1990-ih, brzo je postalo najpopularnije i najuspješnije razvojno podpodručje UI-e zbog lakše dostupnog bržeg hardvera i većih skupova podataka. Strojno učenje povezano je s matematičkom statistikom, ali se od statistike razlikuje na nekoliko važnih načina, u istom smislu kako je medicina povezana s kemijom, ali se ne može svesti na kemiju, jer medicina ima drugačiji predmet proučavanja s vlastitim različitim svojstvima. Za razliku od statistike, strojno učenje se bavi velikim, složenim skupovima podataka (kao što je skup podataka od milijun slika, od kojih se svaka sastoji od desetaka tisuća piksela) za koje bi klasična statistička analiza poput Bayesove analize bila nepraktična. Kao rezultat toga, strojno učenje, a posebno duboko učenje, pokazuje relativno malo matematičke teorije - možda premalo - i u osnovi je inženjerska disciplina. Za razliku od teorijske fizike ili matematike, strojno učenje je vrlo praktično područje vođeno empirijskim nalazima i duboko se oslanja na napredak u softveru i hardveru (Chollet, 2021, 4).

Postoje tri glavna tipa strojnog učenja: nadzirano učenje, nenadzirano učenje i učenje s potkrepljenjem. U nadziranom učenju sustav strojnog učenja uči iz označenih podataka koje je prethodno označio čovjek te koristi metode poput regresije i klasifikacije. U nadziranom učenju se uspješnost mjeri pomoću stope točnih pozitivnih i lažno pozitivnih rezultata prilikom predviđanja novih podataka. U nenadziranom učenju sustav strojnog učenja pronalazi obrasce u neoznačenim podacima i koristi metode poput analize glavnih komponenti i grupiranja. Nenadzirano učenje se često koristi kao uvod u nadzirano učenje. U učenju s potkrepljenjem sustav strojnog učenja ne koristi označene podatke nego se oslanja na povratne informacije u obliku povratnih informacija kao nagrade koja usmjerava strategiju za postizanje određenog cilja. Učenje s potkrepljenjem često se primjenjuje u robotici gdje se robot uči kretati na temelju nagrade koju dobiva za uspješno ispunjen cilj kretanja u određenoj okolini (Bartneck i sur., 2021, 11-12). Ovdje je zanimljivo napomenuti kako je učenje s potkrepljenjem ostvarilo izuzetne rezultate u igri Go, a AlphaGo je prvi program koji je na temelju učenja s potkrepljenjem pobjedio svjetskog prvaka u ovoj igri iako se očekivalo kako će proći nekoliko desetljeća za tako značajne uspjehe (Silver i sur., 2017, 354). AlphaGo je pobijedio Lee Sedola

u igri u kojoj je dokazao kako učenje s potkrepljenjem može stvarati nova znanja. AlphaGo je više puta igrao sam protiv sebe kako bi učeći iz svojih pogrešaka razvijao nove strategije u igri. Nekoliko poteza koje je AlphaGo koristio u igri protiv Sedola nikada nije viđeno u ljudskim igrama, ali potez broj 37 je preokrenuo stoljeća tradicionalne mudrosti igre Go. Potez 37 je bio iznenađujuć i nekonvencionalan, a AlphaGo je time istaknuo potencijal UI-e za inovacije složenih strategija (Hassabis, 2017, 413).

2.3. Duboko učenje

Prije nego što definiramo duboko učenje i razumijemo razliku između dubokog učenja i drugih pristupa strojnog učenja, potrebna nam je preciznija ideja što sustavi strojnog učenja rade. Naveli smo kako strojno učenje prepoznaje pravila na temelju kojih izvršava zadatke, uz primjer očekivanja za uspješno izvršen zadatak. Za strojno učenje su tako potrebna minimalno tri elementa: ulazni podaci, primjeri očekivanih izlaznih podataka i adekvatno mjerenje kvalitete rada sustava strojnog učenja. Ulazni podaci ovisno o vrsti zadatka mogu biti zvučne datoteke, slike, tablični podaci, tekst i slično. Primjeri očekivanih izlaznih podataka mogu biti transkripti zvučnih datoteka ili prijedlozi medicinskih dijagnoza poput karcinoma, Alzheimerove bolesti ili moždanog udara. Mjerenje kvalitete rada sustava strojnog učenja je potrebno kako bi se odredila razlika između trenutnih izlaznih podataka i očekivanih izlaznih podataka. Mjerenje se koristi kao povratni signal za prilagodbu načina na koji sustav strojnog učenja dolazi do izlaznih podataka ili predviđanja. Taj korak prilagodbe nazivamo učenjem. Ukratko, učenje u kontekstu strojnog učenja, znači kako sustav strojnog učenja samo uči kako bi bolje razumio podatke i uspješnije riješio određeni zadatak. To radi na način da mijenja i prilagođava podatke na temelju signala ili povratnih informacija i postupno uči koje su promjene korisne (Chollet, 2021, 6).

Duboko učenje je specifično podpodručje strojnog učenja s novim pristupom učenja iz podataka tako da se podaci postupno otkrivaju kroz više slojeva učenja. Pojam „duboko“ u sintagmi „duboko učenje“ ne označava dublje razumijevanje postignuto novim pristupom nego se odnosi na način učenja kroz više slojeva postupnom obradom i predstavljanjem podataka. Dubina sustava dubokog učenja odnosi se na broj slojeva koji se obrađuje i kroz koje podaci prolaze tijekom obrade. Što je više slojeva, to je model "dublji". Suvremeni oblici dubokog učenja često uključuju desetke ili stotine uzastopnih slojeva podataka na temelju kojih se sustav dubokog

učenja automatski uči iz podataka. Drugi pristupi strojnog učenja, za razliku od dubokog učenja, obično se usredotočuju na učenje samo jednog ili dva sloja podataka (Chollet, 2021, 7).

U dubokom učenju, slojevi podataka se obrađuju putem modela koji se naziva neuronska mreža, koja je strukturirana u slojevima podataka naslaganim jedan na drugi. Duboko učenje i neuronska mreža su blisko povezani s obzirom kako duboko učenje koristi neuronsku mrežu kao osnovni građevni blok za učenje. Pojam „neuronska mreža“ dolazi iz neurobiologije, ali iako su neki od središnjih koncepata dubokog učenja djelomično razvijeni crpeći inspiraciju iz našeg razumijevanja mozga (posebno vizualnog korteksta), sustavi dubokog učenja nisu modeli mozga. Ne postoje dokazi koji tvrde kako mozak koristi slične mehanizme učenja koji se koriste u modernim sustavima dubokog učenja. Iako postoje popularni znanstveni članci koji tvrde kako je duboko učenje modelirano prema mozgu, to nije slučaj. Zbunjujuće je i kontraproduktivno razmišljati o dubokom učenju kao čvrstoj vezi s neurobiologijom. Za naše svrhe, duboko učenje je višeslojeviti način učenja prikaza podataka. Takva jednostavna ideja uz jednostavne mehanizme ipak može ispasti vrlo kompleksna za razumijevanje s obzirom na postupak učenja. Već smo napomenuli kako se postupak kod strojnog učenja odvija tako da uspoređujemo ulazne podatke s odgovarajućim izlaznim podacima te utvrđujemo potrebna pravila na temelju usporedbe ulaznih i izlaznih podataka. Neuronske mreže postupak dubokog učenja provode na način da pretvaraju ulazne podatke u željeni izlaz pomoću više uzastopnih slojeva, pri čemu svaki sloj vrši jednostavnu transformaciju podataka. Te transformacije se ne programiraju ručno, nego se uče automatski analizom velikog broja primjera (Chollet, 2021, 8).

Svaki sloj u neuronskoj mreži nosi određeni skup brojeva koji se nazivaju težine. Težina pojedinog sloja određuje daljnje postupanje s ulaznim podacima. Kada kažemo kako jedan sloj „uči“, to zapravo znači kako zapravo pronalazi prave vrijednosti za određenu težinu sloja. Neuronska mreža uči tako da postupno prilagođava težine svih svojih slojeva kako bi mogla ulazne podatke adekvatno pretvoriti u odgovarajuće izlaze. No, neuronska mreža može imati desetke milijuna takvih težina, pa pronalaženje pravih vrijednosti može biti jako složeno – pogotovo jer promjena jedne težine može utjecati na rad cijele mreže.

Pojednostavljeno, za uspješno upravljanje neuronskom mrežom potrebno je znati koliko dobro ona radi. Zato koristimo funkciju gubitka koja uspoređuje izlazne podatke ili predviđanja neuronske mreže s onim što smo očekivali i na temelju usporedbe funkcija gubitka nam govori o kolikoj je pogrešci riječ. Ova pogreška se koristi kao povratna informacija za prilagodbu rada neuronske mreže. Na temelju toga, prilagodbom se mijenjaju težine u neuronskoj mreži tako da sljedeći put izlazni podaci ili predviđanja budu bliže očekivanim podacima, to jest, da

neuronska mreža pogriješi manje. Proces računanja pogreške i prilagođavanja težine se zove povratno širenje (engl. *backpropagation*) i ključan je za učenje mreže. Na početku su težine slučajne, pa su i rezultati loši. No, mreža uči tako što se kroz mnogo primjera i koraka težine malo po malo prilagođavaju. Kada je funkcija gubitka dovoljno niska, smatramo kako je neuronska mreža naučila zadatak (Chollet, 2021, 9).

Slika 1. Prikaz odnosa između UI-e, strojnog učenja i dubokog učenja.

2.4. Neuronske mreže

Već smo napomenuli kako pojam „neuronska mreža“ dolazi iz neurobiologije. Također, rekli smo kako je duboko učenje, kao specifična podvrsta strojnog učenja, blisko vezana uz neuronsku mrežu. Neuronsku mrežu možemo zamisliti kao višestupanjski proces destilacije informacija, gdje informacije prolaze kroz uzastopne filtere i izlaze sve pročišćenije i korisnije za neki zadatak. Nastanak neuronskih mreža iz perspektive filozofije, uz važnost psihologije i kognitivnih znanosti te povratak neuronskih mreža u računarstvo detaljno je opisao Sandro Skansi u knjizi *Introduction to Deep Learning* (Skansi, 2018, 1-15).

Danas je doprinos neuronskih mreža priznat u javnoj zajednici za što su dodijeljene i dvije Nobelove nagrade za fiziku u 2024. godini. Jedna Nobelova nagrada za fiziku je dodijeljena Geoffreyu Hintonu, zbog otkrića koji omogućuju strojno učenje pomoću umjetnih neuronskih mreža što smo već spomenuli. Druga Nobelova nagrada za fiziku je dodijeljena Johnu Hopfieldu za razvoj Hopfieldove neuronske mreže što je matematički model koji je omogućio asocijativno pamćenje zahvaljujući radu neuronskih mreža (Benani, Messas, 2025, 277). Međutim, kroz povijest potencijal kojeg nose neuronske mreže nije bio prepoznat zbog raznih razloga kao što su ograničena računalna snaga (Thompson i sur., 2022, 1) ili nedostatak podataka (Krizhevsky i sur., 2017, 88).

Prvi matematički model neuronske mreže objavili su 1943. godine neurofiziolog i kibernetičar Warren McCulloch te kognitivni psiholog Walter Pitts u radu pod nazivom *A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics* (McCulloch, Pitts, 1943, 99). Nadovezujući se na ideje o formalizaciji logičkih operacija iz rada Alana Turinga *On computable numbers, with an application to the Entscheidungsproblem* (Turing, 1936, 230), McCulloch i Pitts u svom radu objašnjavaju kako se misaoni procesi mogu opisati pomoću strukture i funkcije neuronskih mreža u ljudskom mozgu. Jedan od ciljeva McCullocha i Pittsa bio je stvaranje poveznice između fiziologije i logike kako bi pokazali da mozak izvršava formalne logičke operacije kroz biološke strukture (Abraham, 2002, 8-9) čime neuronske mreže teoretski mogu simulirati svaku operaciju koju Turingov stroj može izvršiti. Valja istaknuti i kako je Alan Turing bio na tragu velikih otkrića i potencijala kojeg nose neuronske mreže opisujući strukuturu neuronske mreže u radu iz 1948. godine: „Predlažem istražiti pitanje je li moguće da strojevi pokazuju inteligentno ponašanje“. s obzirom kako se do tada bez rasprave pretpostavljalo kako to nije moguće (Turing, 2004, 410). Detaljnije povijesne i bibliografske crtice o razvoju neuronskih mreža i dubokog učenja mogu se naći u knjigama *Artificial Intelligence: A Modern Approach* Stuarta Russella i Petera Norviga

(Russell, Norvig, 2021, 823) te *Introduction to Deep Learning, From Logical Calculus to Artificial Intelligence* Sandra Skansija (Skansi, 2018, 11).

Tijekom povijesti, neuronske mreže su bile osporavane s obzirom na nedostatak rezultata. U isto vrijeme, simbolički sustavi koje su Simon, Shaw i Newell razvijali tijekom 1960-ih pokazuali su impresivne rezultate jer su nudili kontrolu i jednostavno proširivanje. S druge strane, rezultati neuronskih mreža bili su prepoznavanje slika koji se tada nisu smatrali "inteligentnima" (Skansi, 2018, 6). Specifičan izazov za razvoj neuronskih mreža događao se za vrijeme Hladnog rata kada je nakon neuspješne automatizacije prijevoda jezika s ruskog na engleski osnovan Savjetodavni odbor za automatsku obradu jezika koji je 1966. obustavio financiranje svih projekata strojnog prevođenja što je zadalo ozbiljan udarac razvoju UI-e i neuronskih mreža (Poibeau, 2017, 31).

Jedan od posljednih snažnih udaraca koji je gotovo dotukao razvoj neuronskih mreža dolazi iz knjige pod nazivom *Perceptrons: An Introduction to Computational Geometry* čiji su autori Marvin Minsky i Seymour Papert prvi puta 1969. godine opisali XOR pravilo kao ograničenje jednostavnih neuronskih mreža (Minsky, Papert, 1969, 181-182). XOR pravilo opisuje situacije kada rezultat operacije iznosi 1 ukoliko su dva broja različita ili kada rezultat operacije iznosi 0 ukoliko su dva broja ista. Jednostavna neuronska mreža koja se sastoji od samo jednog sloja ne može naučiti ovo pravilo zbog čega se kao potencijalna rješenja ovog problema nude neuronske mreže s više slojeva i nelinearne funkcije. Tada neuronska mreža može naučiti složenija pravila kao što je XOR pravilo. No nemogućnost rješavanja ovog izazova izazvalo je skepticizam i stagnaciju u istraživanju neuronskih mreža (Skansi, 2018, 7).

U kasnijim godinama je ideja neuronskih mreža oživljena kroz pojavu i razvoj kognitivizma, kognitivnih znanosti i promjenu znanstvene paradigme na tragu istraživanja Thomasa Kuhna (Kuhn, 1962, 15). Neuronske mreže postale su sve važnije i za razvoj UI-e. Tako je James Lighthill smatrao kako se neuronske mreže uspješno koriste za modeliranje neurobioloških podataka te kako daljnja istraživanja neuronskih mreža moraju biti u skladu s neuroznanošću i modeliraju stvarne neuronske strukture (Lighthill, 1973, 6). Lighthill je kroz zanimljivu metaforu u kojem kroz odnos softwarea i hardwarea opisuje razumijevanje znanstvenih istraživanja. Smatrao je da kao što je hardver samo dio računalnog sustava, tako je i neuronska aktivnost samo dio onoga što mozak radi. S ciljem razumijevanja čovjeka smatrao je da moramo proučavati mentalne procese i međusobne odnose. Interakcija tih procesa temelj je spoznaje, a UI te interakcije mora detaljno proučavati. Prava znanstvena vrijednost UI-e se prema

Lighthillu tako nalazi u razumijevanju i modeliranju tih procesa, a ne u tehnološkim, ekonomskim i drugim koristima koje su često jedino naglašavane (Lighthill, 1973, 31).

Danas bilježimo rastuću upotrebu neuronskih mreža u brojnim područjima, a jedan od prvih uspješnih primjera primjene u odnosu na čovjeka koji je zaprepastio svijet bio je u području igara. AlphaGo, program temeljen na neuronskoj mreži, pobijedio je europskog prvaka u igri GO u listopadu 2015. godine (Silver i sur., 2016, 484). U zdravstvu se uspješno primjenjuju brojne neuronske mreže koje pospješuju zdravstvenu skrb na razne načine. Konvolucijske neuronske mreže zabilježile su izvanredan uspjeh u segmentaciji i klasifikaciji tri vrste tumora mozga: meningiom, gliom i tumor hipofize. U odnosu na rezultate s drugih sedam pristupa klasifikaciji tumora mozga koji su koristili isti skup podataka, metoda poduprta konvolucijskim neuronskim mrežama postigla je najveću točnost[4] klasifikacije tumora s vrijednošću od 97,3% (Díaz-Pernas i sur., 2021, 13). Konvolucijske neuronske mreže pokazale su se učinkovitima u analizi medicinskih podataka zahvaljujući sposobonosti izdvajanja relevantnih značajki iz medicinskih slika, što je dovelo do poboljšane dijagnostičke točnosti, automatizacije radnih procesa i poboljšanog donošenja odluka u zdravstvu (Mienye i sur., 2025, 6). Konvolucijske neuronske mreže omogućile su napredak u otkrivanju bolesti, segmentaciji i multimodalnom snimanju u ostalim zdravstvenim područjima kao što su neurologija, kardiologija, pulmologija, oftalmologija, dermatologija i ortopedija (Mienye i sur., 2025, 3).

No posljedica ovih neupitno uspješnih neuronskih mreža očituje se, kako u primjeni za igru Go, tako i u zdravstvu. Sustavi UI-e koji svoj rad baziraju na neuronskim mrežama i algoritmima strojnog učenja otežavaju čovjeku da provjeri način rada ili da unaprijed razumije i potvrdi donesene odluke. Zbog toga se kao jedna od glavnih kritika neuronskih mreža često izdvaja kako su neuronske mreže ponašaju kao „crne kutije“, budući kako u većini slučajeva ne postoji zadovoljavajuće objašnjenje rada niti predviđanja neuronskih mreža (Benitez, i sur., 1997, 1156). Način rada neuronskih mreža često se opisuje kao netransparentan (Singh i sur., 2025, 9), opak ili neproziran (Søgaard, 2023, 224), ili subsimboličan (Roth i sur., 2025, 1), zbog nedostatka razumijevanja njihovog načina rada.

[4] Točnost i preciznost su najvažniji kriteriji koji su utemeljeni još u procjeni kvalitete analitičke metode (Karnes, March, 1993, 1420). Razlikujemo „točnost“ od „preciznosti“ na sljedeći način: točnost je oznaka za mjerenje ili rezultat koji nam govori koliko je određeno predviđanje sustava UI-e blizu prave ili stvarne vrijednosti. Preciznost je oznaka mjerenja međusobne konzistentnosti, označava prosječnu blizinu predviđanja, neovisno o tome koliko se određeno predviđanje sustava UI-e razlikuje od stvarnih vrijednosti.

2.5. „Crna kutija“

Različiti sustavi UI-e, posebno oni temeljeni na strojnom i dubokom učenju, dizajnirani su za analizu velikih skupova podataka, pronalaženje obrazaca „skrivenih“ u podacima i prijedlog predviđanja za određenu situaciju (npr. odluka u medicinskom postupku ili odobravanju kredita). Problem se nalazi u tome što iz različitih razloga često ne znamo kako ili zašto je sustav UI-e došao do predloženog rješenja (Suresh, 2021, 1). Usprkos tome, sa sve većim opsegom upotrebe neuronskih mreža u svakodnevnom životu otvara se odnos između stroja i čovjeka kojega je potrebno detaljno propitati. Stroj postaje poput proročišta, a čovjek njegovim predviđanjima mora vjerovati (Nature Editorial, 2016, 437).

Valja napomenuti kako postoje različiti oblici neuronskih mreža i tehnika strojnog i dubokog učenja. Što više slojeva izračuna sustav UI-e primjenjuje na ulaz to je izazovnije ponuditi objašnjenje postupka koji se odvija u „crnoj kutiji“. Jednostavni modeli poput stabla odlučivanja (engl. *decision tree*) koji su transparentni i lako razumljivi pružaju uvid u način donošenja predviđanja jer se temelje na jasnim pravilima ili granama koje vode do određenog predviđanja. Kate Saenko smatra kako takve neuronske mreže zapravo nisu „crne kutije“ jer se radi o skupu pravila na temelju kojih čovjek može lako razumjeti što neuronska mreža radi te što mrežu čini inherentno objašnjivom. No duboke neuronske mreže koje izvode izračune miljardi brojeva su previše složene kako bismo razumjeli pravila ili objasnili način donošenja predviđanja (Savage, 2022).

Ukoliko situaciju pokušamo objasniti jednostavnim jezikom uz primjer iz pravne prakse, možemo pratiti sljedeći primjer. Zamislimo sofisticirani sustav UI-e koji pomaže sucima odrediti optimalnu kaznu za višestrukog počinitelja kaznenog djela. Sustav UI-e koristit će sve dostupne podatke o počinitelju, okolnostima kaznenog djela i veliku bazu podataka sa sličnim slučajevima koji uključuju kazne, ponašanje u zatvoru i rehabilitaciju nakon izvršenja kazne zatvora. Dok bi ljudski stručnjak trebao mnoštvo sati za proučiti navedene podatke, sustav UI-e to može učiniti u nekoliko sekundi. Iako znamo da sustav UI-e traži određen uzorak u podacima, na primjer, ukoliko se radi o inteligentnom počinitelju bez stalnog zaposlenja veća je vjerojatnost za ponovno počinjenje kaznenog djela. No ne znamo točno na temelju kojeg uzorka je sustav UI-e došao do predviđanja niti kojim točno koracima. To izaziva zabrinutost jer u pravosuđu, kao i u ostalim ljudskim djelatnostima težimo transparentnosti (Brožek, 2024, 430). Možemo zamisliti da smo odlučili koristiti sustave UI-e za rješavanje prilično jednostavnih pravnih slučajeva kako bismo smanjili opterećenje sudaca i ubrzali pravne postupke. No pitanje je bismo li bili zadovoljni prihvaćanjem presuda koje dolaze od strane UI-

e bez obzira na sve posljedice ili bismo radije trebali barem osnovno znanje o tome kako sustav UI-e dolazi do svojih odluka. Sustavi dubokog učenja mogu učiti, ali nam ne mogu reći zašto su točno nešto naučili. Isto tako sustavi dubokog učenja mogu dodijeljivati određene vrijednosti, ali ne mogu objasniti zašto je dodijeljena određena vrijednost.

U literaturi se prilikom definicije „crne kutije“ nalaze određene razlike koje naglašavaju određeni aspekt. Christopher Molnar tako definira crnu kutiju s naglaskom na nerazumijevanje unutarnjih mehanizama rada: „Model „crne kutije“ je sustav koji ne otkriva svoje unutarnje mehanizme. U strojnom učenju, „crna kutija“ opisuje modele koji se ne mogu razumjeti gledanjem njihovih parametara (npr. neuronska mreža)“ (Molnar, 2022, 11). Paula Boddington pak naglašava neprozirnost na sljedeći način: „“Crna kutija“: U računarstvu, znanosti i inženjerstvu, sustav koji se razmatra u smislu svojih izlaza i ulaza; u umjetnoj inteligenciji, općenito se odnosi na sustav u kojem može biti teško ili možda nemoguće znati što se točno nalazi u „crnoj kutiji“.“ (Boddington, 2023, 502). S druge strane, u literaturi se pojavljuje sintagma „bijela kutija“ koju kao kontrapost Ajay Thampi definira kao: „Modeli bijele kutije su inherentno transparentni, a tumačenje unutrašnjosti modela je jednostavno.“ (Thampi, 2022, 15). Neki znanstvenici ističu trend udaljavanja od sustava UI-e koji rade na principu „crne kutije“ prema sustavima UI-e koji rade na principu „bijele kutije“, posebice u području zdravstva, bankarstva, vojske kako bi stručnjaci i zainteresirani laici mogli dobiti objašnjenje za predviđanja koja dolaze od strane sustava UI-e (Loyola-González, 2019, 154096). Taj trend ukazuje na posljedicu koja naglašava potrebu za objašnjivim i točnim sustavima UI-e koja se do sada istraživala u sektoru bankarstva za procjenu kreditnog statusa klijenta (Martens i sur., 2007, 1466).

William Rapaport, donosi minimalno 4 razloga zbog kojih se u kontekstu principa „crne kutije“ može razviti nepovjerenje prema predviđanjima UI-e i upotrebi sustava UI-e:

1. kriteriji donošenja odluka koji su kodirani u sustavu UI-e, bilo od strane njegovog programera ili od strane programa strojnog učenja koji je razvio te kriterije iz testnih slučajeva;
2. sami testni slučajevi;
3. sam računalni program;
4. podaci na kojima se temelji određeno predviđanje (Rapaport, 2023, 402).

Ukoliko pretpostavimo kako su testni slučajevi prikladni, podaci točni, a sustavi UI-e ispravno dizajnirani, ključno pitanje postaje: kako i zašto algoritam donosi odluku? Kriteriji za donošenje

odluka se danas najčešće ugrađuju u sustave UI-e putem ljudskog programera ili metodama strojnog učenja. No u oba slučaja ishodi i metode su oblikovane od strane ljudi, izravno ili kroz podatke koje su ljudi prikupili i obradili (Rapaport, 2023, 402).

No ostaje pravo pitanje što se događa kada sustav UI-e radi autonomno prema principima „crne kutije“. Tada se postavlja pitanje: može li i treba li sustav UI-e objasniti svoja predviđanja? Osim toga, važno je pitati trebamo li slijepo vjerovati u točnost sustava UI-e koji rade na principu „crne kutije“ ili bismo zakonski trebali zahtijevati objašnjivost njihovih predviđanja?

Sustavi UI-e temeljeni na dubokom učenju, poput AlphaGo-a, donose iznimno točna predviđanja, ali često ne mogu objasniti kako su došli do tih predviđanja. Uspjeh sustava temeljenih na dubokom učenju možda može biti dovoljan dokaz njihove pouzdanosti, no raste potreba za dodatnom regulacijom ovakvih sustava koji posebice u kontekstu zdravstva pri donošenju odluka zahtijeva veću objašnjivost i odgovornost (Bayor, 2025, 1).

Specifičan etičko-epistemološki izazov koji proizlazi iz fenomena „crne kutije“ u literaturi se pronalazi pod izrazom „praznine u odgovornosti“ (engl. *responsibility gap*) (Lang, 2023, 1) koji opisuje situacije u kojima nitko nije moralno odgovoran za čin ili djelovanje koje dolazi od strane sustava UI-e (Taylor, 2025, 1695). Praznine u odgovornosti se opisuju kao sitacije do kojih dolazi kada postoji očiti rascijep u uobičajenom pripisivanju moralne krivnje ili odgovornosti kada sustav UI-e automatizira kognitivni rad koji inače obavlja čovjek te počini pogrešku (Lang, 2023, 1). Praznine u odgovornosti su proučavane u kontekstu autonomnih vozila u svemiru (Matthias, 2004, 176), dijagnosticiranja karcinoma pluća (Matthias, 2004, 177) ili upotrebe robotskih sustava u vojne (Sparrow, 2007, 62). Na temelju praznina u odgovornosti, izrađene su i različite strategije prema kojima bi se moralna odgovornost trebala drugačije distribuirati što bi dovelo do smanjenja rascijepa odgovornosti (Schur i sur., 2025, 6) ili dobrovoljno raspodijeliti između strana u postupku, preuzimajući određenu moralnu odgovornost za stvari za koje, strogo govoreći, nisu krivi (Lang, 2023, 3).[5]

Neprozirnost sustava UI-e je svakako jedan od izvora praznina u odgovornosti s obzirom kako ni tvorci sustava UI-e često ne mogu razumjeti unutarnji rad sustava UI-e što onemogućuje

[5] Takvi primjeri su vidljivi u slučajevima „prisilne supererogacije“ koja na prvi pogled može zvučati kao oksimoron s obzirom na to da se supererogacija općenito odnosi na moralne činove „izvan poziva dužnosti“. Koncept, posuđen od Schloma Cohena, odnosi se na okolnost u kojoj agent mora izvršiti ili supererogatorni čin ili besprijekornu, ali ipak pogrešnu radnju. Prisilna supererogacija se razlikuje od obične supererogacije, gdje agent zadržava izbor koji predstavlja moralnu osrednjost ili tek dopuštenu radnju između hvalevrijednih i kažnjivih djela. Dakle, slučajevi prisilne supererogacije moraju pokazivati tri parametra: (1) izvršavanje pohvalne radnje je posebno hvalevrijedno; (2) neizvršavanje nije kažnjivo; i (3) neizvršavanje je pogrešno (Cohen, 2015, 1006).

utvrđivanje moralne odgovornosti za posljedice koje proizlaze iz rada sustava UI-e. Napredak koji je vidljiv u sustavima OUI-e, koji pružaju veće znanje o tome kako rade sustavi UI-e, mogao bi pomoći u smanjenju praznina u odgovornosti. Prema Baumu i suradnicima, donositelji odluka će lakše moći premostiti praznine u odgovornosti nakon što mu bude dostupno objašnjenje predviđanja sustava UI-e (Baum i sur., 2022, 14).

Neki znanstvenici ističu kako se izazov „crne kutije“ mora hitno riješiti (Castelvecchi, 2016, 21). U kontekstu zdravstvene skrbi posebno je važno doći do razumijevanja i identifikacije obrazaca koji predviđaju odgovor karcinoma na određeni lijek na temelju obrazaca koji se stvaraju tisućama gena ili predvidjeti prognozu karcinoma pluća analizom mikroskopskih slika - sve bez razumijevanja ili identificiranja zašto ili kako su ti obrasci važni (Xu, Shuttleworth, 2024, 53). Moralne dileme prisutne u kontekstu zdravstvene skrbi često određuju život ili smrt pacijenta, a potreba za smanjenjem kliničke nesigurnosti uz standardizaciju skrbi pomoću alata digitalnih tehnologija potiču razvoj i integraciju sustava UI-e u svakodnevnu zdravstvenu praksu (Lang, 2023, 1).

Europska unija je u okviru Opće uredbe o zaštiti podataka proglasila pravne zahtjeve „prava na informacije“, koji se inicijalno zvao „pravo na objašnjenje“. To znači kako osobe čiji su interesi predmet predviđanja sustava UI-e imaju pravo da im sustav UI-e objasni određeno predviđanje. Očito je kako to predstavlja izazov za sustave UI-e koji rade na principu „crne kutije“ (Bartneck i sur., 2021, 36). No postoje i oni koji nisu zabrinuti zbog metoda strojnog učenja koje se ne mogu objasniti te smatraju kako se takvi sustavi mogu empirijski testirati. Sve dok predviđanja odgovaraju stanju stvari u stvarnosti, za njih nije važno ukoliko se ne mogu razumjeti ili objasniti kako funkcionira u praksi (Fok, Weld, 2024, 317).

Kao što neki autori tvrde, u situacijama kada se preciznost i empirijska potvrda sustava UI-e može postići kroz kliničku validaciju, upotreba netransparentnih sustava UI-e može biti etički i klinički prihvatljiva (London, 2019, 20). Ipak, u visoko rizičnim situacijama u kojima se odlučuje o životu i smrti, i koje bi mogle biti predmet sudskih procesa, razumljivo je kako je moguće putem objašnjenja određenih predviđanja tražiti dodatne informacije, što se potencijalno može proširiti zahtjevom za opravdanjem odluka. Ključni element moralnog djelovanja nije samo činiti pravu stvar, već opravdavati radnju koja je počinjena. Moralno opravdanje ne može se temeljiti na „nerazumljivoj“ crnoj kutiji. Zbog toga su u tijeku istraživanja koja se tiču OUI-e koja nastoje generirati objašnjenja zašto neuronske mreže donose odluke koje donose (Wachter i sur., 2017, 2). Moguće je kako će upravo takva

istraživanja s vremenom omogućiti strojnom učenju i dubokom učenju stvaranje odgovarajućih objašnjenja za predviđanja koje donose.

2.6. Objašnjiva umjetna inteligencija: analiza i razgraničenje pojmova

Od početaka razvoja UI-e postojala je povezanost između mogućnosti objašnjenja što dokazuje rani primjer računalnog programa pod nazivom „Advice Taker“ kojeg je predložio McCarthy 1959. godine kao „programa sa zdravim razumom”. McCarthy se zalagao za transparentan pristup razvoja UI-e u kojem bi ljudi mogli razumjeti način rada sustava UI-e, a ponašanje sustava UI-e trebalo je imati mogućnost prilagodbe prema potrebama (McCarthy, 1958, 78). Prema Vaishaku Belleu, to je jedan od ključnih koraka za uspostavu koncepta objašnjenja sustava UI-e u kojemu se kao ključ razumijevanja UI-e predlaže sposobnost zdravog razuma (Belle, 2017, 5117). Knjiga Petera Jacksona *Introduction to expert systems* značajno je utjecala na razvoj ideje objašnjivosti s obzirom kako ekspertni sustavi, za razliku od sustava UI-e koji se temelje na „crnoj kutiji“, moraju ponuditi razumljiva predviđanja ljudima. Između ostalog, Jackson je predložio mogućnosti za dizajn sustava UI-e koji pruža objašnjenja za svoja predviđanja što je jedan od temelja OUI-e (Jackson, 1998, 9).

Oznaku za OUI na engleskom jeziku „XAI“ (što označava Explainable Artificial Intelligence) prvi puta su ju upotrijebili van Lent i suradnici 2004. godine (van Lent i sur., 2004, 900). Sintagma „Explainable Artificial Intelligence“ je u kasnijim interpelacijama poprimila razne oblike i oznake koje ćemo donijeti u tekstu koji slijedi.

Valja istaknuti kako OUI može omogućiti odgovorniju i učinkovitiju potrebu UI-e u raznim područjima i primjenama (Sharma i sur., 2024, 2) te poboljšati jasnoću i razumljivost (Islam i sur., 2022, 28). No, objašnjivost nije nužna u svim ljudskim djelatnostima i svim situacijama u kojima se primjenjuju sustavi UI-e, a u pojedinim situacijama kao što su: predetaljna objašnjenja ili nesklad između objašnjenja i znanja korisnika što može biti uzrokom kognitivnog preopterećenja (Ai i sur., 2021, 716). Dosadašnja istraživanja su pokazala kako se potreba za objašnjivošću smanjuje u situacijama kada sustavi UI-e rade s visokom razinom autonomije bez ljudske intervencije. U takvim slučajevima se naglasak s objašnjivosti procesa donošenja odluka sustava UI-e premješta na osiguravanje njegove pouzdanosti i performansi. Ova perspektiva sugerira kako u potpuno autonomnim sustavima, posebno onima s uspostavljenim povjerenjem i evidencijom performansi, objašnjivost možda nije ključni zahtjev (Taylor, 2025, 1695). Potreba za objašnjivošću u autonomnim sustavima UI-e je smanjena zbog

minimalnih ili nepostojećih posljedica i velikog povjerenja kojeg imaju takvi sustavi. Zbog sve veće primjene sustava UI-e s povećanim posljedicama i rizikom za ljudske živote zbog donošenja odluka u financijskom sustavu, kaznenom sustavu, zdravstvu i ostalim područjima (Karimi i sur., 2023, 2) došlo je do intenzivnog razvoja OUI-e.

Od velike je važnosti reći kako su brojni znanstvenici kroz istraživanja naglasili kako je za konceptualizaciju razumijevanja i evaluaciju objašnjivosti potrebno koristiti znanja iz područja humanističkih i društvenih znanosti jer razumijevanje i zadovoljstvo ponuđenim objašnjenjima predviđanja proučavaju upravo te znanosti (Langer i sur., 2021B, 6; Langer i sur., 2021A, 167; Speith i sur., 2024, 4). Ponekad je razumijevanje čak potrebno kako bi se zadovoljile potrebe određenog objašnjenja jer se ljudski nadzor nad sustavom može osigurati samo kada određeni stručnjaci razumiju način rada sustava UI-e (Speith i sur., 2024, 4). Zbog toga je potrebno provoditi daljnja istraživanja u području humanističkih i društvenih znanosti kako bismo unaprijedili konceptualizaciju razumijevanja i evaluaciju objašnjivosti.

Poznata je rasprava o uspostavljanju ravnoteže između točnosti i objašnjivosti sustava UI-e o čemu ćemo više govoriti u nadolazećem tekstu (Hammann, Wouters, 2025, 1; Babic i sur., 2021, 285). Ovdje je važno napomenuti kako zahtjev za objašnjivošću sustava UI-e raste u slučajevima kada točnost predviđanja više nije dovoljna, naročito u situacijama kada skupovi podataka za treniranje ili testiranje nisu reprezentativni za stvarne uvjete (Petkovic, 2022, 6). Znanstvena otkrića u prirodnim znanostima podrazumijevaju kako sustave UI-e koristimo s ciljem zaključivanja o uzočnim vezama iz podataka opažanja ili postizanje novih znanstvenih uvida. Računarstvo tako ne zamjenjuju postojeće prirodne znanosti kao što su matematika i fizika već ih unapređuje i proširuje za 21. stoljeće što unapređuje proces znanstvenog otkrića (Brunton, Kutz, 2022, 10) U situacijama kada ne možemo uspostaviti određene veze sa znanjem kojeg su do sada prirodne znanosti ustanovile, javlja se određena nesigurnost te nam je potrebna dodatna mogućnost za potvrditi vjerodostojnost rezultata do kojih smo došli putem predviđanja – u suprotnom visoka točnost predviđanja može sakriti lažne korelacije (Roscher i sur., 2020, 42211). U takvim situacijama objašnjivost osigurava znanstvenu vrijednost rezultata predviđanja (Roscher i sur., 2020, 42200). Važnost objašnjivosti u odnosu na točnost svakako ovisi o potrebama korisnika koji je u interakciji sa sustavom UI-e i rizicima kojima se korisnik može izložiti pri upotrebi sustava UI-e. Interakcije sa sustavima UI-e zahtijevaju veću razinu objašnjivosti modela, a veća nesigurnost predviđanja potiče ljude da više žele komunicirati sa sustavom UI-e. Nesigurnost predviđanja odnosi se na percipiranu razliku između informacija koje su potrebne i onih koje su dostupne za određeno predviđanje (Chapman, 1997, 196). Kada

je nesigurnost predviđanja zadatka veća, rezultati predviđanja troškova pokreću više pitanja i istraživanja o tome zašto sustav UI-e daje određene rezultate, što objašnjivost čini važnijom (Hammann, Wouters, 2025, 3).

Osim toga, rezultati istraživanja iz različitih područja, kao što su znanost (Baron, 2025, 1), industrijski sektor (Nikiforidis i sur., 2025, 145), financijski sektor (Černevičienė, Kabašinskas, 2024, 2), zdravstvo (Mohale, Obagbuwa, 2025, 9), pokazali su kako tehnike OUI-e mogu značajno poboljšati transparentnost i povjerenje u primjenu sustava UI-e zbog pružanja objašnjenja predviđanja. Uobičajeno se tvrdi kako je objašnjivost nužna za razvoj određene vrste povjerenja u UI-u o čemu ćemo detaljnije govoriti u nadolazećem poglavlju (Baron, 2025, 1).

Literatura u zdravstvenom kontekstu ne obiluje radovima koji propituju liječnike, zdravstvene djelatnike i tvorce politika o korisnosti postojećih objašnjenja i preferencija koja bi voljeli vidjeti da posjeduju objašnjenja u budućnosti (Lakkaraju i sur., 2022, 1). No, u zdravstvenom kontekstu, razumijevanje predviđanja sustava UI-e može imati ozbiljne posljedice po zdravlje pacijenata zbog čega je potrebno omogućiti objašnjenje predviđanja sustava UI-e i dodatne informacije zdravstvenim djelatnicima i pacijentu (Mohapatra i sur., 2025, 3).

Danas su neuronske mreže postale jedan od najsnažnijih alata u rješavanju netransparentnih izazova. Prvi pokušaji neuronskih mreža usmjereni na transparentnost odluka u svrhu otklanjanja pogrešaka počinju sredinom 1990-ih godina. Primjerice, 1992. godine Craven i Shavlik predstavili su nekoliko metoda za vizualizaciju numeričkih podataka, poput površina odlučivanja (Craven, Shavlik, 1992, 399) ili Hintonovih dijagrama (Craven, Shavlik, 1992, 409). Važno je napomenuti kako su metode za vizualizaciju numeričkih podataka posebno korisne u zdravstvu (Sarvamangala, Kulkarni, 2022, 2).

Uz to, znanstvena literatura OUI-u često stavlja u kontekst interakciju između čovjeka i računala (engl. *human-computer interactions*) stavljajući naglasak na potrebe korisnika za kojeg su namijenjena objašnjenja (Kim i sur., 2024, 1; Haque i sur., 2023, 2; Liao, Varshney, 2021, 1). OUI se često spominje i u kontekstu kognitivnih znanosti gdje, na primjer, može propitivati kako osobine ličnosti i kognitivne pristranosti utječu na obradu objašnjenja i njihovu učinkovitost (Bertrand i sur., 2022, 80). U kontekstu društvenih znanosti OUI može pomoći pri razumijevanju načina na koji ljudi jedni drugima pružaju objašnjenja te koja društvena očekivanja mogu biti uključena u obradu objašnjenja (Borrego-Díaz, Galán-Páez, 2022, 490).

No i nekoliko desetljeća nakon McCarthyeva pokušaja još uvijek ne postoje dovoljno jasni principi i kriteriji objašnjivosti sustava UI-e na temelju kojih bismo u medicini mogli zadovoljiti etičke principe informiranog pristanka pacijenta.

Filozofska rasprava o kvaliteti objašnjivosti prikazana je u radovima autora Lauren Ross i Jamesa Woodwarda (Ross, Woodward, 2023) te Erasmusa i suradnika (Erasmus i sur., 2021, 834-836).

Znanstvena literatura o UI-i izazov „crne kutije" često smješta u kontekst objašnjivosti i OUI-e (Guidotti i sur., 2018, 1; Miller, 2019, 2; Arrieta i sur., 2020, 10; Vilone, Longo, 2021, 89; Brožek, 2024, 427). Rasprava oko kvalitete objašnjivosti i značenja koje se dodjeljuje objašnjenjima je još uvijek otvorena (Jain, Wallace, 2019, 7; Wiegreffe, Pinter, 2019, 3; Fresz i sur., 2025, 438) i vezana je uz ideju „potreba korisnika" i razvoj etički usklađenih sustava UI-e (Alpsancar i sur., 2025, 3015). Otvoreno se razmatraju i razni pristupi i tehnike, kao što su naknadna objašnjenja, metode transparentnosti modela i interaktivne vizualizacije, koji izazov „crne kutije" pokušavaju riješiti na različite načine s ciljem objašnjenja ponašanja sustava UI-e (Mathew i sur., 2025, 1; Rudin, 2019, 206; Ribeiro i sur., 2016, 1135). Pa ipak, OUI je doživjela značajan rast u posljednjih nekoliko godina zbog široko rasprostranjene primjene strojnog učenja i dubokog učenja, koje je dovelo do razvoja izuzetno preciznih točnih sustava kojima nedostaju objašnjivost i transparentnost (Paliwal i sur., 2025, 2). Bibliografski prikazi potvrđuju kako se tema OUI počela sve više istraživati od 2018. godine, a 2022. godine broj objavljenih radova o OUI-i činio je čak 43,7% do tada što ukazuje na eksponencijalni porast interesa za OUI-u u tom razdoblju (Sharma i sur., 2024, 1).

OUI se primjenjuje u domenama koje uključuju donošenje visokorizičnih odluka gdje su ljudski nadzor i intervencije važne za razvoj povjerenja u digitalne tehnologije (Laux i sur., 2023, 15) kao što su agrokultura (Ryo, 2022, 258), zdravstvo (van Voorst, 2023, 560), kibernetička sigurnost (Atzmueller i sur., 2023, 153), industrija (Ahmed i sur., 2022, 5035)., pravo, (Richmond i sur., 2024, 4).

Iako rezultati pojedinih autora sugeriraju kako su znanstvenici identificirali brojne pojmove i detaljno opisali zahtjeve koje objašnjenja i OUI treba zadovoljiti kako bi ga krajnji korisnici lako razumjeli (Vilone, Longo, 2021, 89), čini se kako još postoji velik prostor za napredak. U pregledima radova znanstvenici su utvrdili kako je u ukupno 18.254 radova koji su objavljeni u razdoblju od 2015. do 2025. godine manje od 1% radova (njih 253; 0,7%) koji govore o OUI-i pružajući empirijski dokaz o ljudskoj objašnjivosti. To znači da je kod samo 253 rada uključen

način koji provjerava ljudsku komponentu u procjeni koliko su metode UI-e zaista razumljive ljudima (Suh i sur., 2025, 1). Takvi podaci upozoravaju na nerazumljiva objašnjenja i nedostatak pažnje koje je potrebno uložiti kako bi se objašnjenja evaluirala na temelju ljudskih povratnih informacija. Bibliografski prikazi također su identificirali 5 glavnih dimenzija utjecaja OUI-e na korisnike: povjerenje, transparentnost, razumljivost, upotrebljivost i pravednost (Bahalul Haque i sur., 2023, 1).

Međutim, unatoč rastućem broju istraživanja u području OUI-e kao i golemom znanju razvijenom oko koncepta objašnjivosti, ne postoji jedinstveni konsenzus o kriterijima i metodologiji evaluacije objašnjenja, te se često koriste različiti pristupi koji otežavaju usporedivost rezultata (Kim i sur., 2024, 2).

Brojni znanstvenici upozoravaju kako nemogućnost razumijevanja rada UI-e utječe na odnos povjerenja prema netransparentnim sustavima UI-e (von Eschenbach, 2021, 1607; Hassija i sur., 2023, 47; Mitchell, 2025, 1). S obzirom na sve veću učestalost uključivanja i delegacije zadataka sustavima UI-e, izazov povjerenja prema primjeni UI-e postaje sve važniji. Filozofska analiza povjerenja pokazat će zašto je transparentnost važna za razvoj povjerenja. Do izgradnje povjerenja u primjeni UI-e može doći ukoliko istodobno izgrađujemo i jačamo pouzdane sustave UI-e (von Eschenbach, 2021, 1608).

Alex DeGrave i Joseph Janizeki su s ciljem izgradnje pouzdanih sustava UI-e uspjeli otvoriti „crnu kutiju“ koristeći mape zasićenosti (engl. *saliency maps*), što je jedna od čestih metoda OUI-e (Rajpurkar, 2017, 5; Mitani i sur., 2020, 23).

Mape zasićenosti ističu dijelove ulaznih podataka koji najviše utječu na odluke modela, što omogućava istraživačima i korisnicima da razumiju koje značajke sustav UI-e smatra ključnima prilikom donošenja predikcija. Time su povećali objašnjivost neuronskih mreža što može pomoći inženjerima unaprijediti sustav UI-e kojeg su stvorili ili im pomoći doći do novih rješenja za određene izazove kao što su otkrivanje potencijalnih pogrešaka ili pristranosti (Yanagawa, Sato, 2024, 2). U zdravstvenom okruženju, tehnike OUI-e mogu pomoći razumjeti zašto je sustav UI-e postavio određenu dijagnozu i podržati liječnika u uspostavljanju dijagnoze (Savage, 2022). No potrebno je zadržati oprez jer se prednosti OUI-e mogu ostvariti samo ako su objašnjenja koja daje OUI razumljiva i provjerljiva, a postoje i primjeri u kojima mape zasićenosti zbog vlastitih ograničenja nisu pružile objašnjive rezultate kakvi su se očekivali (Zhang i sur., 2025, 2).

Općenito, sustav UI-e možemo smatrati objašnjivim ako ga ljudi mogu razumjeti putem pojednostavljenog prikaza (Brožek, 2024, 428). Objašnjivost se može shvatiti kao karakteristika sustava UI-e koja čovjeku omogućuje rekonstruirati zašto je određen sustav UI-e došao do predstavljenih predviđanja (Amann i sur., 2020, 2). Međutim, definicija objašnjivosti nije u potpunosti jasna u računarskoj zajednici zbog nedostatka matematičke strogosti (Roscher i sur., 2020, 42203).

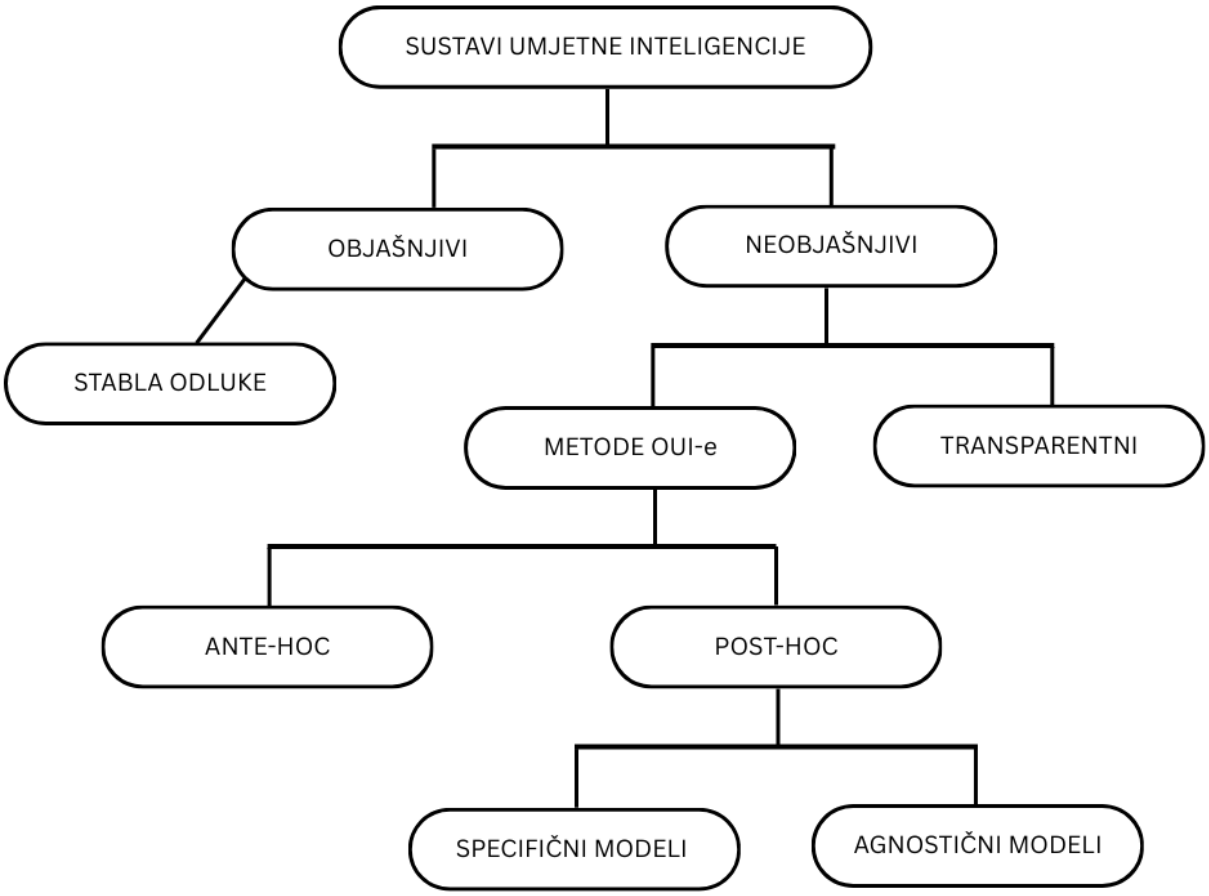
Objašnjenja liječnicima i zdravstvenom osoblju mogu pomoći pri razumijevanju rezultata predviđanja do kojih je došao sustav UI-e u zdravstvu, kao što to čine primjeri sustava OUI-e u radiologiji ili patologiji (Sharma i sur., 2024, 2). Ali objašnjenja mogu biti korisna pri povećanju transparentnosti i razumijevanju sustava temeljnih na UI-i za otkrivanje prijevara (Aljunaid i sur., 2025, 1) i upravljenje rizicima (Wildi, Hadji Misheva, 2022, 1). Razvoj autonomnih vozila također može imati koristi od vizualnih objašnjenja u obliku karte pažnje koje ističu dijelove slike na koje sustav UI-e obraća najviše pažnji pri donošenju odluka (Cultrera i sur., 2023, 10). Na takav način OUI može omogućiti odgovorniju i učinkovitiju upotrebu UI-e u raznim područjima i primjenama.

No važno je napomenuti kako objašnjivost ima mnogo aspekata i, nažalost, terminologija objašnjivosti nije dobro definirana. Područje istraživanja OUI-e razvilo je raznu terminologiju, motivaciju, pristupe i kriterije vrednovanja (Speith, 2022, 2246). Kako broj metoda OUI-e raste, istraživači trebaju taksonomiju kako bi razumjeli određene rezultate, usporedili metode i odabrali onu pravu za određeni kontekst uporabe (Schwalbe, Finzel, 2024, 3063-3088).

Taksonomija OUI-e predstavlja klasifikaciju i sistematizaciju različitih pristupa i metoda koje se koriste za objašnjavanje predviđanja i ponašanja sustava UI-e, posebno onih temeljenih na dubokom učenju i drugim složenim sustavima. Cilj OUI-e je učiniti sustave UI-e transparentnijim i razumljivijim ljudima, što je ključno za razvoj povjerenja u primjenu UI-e u zdravstvu, razvoj etičkih teorija i zakonodavstva (Markus i sur., 2021, 2). Taksonomski se sustavi UI-e najčešće razlikuju na transparentne (prozirne) i netransparentne (neprozirne) sustave, te daljnja taksonomija razlikuje specifične pristupe za model i agnostičke pristupe za model, kao i *post-hoc* pristupe objašnjivosti. Pojam *post-hoc* odražava činjenicu kako takav pristup pruža objašnjenje predviđanja nakon što je obuka sustava UI-e završena, stoga takav pristup ne utječe niti se miješa u proces obuke, već samo revidira postojeći sustav UI-e kako bi procijenio njegovu kvalitetu (Belle, Papantonis, 2021, 3).

Važno je napomenuti kako postoje različita objašnjenja kao što su tekstualna, vizualna, lokalna, objašnjenja primjerima, objašnjenja pojednostavljenjem te objašnjenja relevantnosti značajki (Arrieta i sur., 2020, 12).

U nastavku ćemo pokušati sažeti glavna zapažanja i tehnike OUI-e iz literature. Nećemo uvijek moći pokriti cijeli raspon dimenzija koje se razmatraju u ovom odjeljku, ali one služe kao vodič za sljedeća razmatranja.



Slika 2. Taksonomski prikaz sustava UI-e[6]

Taksonomija OUI-e izgrađena je na principima objašnjivosti i osnovnim pitanjima koja je potrebno postaviti prije nego se izgrađuju sustavi OUI-e.

Osnovna pitanja na koja pokušavamo odgovoriti pružanjem objašnjenja su:

[6] Za više informacija vidi: Saranya, Subhashini, 2023, 9.

1. Kome točno objašnjavamo?

2. Zašto objašnjavamo?

3. Kako objašnjavamo?

4. Kada objašnjavamo?

5. Koje tehnike objašnjivosti koristimo?

Slika 3. Osnovna pitanja na koja pokušavamo odgovoriti objašnjenjima

	Intrinsic metode	Transparentni sustavi	Regresija, stabla odlučivanja, Bayes, skupovi odluka, modeli

Ante-hoc metode		koje ljudi mogu sami razumjeti do određene mjere	temeljeni na pravilima, skupovi pravila i popisi pravila, linearni modeli, hibridni modeli
	Decomposable metode	Sustavi koji se mogu objasniti	Kompozicijski modeli, generativni modeli
	Reasoning-based metode	Sustavi koji koriste prikladne formalizme za predstavljanje, analiziranje i evaluaciju informacija i argumenata kako bi se došlo do rezultata ili predviđanja.	Rasuđivanje temeljeno na argumentima, rasuđivanje temeljeno na slučajevima
Post-hoc metode	*Importance-based* i *Relevance-based* metode	Metode i tehnike koje procjenjuju važnost ili relevantnost ulaznih podataka sustava.	Dijagram parcijalne ovisnosti, Individualno uvjetno očekivanje, analiza osjetljivosti, mapa istaknutosti, širenje relevantnosti po slojevima, SHAP (*SHapley Additive exPlanations*)
	Knowledge Extraction metode	Metode koje imaju za cilj izvući znanje iz sustava „crne kutije“ u oblicima koje ljudi mogu razumjeti	Ekstrakcija pravila, destilacija znanja

	Surrogate metode	Metode koje treniraju surogatni model kako bi se aproksimirali i objasnili rezultati sustava UI-e koji radi na principu „crne kutije“	LIME (*Local Interpretable Model-Agnostic Explanations*), simbolični surogati
	Example-based metode	Metode koje objašnjavaju izlazne rezultate sustava prikazivanjem izbora sličnih ili reprezentativnih slučajeva	Najbliži susjedi, rijetki skup konceptualnih riječi, prototipovi

Tablica 1. Pregled trenutno postojećih metoda OUI-e[7]

Ante-hoc sustavi inherentno su transparentni, dok post-hoc sustavi naknadno objašnjavaju složene „crne kutije“ koristeći tehnike poput SHAP-a, LIME-a i ekstrakcije pravila.

Povijesni razvoj je važno obrazložiti kako bismo objasnili nastanak zadataka koji su specifično kompleksni zbog suradnje između čovjeka i stroja. Ovdje donosimo primjer aktivnosti i zadataka koji su specifični čovjeku, stroju ili se izvršavaju hibridno između čovjeka i stroja (Daugherty, Wilson, 2018, 8).

Donošenje odluka	Treniranje	Provođenje transakcije
Suosjećanje	Objašnjenje	Iteriranje
Stvaranje	Simuliranje	Predviđanje
Prosuđivanje	Koordinacija	Prilagođavanje
Preuzimanje odgovornosti	Pojačanje učenja	Imitiranje

[7] Za više informacija vidi: Speith, 2022, 2244.

	Čovjek nadograđuje UI-u	UI daje čovjeku supermoći	
Isključivo aktivnosti čovjeka	Hibridne aktivnosti čovjeka i stroja		Isključivo aktivnosti stroja

Tablica 2. Suradnja između čovjeka i stroja

Osim navedenog, ostali etičko-epistemološki izazovi OUI-e vezani su uz pitanja: moralne odgovornosti (engl. *responsibility*), pravne odgovornosti (engl. *liability*) i odgovornosti povezane s dužnošću koju netko obavlja (engl. *accountability*) za razvoj UI-e, pristranosti, povjerenja, autonomije, odnosa pacijent-liječnik, pretjeranog oslanjanja na tehnologiju (engl. *overreliance*).

Sistematizacija temeljnih pojmova značajno doprinosi sustavnom razumijevanju i konceptualizaciji pojmova koji su vezani uz UI-e u zdravstvu putem prikaza sveobuhvatne analize terminologije, povijesnog razvoja i metodoloških pristupa. U ovom poglavlju su sustavno definirani ključni pojmovi poput umjetne inteligencije, strojnog učenja, dubokog učenja, neuronskih mreža, objašnjive umjetne inteligencije i koncept „crne kutije“, što omogućuje jasniju i precizniju raspravu o tehničkim, etičkim i pravnim pitanjima te smanjuje rizik od pogrešnih interpretacija i nesporazuma u kliničkoj praksi. Razvoj UI-e prikazan je kroz povijesnu evoluciju koncepta UI-e od mehaničkih računala i Babbageova Analitičkog stroja do modernih sustava UI-e, naglašavajući prijelaz od eksplicitnog programiranja prema autonomnom učenju sustava, što doprinosi razumijevanju novih dimenzija kreativnosti i inovacije u složenim zadacima (primjerice kroz AlphaGo sustav). Ovo poglavlje objašnjava kako su definicije UI-e fragmentirane zbog različitih epistemoloških ciljeva i istraživačkih metoda, te naglašava važnost eksplicitnog definiranja UI-e za razumijevanje njezinih ograničenja, praktičnih mogućnosti i etičkih implikacija, osobito u kontekstu zdravstva, gdje sustavi UI-e ne mogu biti moralno ili pravno odgovorni. Poglavlje posebice ističe doprinos filozofije, psihologije, neuroznanosti, računarstva i kibernetike u oblikovanju konceptualnog okvira UI-e te razlikuje ljudsku inteligenciju od artefakta umjetne inteligencije, naglašavajući potrebu da ljudski agenti preuzimaju odgovornost u primjeni UI-e.

U tom kontekstu, poglavlje dodatno integrira pojmovnu i realnu definiciju UI-e. Pojmovna definicija naglašava kako UI obuhvaća sustave sposobne za ostvarivanje ciljeva putem procesa

koji oponašaju ili modeliraju inteligentno djelovanje, što uključuje mogućnost učenja, planiranja i rješavanja problema. Realna definicija usredotočuje se na stvarne, operativne sposobnosti UI sustava, na to kako se ponašaju u praksi, koje zadatke mogu izvršavati, koja ograničenja imaju te u kojim kontekstima njihova uporaba zahtijeva povećan oprez, osobito u visokorizičnim područjima poput zdravstva, gdje nije poželjno da UI autonomno uči i mijenja ponašanje tijekom izvođenja zahvata.

Poglavlje detaljno pojašnjava principe strojnog učenja, dubokog učenja, uključujući višeslojnu strukturu neuronskih mreža, automatsko učenje transformacija podataka kroz slojeve, prilagodbu težina na temelju funkcije gubitka i povratnog širenja, te razlikuje ove sustave od tradicionalnog strojnog učenja s manjim brojem slojeva i ručno definiranim pravilima, čime doprinosi tehničkom razumijevanju modernih metoda UI-e.

Ovo poglavlje sintetizira teorijske temelje OUI-e, opisuje metode evaluacije objašnjivosti te ističe važnost ljudske komponente u procjeni i interpretaciji rezultata, čime doprinosi razvoju transparentnijih, odgovornijih i pouzdanijih sustava UI-e u domeni visokog rizika. Na taj način, ovo poglavlje doprinosi sustavnom povezivanju teorijskih, metodoloških i praktičnih aspekata UIe, produbljujući razumijevanje UI-e, te doprinosi razvoju etički odgovornih pristupa primjeni UI-e u zdravstvu, olakšavajući suradnju između računarstva, zdravstva i filozofije.

3. Umjetna inteligencija u zdravstvu

Razvoj zdravstvene skrbi jedan je od najvažnijih pokazatelja brige za ljudsko zdravlje. Prema procjenama Svjetskog zdravstvenog foruma u 2021. godini globalna potrošnja na zdravstvo dosegla je 9.800 milijardi dolara, što čini 10,3% globalnog bruto domaćeg proizvoda (Sterlin, 2024). Financijska ulaganja za zdravstvo unutar Europske unije u 2022. godini iznosila su 1.648 milijardi eura što je ekvivalentno 10,4% bruto domaćeg proizvoda (Eurostat, 2024).

U kontekstu Europe, sa svrhom osiguranja sigurnosti sustava UI-e stavljenih na tržište EU te poštivanja vrijednosti Europske Unije, tema regulacije UI-e je aktualizirana zakonskom inicijativom Akta o umjetnoj inteligenciji. Akt je stupio na snagu u srpnju 2024. godine te je pionirski regulatorni okvir usmjeren na osiguranje etične i sigurne integracije sustava UI-e u svim sektorima, uključujući zdravstvo (Europski parlament i Vijeće Europske unije, 2024). Prije toga, u 2019. godini Europska komisija objavila je Etičke smjernice za pouzdanu umjetnu inteligenciju za razvoj i sigurnu primjenu UI-e u skladu s europskim vrijednostima. Između

ostalog, Smjernice napominju kako je za ostvarenje pouzdane UI-e potrebno ispuniti nekoliko različitih zahtjeva kao što su: ljudsko djelovanje i nadzor, tehnička otpornost i sigurnost, privatnost i upravljanje podacima i drugo. Smjernice napominju kako je korisnicima nužno približiti informacije o sposobnostima i ograničenjima sustava UI-e te kako korisnik mora biti informiran na transparentan način da komunicira sa sustavom UI-e. Smjernice uključuju i načelo objašnjivosti i zahtjev za transparentnošću jer omogućuju odgovorno upravljanje sustavima UI-e te izgradnju povjerenja korisnika. Načelo objašnjivosti osigurava razumijevanje predviđanja čime može povećati povjerenje u procese UI-e, dok zahtjev za transparentnošću naglašava važnost nadzora i odgovornosti (Visoka stručna skupina za umjetnu inteligenciju Europske komisije, 2019).

Europska unija je jedno od svojih stručnih izvješća pod nazivom *TechDispatch* Europskog nadzornika za zaštitu podataka, kao neovisnog nadzornog tijela, posvetila temi OUI-e. Izvješće koje je objavljeno 2023. godine istražuje izazove i važnost transparentnosti u radu sustava UI-e, s posebnim naglaskom na poštivanje etičkih normi i ljudskih prava. U izvješću je posebno naglašeno kako OUI nije samo tehnička potreba nego i moralna i pravna obveza, a razvoj OUI-e predstavlja korak prema odgovornijoj i pravednoj upotrebi UI-e u društvu (European Data Protection Supervisor, 2023). Sva rješenja sustava UI-e implementirana u zdravstvu označena su kao visokorizična, a za takva rješenja objašnjivost i transparentnost su posebno važni (Panigutti i sur., 2023, 1139).

Razvoj zdravstvene skrbi dodatno se ubrzava primjenom UI-e. S obzirom na široku primjenu i prednosti koje primjena UI-e u zdravstvu donosi, poput ranog otkrivanja bolesti (van der Schee i sur., 2018, 2) i osmišljavanja personaliziranog postupka (Alum, Ugwu, 2025, 1) postaje sve uobičajenije da sustavi UI-e nadmašuju liječnike u izvršavanju specifičnih zadataka koji uključuju prepoznavanje određenog uzorka. Sustavi UI-e pomažu i pri dijagnosticiranju mnogih bolesti i stanja, kao što su Alzheimerova bolest (Ding i sur., 2019, 456), metastatski rak (Liu i sur., 2017, 1) i kožne bolesti (Liu i sur., 2020, 900).

Takvi sustavi UI-e trenirani su za prepoznavanje uzoraka na temelju tisuća primjera podataka koji dolaze u obliku teksta, govora ili slike (Topol, 2019A, 12). Primjena UI-e u zdravstvu upućuje nas na usporedbu između sposobnosti čovjeka i stroja, pa je tako 2015. godine mogućnost pogreške pri upotrebi sustava UI-e koji prepoznaju uzorke dokazano smanjena ispod 5% koliko iznosi prag ljudske pogreške (Sze i sur, 2017, 8). Osim toga, ljudi zbog nedostatka sna mogu imati lošije dane, brže se umoriti, a emocije im mogu negativno utjecati na koncentraciju, dok strojevi mogu raditi efikasno 24 sata dnevno bez uzimanja godišnjeg odmora

(Topol, 2019A, 12). Način na koji definiramo koncept UI-e utječe na našu percepciju utjecaja UI-e i načina kako taj utjecaj razumijemo. Sustavi UI-e se tradicionalno proučavaju iz tehničke i matematičke perspektive (Martínez i sur., 2022, 3), a sustavi UI-e dramatično su poboljšali svoju izvedbu stjecanjem razvijenih kognitivnih vještina, kao što su percepcija, rasuđivanje i donošenje odluka (Eiband i sur., 2019, 1). Istraživanje koje je ispitivalo definicije sustava UI-e koje su dolazile od strane ispitanika otkrila je kako se više od 80% ispitanika usredotočilo na kategorije vezane uz matematiku, jednadžbe, izračun, logike i formule (Logg, 2017, 29). Česta je pretpostavka kako matematičke i statističke metode nadmašuju ljudsku prosudbu (Dawes i sur, 1989, 12), zbog čega prilikom donošenja odluka ljudi u pojedinim kontekstima mogu više vjerovati preporukama koje predlaže sustav UI-e nego čovjek (Logg i sur, 2018) jer smatraju kako sustavi UI-e predviđanja mogu donositi objektivnije, neutralnije, racionalnije i bez predrasuda za razliku od ljudi (Araujo i sur, 2020, 613; Kahneman i sur, 2021, 275; O'Neil, 2016, 116).

Neka od najnovijih istraživanja najavljuju uspješnu inovaciju i potencijalne trendove koji će promijeniti način na koji razumijevamo ulogu UI-e u zdravstvu. Tu su i suradnici potvrdili kako Googleov sustav UI-e za zdravstvo pod nazivom *AMIE* nije samo bolji od liječnika u postavljanju dijagnoza nego ga i pacijenti preferiraju u interakciji (Tu, 2025, 2). Sustav *AMIE*, temeljen na velikim jezičnim modelima, pokazao je kako može unaprijediti dijagnostičko razmišljanje i preciznost u samostalnom uspostavljanju precizne diferencijalne dijagnoze u usporedbi s liječnikom. Kroz analizu 302 složena medicinska slučaja, sustav *AMIE* je pružio točnu dijagnozu u 177 slučaja (59%) s popisa diferencijalnih dijagnoza, a u 89 slučaja (29%) ta je točna dijagnoza bila prva na popisu što je znatno bolji rezultat od liječnika. Svaki je liječnik prvo dao vlastitu dijagnozu bez pomoći, a zatim je koristio pomoć ili standardnih izvora (poput tražilica) ili *AMIE*-ja. Preciznost *AMIE* je bolja i u kreiranju popisa 10 najtočnijih dijagnoza koji je postigao 59,1% točnih dijagnoza u usporedbi sa 33,6% točnih dijagnoza koje su uspostavili liječnici (McDuff i sur., 2025, 3). Osim toga, sustav *AMIE* je uspoređen s liječnicima primarne zdravstvene zaštite u istraživanju sa 159 simuliranih slučajeva. *AMIE* je pokazao veću dijagnostičku preciznost i superiorne performanse u odnosu na liječnika primarne zdravstvene zaštite u 30 od 32 usporedne kategorije prema odabiru liječnika specijalista te u 25 od 26 usporedne kategorije prema odabiru pacijenata glumaca (Tu, 2025, 1).

Značaj sve preciznijih i uspješnijih metoda UI-e u radiologiji velik je zbog starenja populacije i preporuka za ranijim početkom probira karcinoma (Patel i sur., 2023, 10). Brojni radiološki sustavi UI-e pružaju dodanu vrijednost u kliničkoj interpretaciji, a sustav UI-e pod nazivom

Lunit INSIGHT unaprijedio je uspješnost probira karcinoma dojke. Kroz analizu 211 mamograma stopa probira raka bila je 67,3% za radiologa, 72,7% za Lunit INSIGHT i 83,6% kada su i Lunit INSIGHT i radiolog zajednički tumačili mamogram (Kizildag Yirgin i sur., 2022, 5). Na temelju ovog rezultata očito je kako sustavi UI-e mogu značajno koristiti radiolozima, no potrebno je dodatno istražiti koristi UI-e unutar radiologije. Sustavi UI-e imaju sklonost proizvesti više lažno pozitivnih ili lažno negativnih rezultata, što ograničava njihovu samostalnu sposobnost tumačenja slika. Istraživanje u Nastavnoj bolnici u Sheffieldu NHS Foundation Trust, sustav UI-e pokazao je znatno lošiju izvedbu u stvarnom svijetu pri prepoznavanju plućnih nodula. Softver je od ukupno 5722 analizirana slučaja prepoznao 1120 potencijalno kancerogenih nodula. Međutim, među 943 lažno pozitivna slučaja, u čak 69,9% njih softver je pogrešno označio nalaz kao abnormalan (Maiter i sur., 2023, 4).

Neki autori ističu kako razvoj UI-e u zdravstvu ne donosi revolucionaran razvoj samo u području prakse nego značajno utječe i na obrazovne programe u zdravstvu te zdravstvenu infrastrukturu. Prema Wartmanu i Combsu, medicina prelazi iz „Razdoblja Informacija“ u „Razdoblje Umjetne Inteligencije“ (Wartman, 2018, 1107) što mijenja paradigmu tradicionalnih medicinskih kurikuluma koji su se temeljili na pamćenju opsežne količine informacija. Međutim, s razvojem medicine količina podataka postala je nepremostiv izazov za ljude posebice s obzirom kako se vrijeme udvostručenja zdravstvenih podataka značajno smanjilo. Vrijeme za udvostručenje zdravstvenih podataka iznosilo je 50 godina 1950. godine, 7 godina 1980. godine, 3,5 godine 2010. godine. Predviđanja za 2020. godinu sugerirala su kako će za udvostručenje zdravstvenih podataka trebati nevjerojatnih 73 dana (Densen, 2011, 50). Uz to, neki znanstvenici smatraju kako će razvoj UI-e toliko utjecati na obrazovanje liječnika da će se njihove uloge znatno promijeniti i zahtijevati specifične vještine. U budućnosti će zbog toga kod liječnika biti sve veća potreba za razvojem emocionalne inteligencije i komunikacijskih vještina kako bi prilikom donošenja odluka u obzir mogli uzeti sva fizička i emocionalna stanja pacijenta koja sustavi UI-e neće moći uzeti u obzir (Wartman, 2018, 1107).

Usprkos prednostima koje primjena UI-e može donijeti, sustavi UI-e mogu uzrokovati brojne pogreške. Primjeri pogrešaka vezanih uz primjenu UI-e pojavljuju se u brojnim područjima, a posebno zanimljiva pogreška se pojavila prilikom primjene UI-e za objavljivanje znanstvenih radova. U ovom primjeru, sustav UI-e je pogrešno spojio pojam „vegetativna elektronska mikroskopija“ za kojeg se pretpostavlja da je nastao kao pogreška prilikom digitalizacije starog članka iz 1959. godine u kojem su riječi „vegetativna“ i „elektronska mikroskopija“ bile na

suprotnim stranama dvokolonske stranice. Sporni izraz prvo je primijetio ruski kemičar, pod pseudonimom „Paralabrax clathratus“, u jednom radu iz 2022. godine. Naknadno istraživanje je otkrilo preko 20 radova koji koriste isti izraz, najčešće s autorima iz Irana, uključujući i jednog uvaženog urednika iz Elseviera (Bradley, 2025).

U kontekstu zdravstvenog sustava primjena UI-e može uzrokovati pogreške, kvarove ili slom zdravstvenog sustava. Uzroci pogrešaka mogu uključivati pristranost, halucinacije i neželjeno skretanje pod utjecajem UI-e, što može ozbiljno naštetiti pacijentima (Jongsma, 2024, 1). Neki stručnjaci smatraju kako se zbog pogrešaka do kojih dolazi tijekom procesa donošenja odluka u kojemu sudjeluju sustavi UI-e u ovom trenutku ne možemo u potpunosti osloniti na sustave UI-e (McKendrick, Thurai, 2022). Rosenbacke i suradnici su predstavili tri tipa pogrešaka do kojih dolazi prilikom donošenja odluka u zdravstvu: prvo, pogreška lažne potvrde koja opisuje situacije kada se liječnik i UI slažu, ali su oboje u krivu, drugo, pogreška lažnog sukoba koja opisuje situacije kada liječnik mijenja mišljenje pod utjecajem UI-e koja daje netočno predviđanje, i treće, stvarna pogreška sukoba koja opisuje situacije kada liječnik nije u pravu i ne mijenja mišljenje, usprkos točnom predviđanju UI-e (Rosenbacke i sur., 2024, 1).

3.1. Donošenje odluka u zdravstvu

Donošenje odluka može biti poprilično zahtjevan i iscrpljujuć zadatak. Na primjer, ukoliko je ograničenje brzine postavljeno na 50 km/h, policijski službenici ne moraju mnogo razmišljati o tome da li je netko počinio prekršaj i vozio prebrzo. No, ako pravilo glasi kako se ne smije voziti "nerazumno brzo", policajci će morati više razmišljati i procjenjivati situaciju. Provođenje takvog pravila gotovo sigurno će se razlikovati od slučaja do slučaja. U sličnoj situaciji bismo se našli ako bismo liječnike tražili da daju najbolju moguću prosudbu. Tada će liječnici morati potrošiti mnogo vremena razmišljajući o svakom pojedinačnom slučaju, a njihove prosudbe bi se mogle razlikovati od slučaja do slučaja. S druge strane, ako liječnici dobiju jasne smjernice za odlučivanje o tome ima li pacijent upalu grla, njihova će odluka biti brža i jednostavnija (Kahneman i sur., 2021, 295).

Donošenje odluka u zdravstvu često je složeno i do sada su detaljno istraženi brojni čimbenici koji na njega utječu na razne načine, kao što su: pacijenti, obitelj pacijenta, tip zdravstvene ustanove, liječenje bolesti (Schuler i sur., 2025, 8), kliničke indikacije, kvaliteta života, očekivanja pacijenta (McCracken i sur., 2025, 3), liječničko usavršavanje, dob pacijenta (Kelly i sur., 2024, 813). Donošenje odluka u zdravstvu često je otežano različitim izazovima kao što

su zdravstvene nesigurnosti (Han i sur., 2021, 289), pristranosti (Kunitomo i sur., 2022, 6), dijagnostička, terapeutska ili prognostička nejasnoća (Scott i sur., 2023, 419) ili nepotpuni podaci (Kramer i sur., 2021, 257). Uz to, donošenje odluka u zdravstvu može varirati ovisno o raznim individualnim ili sustavnim čimbenicima, kao što su stupanj boli i patnje pacijenta ili troškovi zdravstvene skrbi, te podaci koji se koriste za olakšavanje donošenja odluka mogu biti proturječni, dvosmisleni ili oskudni (Helou i sur., 2020, 2). Ovakve izazove u donošenju odluka znanstvenici opisuju kao uobičajene izazove koji su svojstveni zdravstvenoj skrbi (Luther, Crandall, 2011, 800).

U svakodnevnici, liječnici moraju donositi odluke usprkos činjenici kako se suočavaju s različitim izazovima, kao što su: nesigurnosti, nejasnoća, pristranosti ili nepotpuni podaci. Kako bi se utjecaj navednih izazova pri donošenju odluka smanjio, posljednih godina je objavljen cijeli niz visokokvalitetnih istraživanja koja su potvrdila kako sustavi UI-e nadmašuju liječnike pri izvršavanju pojedinih zadataka u zdravstvu. Tako su, na primjer, sustavi UI-e koji uče putem nadziranog učenja pokazali iznimne rezultate u interpretaciji medicinskih slika i izradi izvješća koja su kvalitetnija i točnija od pojedinih radiologa (Huang i sur., 2023, 8). Utvrđeni su i primjeri u kojima je ChatGPT4 pokazao veću dijagnostičku točnost kod 100 pacijenata u odnosu na liječnike specijalizante (Hoppe i sur., 2024, 5). Iako sustavi UI-e mogu poboljšati točnost medicinske dijagnoze i pomoći pri interpretaciji medicinske dokumentacije, važno je napomenuti kako takva vrsta uključivanja sustava UI-e donosi kompromise na epistemološkoj i normativnoj razini. Jer iako sustavi UI-e mogu poboljšati točnost medicinske dijagnoze, predviđanja koje dolaze od sustava UI-e koji rade na principu „crne kutije“ donose netransparentna predviđanja što otežava procjenu pouzdanosti medicinske dijagnoze (Grote, Berens, 2020, 205).

Osim toga, kada su za donošenje odluka i precizne dijagnoze potrebni specifično znanje i iskustvo te specifični medicinski uređaji, može doći do različitih rezultata. U slučaju pokušaja procjene endometrijskih lezija 108 ginekoloških kirurga je dramatično drugačije procijenilo broj i lokaciju lezija kod ženskih pacijentica (Buchweitz i sur., 2005, 214). Ginekolozi su procjenu vršili putem digitalnog videozapisa laparaskopije kod tri pacijentice. Jedna pacijentica je bolovala od minimalne endometrioze, druga od blage endometrioze, a treća nije imala endometriozu. Kod prve pacijentice broj otkrivenih lezija se kretao između 0 i 9. Većina liječnika (31%) identificirala je tri endometrijske lezije, iako su histološki potvrđene ukupno dvije lezije. Kod druge pacijentice broj otkrivenih lezija se kretao između 0 i 15. Većina liječnika (18%) identificirala je četiri endometrijske lezije iako je histološki potvrđeno ukupno

5 lezija. Kod treće pacijentice bez endometrioze, 53% liječnika nije otkrilo endometrijske lezije, dok je 47% otkrilo jednu ili više lezija (Buchweitz i sur., 2005, 215).

Osim toga, Tenner tvrdi kako tehnologija nije savršena te ne možemo slijepo vjerovati tehnološkom napretku, već moramo razumjeti njegove društvene posljedice. Osim što tehnologija olakšava život ili rješava određeni problem, često proizvodi neočekivane negativne posljedice koje Tenner naziva „učinak osvete“. Kao primjer učinka osvete tehnologije Tenner navodi uvođenje antibiotika koji je spasio brojne živote, ali i doveo do razvoja otpornijih bakterija (Tenner, 1997, 59). Uz to, napredak u dijagnostici i liječenju bolesti trebao je rezultirati uklanjanjem patnje i produljenjem života, no često se događa da pacijenti dulje žive uz kronične bolesti koje narušavaju kvalitetu života (Tenner, 1997, 47).

Osim sposobnosti objektivnijeg, neutralnijeg, racionalnijeg donošenja odluka koje nisu utemeljene na predrasudama, povećanje učinkovitosti se spominje kao čest razlog za uvođenje široke primjene UI-e (Wenderott, 2024, 1). Uz to, dostupnost velikih količina zdravstvenih podataka, napredak računalne snage i proboji u strojnom učenju ubrzali su razvoj, povećali prediktivnu moć i proširili opseg domene rada sustava UI-e (London, 2019, 15).

No, valja imati na umu i kako se inovacije kratkoročno mogu činiti kao dobit za učinkovitost, ali dugoročno mogu postati izvor neučinkovitosti zbog sustavnih promjena, kao što se to može vidjeti na primjeru vojnih ili drugih tehnologija. Radiokomunikacije, zračna obavještajna služba i novi način zapovijedanja revolucionirali su način borbe u ratu koji je uvijek podložan inovacijama (Rochlin, 1997, 197). Također, komunikacija putem e-pošte zamijenila je pisma i omogućila brzu razmjenu pisanog teksta. No istovremeno, ljudi danas više vremena provode pišući i čitajući e-poštu nego što su ikad trošili na pisma (Newport, 2021, 18). Ovi primjeri bi nas trebali podsjetiti kako čak i ako se UI-a pokaže adekvatnim alatom koji dovodi do povećanja učinkovitosti u užem smislu, potrebno je promotriti ostale sustavne pomake koji mogu poništiti prednosti koje učinkovitost može donijeti kratkoročno. Povećanje dostupnosti preciznih sustava UI-e može, na primjer, rezultirati institucionalnim preporukama za njihovu češću primjenu, što bi na kraju moglo povećati, a ne smanjiti radno opterećenje liječnika bez obzira na prisutnost sustava UI-e (Jongsma, 2024, 1).

Sve veću ulogu u donošenju odluka predstavljat će upotreba sustava UI-e koji rade na principu „crne kutije“ zbog čega povećano oslanjanje na predviđanja UI-e u zdravstvu predstavlja izazov za zagovornike koji predlažu zdravstvo usmjereno na pacijenta (Bjerring, Busch, 2021, 352). Većina današnjih sustava UI-e temelji se na umjetnim neuronskim mrežama za koje se

tradicionalno smatra kako rade na principu „crne kutije“ (Castelvecchi, 2016, 22). To znači kako neuronske mreže ne pružaju korisniku uvid u proces donošenja odluka sustava UI-e, odnosno ne pružaju uvid na temelju kojeg bi prosječnom korisniku bilo jasno zašto je sustav UI-e došao do određenog predviđanja (Hassija i sur., 2023, 47). Sama ideja „crne kutije“ nije novost i prvi put se spominje u priručniku Rossa Ashbyja *Uvod u kibernetiku* iz 1956. (Ashby, 1956, 86), ali i u članku Maria Bungea *Opća teorija crne kutije* iz 1963. (Bunge, 1963, 346) te priručnicima poput *Teorija i računalni model slušnog monoauralnog odvajanja zvuka* iz 1985. (Weintraub, 1985, 15) te *Uvoda u stohastičko modeliranje* iz 1998. (Taylor, Karlin, 1998, 221).

Napredak UI-e počinje od 1950-ih godina pod utjecajem Alana Turinga koji je 1950. godine u knjizi „Computers and Intelligence“ opisao Turingov test s ciljem utvrđivanja sposobnosti oponašanja računala ljudske inteligencije kroz razgovor (Turing, 1950, 433). John McCarthy je 1956. godine definirao UI-u kao znanost i inženjerstvo stvaranja inteligentnih strojeva (McCarthy, 1956, 2). Prvi novinski natpis o stroju koji je mogao hodati, govoriti, vidjeti, pisati, reproducirati se i biti svjestan postojanja objavljen je 1958. godine (The New York Times, 1958). Prvi rad indeksiran u MEDLINE-u, vodećoj bibliografskoj bazi podataka iz područja biomedicine i zdravstvenih znanosti, s pojmom „umjetna inteligencija“ datira u 1951. godinu kada je K. H. Fletcher opisao robota u svom radu *Matter with mind; a neurological research robot* (Fletcher, 1951, 305). Utjecaj UI-e vidljiv je i u popisu najcitiranijih radova u 21. stoljeću u sklopu kojih je među prvih deset radova čak pet njih vezano uz različite metode umjetne inteligencije. Takav utjecaj se ponajviše može zahvaliti širokoj primjeni neuronskih mreža koje nadmašuju druge metode prilikom specifičnih zadataka kao što su prepoznavanje ili označavanje objekata na slikama (Pearson i sur., 2025, 588). Najnovija istraživanja pokazuju kako bi se u nadolazećih 50 godina životni vijek čovjeka u razvijenim zemljama mogao produljiti na 100 do 120 godina, a u tom procesu UI bi trebala igrati jednu od ključnih uloga (Kulkova, 2023, 12).

Zabrinutost zbog nedostatka objašnjenja načina na koji rade sustavi UI-e izravna je posljedica neprozirnosti sustava UI-e, što inženjerima i pružateljima usluga onemogućuje objašnjenje predviđanja i načina rada sustava UI-e. Izraz „crna kutija“ koristi se kao sinonim za neprozirnost. Sustavi UI-e koji su loše izgrađeni, testirani ili planirani mogu dati rezultate koji su štetni ili diskriminirajući za određene skupine ljudi ili pojedince. Za početnu karakterizaciju situacije, možemo prihvatiti termin „medicinska crna kutija“ autora Price II kako bismo opisali zdravstvene prakse u kojima sustavi UI-e temeljeni na „crnoj kutiji“ dominiraju pri pružanju podrške za odlučivanje (Price, 2018, 295).

Kada sustav UI-e radi na principu „crne kutije“ liječnici i zdravstveno osoblje ne mogu razumjeti kako je sustav UI-e došao do određenog predviđanja. Kao rezultat toga, u situacijama kada se pojavljuju određene greške u predviđanjima nemoguće je identificirati potencijalni izvor nedostataka sustava UI-e. To je posebno problematično ukoliko su pogrešno uspostavljena medicinska dijagnoza, prognoza ili terapija bolesti (Babushkina, 2023, 569). Medicinske pogreške uzrokovane sustavima UI-e mogu donijeti negativne zdravstvene ishode pacijentima i negativno utjecati na povjerenje javnosti u primjenu algoritamskih rješenja u zdravstvu (Quinn i sur., 2021, 890). Zdravstveni sustav oslanja se na pretpostavku kako pacijenti mogu razumjeti složene pisane i usmene informacije koje pružaju liječnici i zdravstveno osoblje (Graham, Brookey, 2008, 67). Time se naglasak uvelike stavlja na zdravstvenu pismenost pacijenta od kojeg se očekuje da se može snaći u složenim zdravstvenim procedurama i složenom zdravstvenom sustavu. Negativne posljedice ograničene zdravstvene pismenosti mogu utjecati na zdravstveno stanje, zdravstvene ishode, korištenje zdravstvene skrbi i troškove zdravstvene skrbi (Weiss, 2007, 13). Ukoliko pacijenti ne razumiju zdravstvene informacije, ne mogu poduzeti potrebne radnje koje su potrebne za očuvanje zdravlja ili donošenje odgovarajućih zdravstvenih odluka (Graham, Brookey, 2008, 67). U većini situacija kada liječnik uspostavlja dijagnozu, prognozu ili terapiju bolesti, moći će pružiti objašnjenje za navedeni prijedlog. U takvim situacijama prekršeno je načelo autonomije koje se često navodi kao jedno od četiri središnja etička načela za identifikaciju i rješavanje etičkih izazova u zdravstvu (Beauchamp, Childress, 2001, 57). Na temelju poštovanja autonomije proizlazi kako pacijenti moraju biti informirani kako bi mogli dati svoj informirani pristanak za liječenje (Wolkenstein, 2024, 371).

Osim toga, povećano oslanjanje na predviđanja UI-e prilikom donošenja odluka u zdravstvu jest u sukobu s temeljnim idealima medicine usmjerene na pacijenta. Iako ćemo u kasnijem poglavlju detaljnije analizirati navedeni izazov, osnovna ideja jest kako zdravstvena praksa treba služiti za razvoj informiranog i autonomnog donošenja odluka pacijenta uspostavljanjem stanja zajedničkog promišljanja, dijeljenjem informacija i zajedničkim umom između liječnika i pacijenta (Epstein i sur., 2010, 1491). Razni znanstvenici već su upozoravali na ekstremne rizike od primjene UI-e (Weidinger i sur., 2022, 216), no ne postoji konsenzus o tome kako upravljati takvim rizicima. S obzirom na povećanje sposobnosti i autonomije koji bi trebali povećati utjecaj UI-e, nadolazeći rizici mogu uključivati zlonamjernu upotrebu, društvene štete i gubitak ljudske kontrole nad autonomnim sustavima UI-e (Bengio, 2024, 1).

U kontekstu donošenja odluka, liječnik će svoju odluku temeljiti na radiološkim slikama, kliničkim podacima, anamnezi pacijenta, laboratorijskim nalazima i podacima prikupljenim iz mnogih drugih izvora. Ljudski um ima sposobnost sažeti sve navedeno u racionalnu odluku (Khader i sur., 2023, 1). S druge strane, sustavi UI-e dugo se smatraju sposobnima pomoći liječnicima u određenim zadacima, a neki od njih pokazuju jednake ili bolje sposobnosti od liječnika (Shmatko i sur., 2022, 1028). Zbog toga se ponekad događa da u području zdravstva liječnici i drugi zdravstveni djelatnici prednost daju sustavima UI-e koji mogu objasniti predviđanja, dok programeri i istraživači preferiraju modele koji mogu proizvesti predviđanja što je preciznije moguće. S druge strane, stavljanje preciznosti na prvo mjesto može rezultirati složenijim modelima koje je teško razumjeti (Hatherley i sur., 2022, 7). Budućnost razvoja sustava UI-e i učinkovitog donošenja odluka u zdravstvu može ovisiti o postizanju ravnoteže između objašnjivosti i preciznosti.

Iako je niz autora objavio radove koji detaljno proučavaju etičke, pravne, sociološke i ostale standarde te stvorio smjernice za korištenje UI-e kojih se potrebno pridržavati prilikom donošenja odluka u zdravstvu uz podršku sustava UI-e (Pham, 2025, 18; Beheshti, Kerridge, 2025, 7; Čartolovni i sur., 2022, 4), još uvijek se vodi rasprava o zahtjevima i najboljim načinima korištenja sustava UI-e na moralno odgovoran način. Rastuća upotreba sustava UI-e u zdravstvu znak je kako će tehnički napredak morati biti popraćen odgovornim upravljanjem u svrhu održivosti. Virginia Dignum definira odgovornost kao dužnost odgovaranja za svoje postupke i određeni agent ju ima prije nego što izvrši određeni čin ili radnju. Prema Dignum, kada osoba delegira neki zadatak umjetnom ili ljudskom agentu, rezultat tog zadatka i dalje je odgovornost osobe koja ga delegira (Dignum, 2020, 219). Stoga, kada se zadaci delegiraju sustavima UI-e, ljudski stručnjaci i dalje nisu lišeni odgovornosti.

Brojni radovi ističu ključnu ulogu sustava UI-e u podršci i nadopunjavanju rada liječnika, ali također ističu kako isti sustavi ne bi trebali zamijeniti rad liječnika (Winkler i sur., 2025, 7; Sezgin, 2023, 2; Chustecki, 2024, 8). Poseban naglasak se stavlja na nezamjenjivu važnost kliničke prosudbe i empatije liječnika (Krishna i sur., 2025, 2), suosjećanja i povjerenja liječnika (Kerasidou, 2020, 245), te komunikacijske vještine i terapeutske intervencije zdravstvenog osoblja (Morrow i sur., 2023, 22).

Brojni autori ističu kako je budućnost sustava UI-e multimodalna (Schouten i sur., 2025, 1; Simon i sur., 2025, 303; Parvin i sur., 2025, 1) jer će morati oslanjati na više tipova podataka, slično kao i liječnici koji u praksi primaju podatke iz slike, teksta, govora (Ferber, 2025, 1337). Primjeri multimodalnih sustava UI-e uključuju one sustave koji analiziraju radiološke slike

zajedno s kliničkim podacima (Khader i sur., 2023, 5) ili obrađuju histopatološke podatke s genomskim (Chen i sur., 2022, 13) i tekstualnim podacima (Lu i sur., 2024, 466).

S obzirom da je provedeno mnogo istraživanja koji proučavaju multimodalne ulazne podatke, malo je onih koji su proučavali multimodalne izlazne podatke sa svrhom poboljšanja objašnjivosti sustava i diverzifikacije kliničke upotrebe (Tanida i sur., 2023, 7440). Zbog toga posebno valja istaknuti sustave UI-e koji koriste objašnjive metode sa svrhom tumačenja medicinskih prikaza i stvaranja izvješća iz područja radiologije, patologije i dermatologije kao što je GenMI (Rao i sur., 2025, 888). Takvi sustavi u zdravstvu naglašavaju potrebu objektivnih predviđanja sustava UI-e koja su potkrijepljena obrazloženjima s mogućnošću validacije predviđanja, što je posebno važno za zdravstvo koje obiluje složenim slučajevima u kojima detalji čine bitnu razliku.

3.2. Pouzdanost predviđanja sustava umjetne inteligencije

Jedan od temeljnih izazova vezanih uz interakciju sustava UI-e i čovjeka danas tiče se, ne samo sustava UI-e koji donose pogrešne prosudbe, nego sustava UI-e koji donose pogrešne prosudbe s nevjerojatno visokom pouzdanošću (Christian, 2020, 279). Brojni znanstvenici trenutno proučavaju mogućnost implementacije sustava UI-e u zdravstvo te suradnje zdravstvenog osoblja i sustava UI-e posebice u području radiologije s obzirom kako sustavi UI-e u radiologiji značajno povećavaju točnost i brzinu dijagnostičkog postupka zbog brze analize stotina tisuća medicinskih slika te mogu uzrokovati značajne uštede u troškovima (Muyskens i sur., 2025, 207; Jha, Topol, 2016, E1; Hayashi, 2021, 139). Neki autori tvrde kako pouzdanost predviđanja sustava UI-e nosi ključnu ulogu u izgradnji povjerenja javnosti u primjenu UI-e u zdravstvu jer pogreške ili propusti sustava UI-e nose ozbiljne posljedice za kliničke ishode i iskustvo pacijenta (Quinn i sur., 2021, 890).

Na ovom mjestu je samo korisno napomenuti kako se u hrvatskom jeziku često upotrebljavaju termini „pouzdanost“ i „povjerenje“ ne razlikujući jedan termin od drugog. Ukratko, povjerenje podrazumijeva odnos između dviju strana (onoga tko vjeruje i onome kojemu se vjeruje), u kojem prva osoba polaže povjerenje u drugu. Pouzdanost, s druge strane, označava sposobnost osobe u postizanju cilja za koji je napravljen ili sposobnost alata da postane funkcionalan (Fossa, 2019, 70). U literaturi na engleskom jeziku pouzdanost se najčešće označava engleskom riječju „reliability“. Potter ističe primjer iz radnog odnosa koji oslikava razlike između pouzdanosti i povjerenja. Potter uzima za primjer poslodavca koji krši radnička prava, ali se

dobro odnosi prema zaposlenicima u strahu zbog pravnih sankcija koje mu prijete ukoliko ugrozi radnička prava. Budući da poslodavac, koji inače krši radnička prava, nije učinio ništa neprimjereno svojim trenutnim zaposlenicima, oni ga mogu smatrati pouzdanim, ali ne vjerodostojnim (Potter, 2002, 5). Tako su na primjer, moguće situacije u kojima neku osobu smatramo pouzdanom, ali ne i vjerodostojnom. Zaposlenici možda znaju kako se poslodavac dobro odnosi prema njima zato što se boji sankcija i u tom slučaju poslodavac svoje zaposlenike ne bi mogao izdati (jer mu uopće nisu niti vjerovali), ali bi ih mogao razočarati (Ryan, 2020, 10). Povjerenje onoga kome se vjeruje uključuje određeni stupanj rizika ili ranjivosti od strane onoga tko vjeruje, jer se oslanja na pomoć onoga kome se vjeruje. U kontekstu UI-e, pouzdanost se odnosi na funkcionalna očekivanja koja nisu ispunjena i često se proučavala u kontekstu umjetnih moralnih agenata (Fossa, 2019, 75).

Pouzdanost se određuje kao faktor koji označuje koliko treba vjerovati određenoj osobi ili alatu, a prilikom procjene pouzdanosti analiziramo dosadašnje rezultate, te se putem poklanjanja povjerenja obvezujemo oslanjati na buduće rezultate (O'Neill, 2002, 14). Iako se pouzdanost mjeri isključivo na temelju dosadašnjih performansi osobe ili alata, to nije jedina stvar koja se uzima u obzir kada se nekome poklanja povjerenje. O'Neill smatra kako se izgradnja povjerenja temelji na dodatnim elementima kao što su osobne prosudbe, odnosi, očekivanja rizika kojeg smo spremni prihvatiti (O'Neill, 2002, 15). U zdravstvu pouzdanost sustava UI-e uključuje točnost sustava UI-e, relevantnost za kliničku interpretaciju, pravednost u kontekstu donošenja odluka te prihvaćanje takvog sustava UI-e od strane zdravstvenog osoblja ili pacijenta (Balagurunathan i sur., 2021, 75). Jedan od vrlo čestih instrumenata za procjenu pouzdanosti zove se Cronbachov alfa koeficijent pouzdanosti koji se koristi u znanosti (Zhang i sur., 2023, 6; Wang i sur., 2022, 1332), ali i u zdravstvu (Karaca i sur., 2021, 5-6). Cronbachov alfa koeficijent se koristi kada numeričkom vrijednošću želimo izraziti u kojoj mjeri su određeni test ili skala pouzdano konstruirani kako bi ostvarili svoju svrhu (Taber, 2018, 1273).

Iako se često očekuje nulta stopa tolerancije na pogreške uzrokovane od strane UI-e u zdravstvu, to nije realno, jer sustavi UI-e uvijek moraju balansirati između osjetljivosti i specifičnosti ovisno o slučaju upotrebe. Na primjer, u radiologiji se performanse sustava UI-e podešavaju tako da prioritet bude otkrivanje bolesti, čak ako to znači veći broj lažno pozitivnih predviđanja. Sličan pristup koristi se i u drugim zdravstvenim područjima, gdje je cilj maksimizirati korisnost sustava UI-e unatoč neizbježnim pogreškama (Rubin, 2019, 5). Osim toga, obrasci pogrešaka koje čine sustavi UI-e u zdravstvu razlikuju se od onih koje čine ljudi. Na primjer, u radiologiji prilikom izrade medicinskih snimki razlog za 60 – 70% ljudskih

pogrešaka ima uzrok u neuspješnoj percepciji koja je uzrokovana čimbenicima kao što su suptilnost vizualnih nalaza, nepotpuno pretraživanje cijele snimke i takozvani „sindrom zadovoljstva“, koji opisuje situacije u kojima pronalazak jedne abnormalnosti u radiološkoj snimci smanjuje vjerojatnost da će liječnik pronaći druge abnormalnosti (Degnan, 2019, 840). Nasuprot tome, sustavi UI-e nisu podložni nepotpunim pretragama ili „sindromu zadovoljstva“, ali imaju sklonost učenju pogrešnih korelacija kroz podatke za obuku te mogu isporučiti nedovoljno pouzdane mehanizme objašnjenja pa takve različitosti u obrascima pogrešaka između čovjeka i sustava UI-e mogu dovesti do različitih kliničkih ishoda (Liu i sur., 2022, e384). Uz sve to, postoji i rizik privremenog kvara sustava UI-e zbog neočekivanih problema s ulaznim podacima (pristranost podataka, skrivena stratifikacija, nedostatak generalizacije) ili pristranost automatizacije o čemu svjedoče primjeri iz patologije (Evans, Snead, 2024, 279).

Zbog svega toga, trenutno znanstvenici smatraju kako niti u radiologiji ne smijemo zanemariti ljudske stručnjake koji moraju donositi odluke umjesto sustava UI-e jer kliničke odluke još uvijek zahtijevaju ljudsku prosudbu, interpretaciju konteksta i iskustva ljudskog eskperta zbog čega je ljudski nadzor u procesu donošenja odluka još uvijek potreban (Nakamura i sur., 2024, 9).

Epistemologija, kao jedna od temeljnih filozofskih disciplina, pomaže nam u istraživanju uvjeta, mogućnosti i granica znanstvene spoznaje. Klasična definicija znanja seže još u razdoblje Stare Grčke i Platona koji je znanje definirao kao opravdano istinito vjerovanje. Kako bismo došli do određenog znanja, moramo smatrati određenu tvrdnju istinitom (vjerovanje), moramo biti u pravu oko te tvrdnje (istinito) i moramo imati opravdane razloge za vjerovati u određenu tvrdnju (opravdanje) (Chalmers, 2022, 44). Drugim riječima, epistemologija pokušava pronaći temelje za tvrdnje o znanju unutar okvira računalnih operacija. Metode verifikacije i validacije koje su proizašle iz epistemologije koriste se i danas u suvremenoj znanosti i to na način da se verifikacija odnosi na proces utvrđivanja da sustav UI-e točno predstavlja temeljni matematički model i njegovo rješenje, a validacija na proces određivanja stupnja u kojem sustav UI-e točno prikazuje „stvarni svijet“ iz perspektive namjeravane upotrebe sustava UI-e (Russo i sur., 2024, 1589).

Kako bi se izbjegli scenariji diskriminacije, nejednakosti i pristranosti te ih smanjili na najmanju moguću mjeru, pri donošenju odluka uz podršku sustava UI-e u zdravstvu posebna pozornost se pridaje procesima validacije i verifikacije sustava UI-e. Nejednakosti koje uzrokuje UI mogu biti posebno štetne jer se naizgled rezultati UI čine objektivnima (Bracic i sur., 2022, 1159). Za razliku od ljudi, sustavi UI-e ne mogu imati rasističke stavove, ali ljudi koji stvaraju modele

mogu (Sagona i sur., 2025, 3). Procesi validacije i verifikacije posebno su važni za one sustave UI-e koji su kroz povijest pokazivali diskriminatornu praksu prema određenim društvenim skupinama (Shavers i sur., 2000, 572; Wall, 2006, 349) ili ju još uvijek pokazuju (Averbach i sur., 2023, 1; Fawzy i sur., 2022, 730). Diskriminatorna praksa na taj način se odražava u podacima koji postaju dio njih (Bracic i sur., 2022, 1159).

Metode verifikacije i validacije sustava UI-e sastoje se od nekoliko koraka. Prvo, svrha sustava UI-e mora biti jasno određena kako bi cilj verifikacije i validacije bio nedvosmislen. Drugo, skup podataka za internu i eksternu validaciju mora biti unaprijed definiran, a skup podataka bi trebao sadržavati nepristrane podatke iz „stvarnog svijeta“ što je izuzetno važno s obzirom kako kvaliteta podataka utječe na generalizaciju sustava UI-e. Zatim verifikaciju i validaciju treba provoditi na kontroliran i transparentan način, a rezultati moraju biti detaljno analizirani. U ovom koraku je potrebno identificirati ključne snage i slabosti sustava UI-e te označiti potencijalne uzroke pogrešaka što omogućuje optimizaciju sustava i smanjuje operativne rizike. Konačno, stručnjaci trebaju analizirati učinkovitost verifikacije i validacije (Winkler i sur., 2025, 7), a poželjno ih je usporediti sa standardima koji osiguravaju praksu verifikacije i validacije (Collins i sur., 2024, 6).

Važnost epistemologije postala je sve veća s razvojem svijesti kako sustavi UI-e mogu biti u krivu. AlexNet kao jedna od prvih dubokih neuronskih mreža koja je imala 8 slojeva i uspješno se trenirala na velikom skupu podataka pokazala je kako neuronske mreže mogu znatno nadmašiti klasične metode učenja koje su se primjenjivale za računalni vid (Krizhevsky i sur., 2017, 84). No AlexNet je također pokazala i da može biti krhak. Kao izuzetno moćna neuronska mreža, u stanju je prepoznati obrasce toliko precizno da može klasificirati životinje, automobile i ljude koje nikad prije nije vidio. Ova neuronska mreža pokazala je slabosti jer može halucinirati, to jest, klasificirat će i one slike koje nije potrebno klasificirati i označiti ih kao određeni objekt ili životinju s visokim stupnjem sigurnosti od 99% iako to uopće nije taj objekt ili životinja. Uz to, AlexNet je osjetljiv na zavaravajuće primjere koji upućuju kako neuronska mreža zapravo ne razumije slike na ispravan način, odnosno, ne gleda na njih kao čovjek (Nguyen i sur., 2015, 428). AlexNet tako na jednoj slici može ispravno prepoznati pandu u prvom primjeru. U drugom primjeru, nakon što napravimo minimalne izmjene koje čovjek ne može primjetiti, neuronska mreža je odjednom uvjerena s visokom sigurnošću od 99% da se radi o gibonu. To može biti izuzetno opasno u situacijama kada je sigurnost ključna (Christian, 2020, 279). Na primjer, autonomno vozilo može vidjeti znak STOP koje je malo izmijenjeno naljepnicama i prepoznati ga kao ograničenje brzine do 80km/h. Dakle, iako rezultati sustava

UI-e mogu djelovati pametno, oni zapravo rade na osnovi statističkih uzoraka, a ne zahvaljujući stvarnom razumijevanju.

Zbog toga je jedan od osnovnih zadataka epistemologije unutar UI-e pružiti objašnjenje pouzdanosti i točnosti sustava UI-e ili pružiti uvjete koji trebaju biti zadovoljeni kako bismo vjerovali predviđanjima sustava UI-e. U literaturi se navodi kako se epistemološko povjerenje najčešće temelji na značajkama kao što su transparentnost (Creel, 2020, 569), točnost (Durán, Formanek, 2018, 658) i objašnjivost (Mittelstadt i sur., 2019, 2). Postoje različiti epistemološki pristupi koji nam omogućuju razumjeti zašto je opravdano vjerovati rezultatima računalnih operacija i sustavima UI-e usprkos neprozirnom načinu rada sustava UI-e. Jedan od načina koji koristi metode verifikacije i validacije te naglasak stavlja na pouzdane procedure zove se računalni reliabilizam. Na računalni reliabilizam su najviše utjecali Juan Durán i Nico Formanek pod utjecajem Alvina Goldmana i Boba Beddora (Goldman, Beddor, 2016, 90). U najjednostavnijem obliku, reliabilizam se može izraziti na sljedeći način:

> ako *S*-ovo vjerovanje u *p* u trenutku *t* proizlazi iz *m*, tada je *S*-ovo vjerovanje u *p* u trenutku *t* opravdano.

Pri tome je *S* kognitivni agent, *p* je bilo koja propozicija s istinosnom vrijednošću, *t* je bilo koje zadano vrijeme, a *m* je pouzdan proces. Računalni reliabilizam prilagođava ovu ideju specifičnim potrebama računalnih operacija što rezultira sljedećom definicijom:

> ako S-ovo uvjerenje da *p* u *t* proizlazi iz *m*, tada je *S*-ovo uvjerenje u *p* u *t* opravdano.

Pri tome je *S* kognitivni agent, *p* je bilo koja istinosna propozicija povezana s rezultatima računalne simulacije, *t* je bilo koje zadano vrijeme, a *m* je pouzdana računalna simulacija (Durán, Formanek, 2018, 654).

Postoje i druge metode kojima se proučava pouzdanost računalnih operacija, a jedna od njih je *epistemology-cum-ethics* koja predlaže i unapređuje metodu računalnog reliabilizma putem holističke validacije sustava u 4 kategorije: (1) verifikaciju i validaciju, (2) analizu robusnosti, (3) povijest prethodnih implementacija te (4) stručnost onih koji su razvili ili procjenjuju sustav (Russo i sur., 2024, 1589). U nastavku teksta analizirat ćemo ulogu OUI-e koja poštivanjem etičkih i epistemoloških principa doprinosi pouzdanoj upotrebi UI-e u zdravstvu.

4. Razvoj i primjena objašnjive umjetne inteligencije

Postizanje visoke točnosti predviđanja sustava UI-e jedan je od temeljnih zahtjeva za upotrebu UI-e u praksi i ključno je za razvoj UI-e (Dixon i sur., 2024, 5). Ali sve veća rasprostanjenost sustava UI-e uzrokovala je zabrinutost kod donositelja odluka, znanstvenika i društva zbog nedostatka razumijevanja i ograničenja kojeg imaju sustavi UI-e koji rade na principu „crnih kutija“ (Goodman, Flaxman, 2017, 51). To je potaknulo razvoj OUI-e zbog pretpostavke kako će objašnjenja zauzvrat poboljšati razumijevanje korisnika sustava UI-e i istraživača (Renftle i sur., 2024, 2), a dodatna motivacija za razvoj OUI-e nekih znanstvenika nalazi se u uspostavi novih metoda koji će doprinijeti razvoju odgovornijih, pravednijih i pouzdanijih procesa odlučivanja (Ali i sur., 2023, 4). Potreba za poštenom i pravednom UI-om često je povezana s mogućnošću razumijevanja same UI-e. Drugim riječima, kako bismo došli do pravedne upotrebe UI-e, potrebno je pretvoriti „crnu kutiju“ u „staklenu kutiju“, što je moguće pregledniju. Važno je istaknuti kako se pristup rješavanju izazova objašnjivosti i transparentnosti u većini slučajeva ograničava samo na tehničku domenu i domenu filozofije znanosti, ostavljajući etičke i epistemološke izazove u pogledu UI-e uglavnom zapostavljenima. Kako bismo tome doskočili, u ovom radu ćemo razmotriti kako uključiti vrijednosti i druga normativna razmatranja u kritične faze razvoja procesa, od dizajna i implementacije do upotrebe i etičke analize.

Pojam „objašnjenje“ ima više različitih značenja koja mogu biti dvosmislena, što je spomenuto u prethodnom tekstu, i značenje se može promijeniti u ovisnosti od znanstvene discipline unutar koje proučavamo određeno objašnjenje (Renftle i sur., 2024, 8). Neki autori, poput Andrésa Páeza, ističu kako je svrha objašnjenja omogućiti razumljivost predviđanja relevantnim dionicima, te bismo se umjesto potrage za objašnjenjima i interpretabilnošću trebali usmjeriti na razvoj razumijevanja objašnjenja (Páez, 2019, 448).

Pojava pogrešaka naglašava činjenicu kako složeni sustavi digitalnih tehnologija kao što su sustavi UI-e u zdravstvu mogu dobro funkcionirati samo ako su podržani opsežnom i pouzdanom tehničkom infrastrukturom i stručnošću ljudi poput IT stručnjaka i podatkovnih znanstvenika. Naravno, to je moguće uz filozofsku pretpostavku kako možemo objasniti ljudsko ponašanje, suprotno nekim autorima koji tvrde kako je nemoguće objasniti ponašanja teorijskim konstruktima koji se odnose na unutarnja ljudska stanja poput „razumijevanja", "osjećanja", "znanja", "shvaćanja", "željenja", "trebanja", "vjerovanja", "mišljenja", "očekivanja" (Shermer, 2011, 50).

4.1. Razvoj objašnjive umjetne inteligencije

Kao što smo ranije spomenuli, razvoj neuronskih mreža značajno je pomogao u razvoju UI-e u zdravstvu. Neuronske mreže su posebice pomogle prilikom razvoja računalnog vida pri obradi medicinskih slika, pri obradi prirodnog jezika u domeni podataka elektroničkog zdravstvenog kartona ili pri strojnom podržanom učenju u kontekstu robotske kirurgije (Esteva i sur., 2019, 24). Metode dubokog učenja pokazale su napredak i u poboljšanju preciznosti u radiologiji pri probiru karcinoma dojke što radiolozima smanjuje radnu opterećenost i poboljšava uspješnost probira karcinoma dojke (Yala i sur., 2019, 38). Uz to, metode dubokog učenja mogu biti jednako dobre u dijagnostičkoj izvedbi, ali uz mogućnost potvrđivanja dijagnoze na vanjskim neovisnim podacima koji nisu korišteni pri izradi sustava dubokog učenja što doprinosi većoj vjerodostojnosti dijagnoze (Hannun i sur., 2019, 5). Sustav UI-e treniran na više od 120.000 slika temeljen na dubokom učenju pokazao je visoku osjetljivost i uspješnost u detekciji dijabetičke retinopatije (Gulshan i sur., 2016, E5). Također, neuronske mreže su uspješne i prilikom klasifikacije promjena na koži te razlikovanja karcinoma keratinocita od benignih seboreičnih keratoza i razlikovanja malignog melanoma od benignih madeža (Esteva i sur., 2017, 115). Osim neuronskih mreža, dostupnost zdravstvenih podataka u zadnjih je nekoliko godina pomogla značajno unaprijediti kvalitetu zdravstvene skrbi i istraživanja u području zdravstva. Podaci se prikupljaju ne samo u bolničkim centrima nego i putem mobilnih aplikacija i internet stranica, pa je tako u drugom kvartalu 2020. godine, baza podataka iz područja biomedicine i zdravstva PubMed prikazivala preko 11.000 rezultata za pojam "deep learning" od čega je oko 90% takvih publikacija objavljeno u razdoblju od 2017. do 2020. godine (Egger i sur., 2022, 1). Broj objavljenih znanstvenih radova porastao je u zadnjih nekoliko godina te je 2022. godine dosegao vrhunac od 2286 objavljenih radova, što je 36,3 puta više od broja objavljenih radova 2013. godine i ranije (Yao i sur., 2025, 3).

Literatura koja govori o primjeni UI-e u zdravstvu promiče narativ koji podržava primjenu UI-e u zdravstvu na individualnoj i institucionalnoj razini (Norgeot, 2019, 14; Topol, 2019B, 44; Al Kuwaiti, 2023, 12). Na individualnoj razini, stručnjaci tvrde kako će implementacija sustava UI-e poboljšati donošenje medicinskih odluka što znači kako će odluke o dijagnozama, prognozama i liječenju učiniti bržim i pouzdanijim. Osim toga, zagovornici primjene sustava UI-e u zdravstvu ne ustručavaju se isticati nedostatke liječnika, poput njihove podložnosti kognitivnim predrasudama i činjenju dijagnostičkih pogrešaka (Topol, 2019B, 52). Sustavi UI-

e bi tako mogli kompenzirati slabosti ili čak poboljšati sposobnosti donošenja odluka pojedinačnih liječnika. S obzirom na institucionalnu razinu zdravstvene zaštite, ovdje se govori o neučinkovitosti u tijeku rada, potencijalnom rasipanju resursa, nejednakosti i rastućim troškovima.

Sustavi UI-e za donošenje odluka pružaju podršku u zdravstvu već dulje vrijeme, a brojna istraživanja potvrđuju da veća prediktivna moć poboljšava njihovu uspješnost i točnost u raznim područjima kao što je obrada medicinskih slika, obrada prirodnog jezika, smanjenje radnog opterećenja, uspostavljanje dijagnoza ili klasifikacija promjena na koži. Takvi rezultati potaknuli su značajan porast interesa za još jaču integraciju sustava UI-e u medicinske odluke. S druge strane, stručnjaci su također opravdano započeli propitivati etičke i epistemološke posljedice relevantne za budući razvoj upotrebe UI-e u zdravstvu. Jedno od osnovnih pitanja propituje vrste kompromisa na epistemološkoj i etičkoj razini koji se uspostavljaju prilikom uvođenja sustava UI-e u zdravstvu (Grote, Berens, 2020, 205).

Etički izazovi koji analiziraju upotrebu UI-e u zdravstvu propituje tezu prema kojoj se sustavi UI-e smatraju epistemološkim autoritetima te kako rezultati medicinskog algoritma moraju u potpunosti zamijeniti uvjerenja pacijenta što se označava preempcionizmom (Wolkenstein, 2024, 370). Jedan od razloga za motivaciju istraživanja koji propituju epistemološki autoritet sustava UI-e jest princip „crne kutije“ i algoritamska neprozirnost koji označavaju kako zapravo ne razumijemo način rada sustava UI-e niti možemo objasniti zašto sustav UI-e dolazi do određenog predviđanja (Smith, 2021, 540).

Neprozirnost nije jedini pojam koji se koristi za opisivanje ovog problema. Na primjer, kada je proces donošenja odluka nejasan, može se opisati principom „crne kutije“ (Mukherjee, 2017) ili kao netransparentan način rada (Schiff, Borenstein, 2019, 140). Za potrebe ovog rada, ovi se pojmovi mogu koristiti naizmjenično, ali značenje ostaje dosljedno.

Razumijevanje načina na koji sustav UI-e dolazi do predviđanja predstavlja izazov koji ne obuhvaća samo računalne znanosti, već zahtijeva i suradnju različitih disciplina. Potreba za objašnjivim i odgovornim predviđanjima povezana je s mogućnošću razumijevanja rada UI-e, a ključnu ulogu u tumačenju ovog razumijevanja imaju etika i epistemologija u pogledu izazova UI-e (Russo i sur., 2024, 1585).

Kasnije u tekstu ćemo analizirati možemo li sustave UI-e koji rade na principu „crne kutije“ smatrati epistemološkim autoritetima, no važno je napomenuti kako brojni znanstvenici i

istraživači prema sustavima UI-e koji rade na principu „crne kutije“ zauzimaju skeptičan stav nepovjerenja (Rudin, 2019, 210; Vayena i sur., 2018, 3; Bjerring, Busch, 2021, 365).

Kao što smo spomenuli, računalni program *Advice Taker* čiji je tvorac John McCarthy dokaz je potrage za objašnjenjem još od 1958. godine (McCarthy, 1958, 78). To je jedan od prvih pokušaja da se kao ključ razumijevanja UI-e predloži sposobnost zdravog razuma. No i nekoliko desetljeća nakon tog pokušaja još uvijek ne postoje dovoljno jasni principi i kriteriji transparentnosti i objašnjivosti sustava UI-e na temelju kojih bismo u medicini mogli zadovoljiti etičke predispozicije informiranog pristanka pacijenta. Nedavna istraživanja sve više naglašavaju kako bi sustavi UI-e trebali moći izgraditi sliku svijeta koju podržavaju objašnjenje i razumijevanje, umjesto da samo rješavaju izazove prepoznavanja uzoraka (Lake i sur., 2017, 51).

Sve je to utjecalo na razvoj sustava OUI-e za koje se zadnjih nekoliko godina ponovno razvio interes u istraživanju i primjeni u brojnim područjima nakon što područje OUI-e nije proučavano gotovo trideset godina (Buchanan, Shortliffe, 1985, 331; Chandrasekaran i sur., 1989, 10).

Razvoj OUI-e prati razvoj tri različite generacije sustava UI-e na temelju složenosti sustava UI-e (Ali i sur., 2023, 4). Prva generacija razvoja UI-e započinje kada su Chandrasekaran i suradnici (Chandrasekaran i sur., 1989, 10) te Swartout i suradnici (Swartout, 1981, 397) kao jedni od prvih znanstvenika opisali proces donošenja odluka ekspertnih sustava. Ekspertni sustavi su 1970-ih godina (Zadeh, 1973, 29) potaknuli zanimanje za objašnjenje sustava UI-e, a isto zanimanje održano je kroz 1980-te razvojem neuronskih mreža (Andrews i sur., 1995, 2) te razvojem Bayesovih mreža (Lacave, Díez, 2002, 5). Kasnije u 2000-im su inteligentni sustavi za preporuke bili predmet proučavanja objašnjenja (Cramer i sur., 2008, 468; Herlocker i sur., 2000, 242).

Druga generacija razvoja nastupa početkom 2000-ih u kontekstu snažnijih sustava UI-e čiji je proces donošenja odluka i stvaranja predviđanja toliko kompleksan da ga nisu razumjeli niti stručnjaci iz pojedinih područja (Doyle i sur., 2003, 7). Već se u drugoj generaciji sustava UI-e pojavio rizik nenamjernog donošenja netočnih predviđanja kao rezultata pristranosti ili pogrešnih korelacija u podacima. Ovaj rizik od posebne je važnosti kada su se sustavi druge generacije koristili u visokorizičnim primjenama kao što su autonomna vozila i zdravstvo, gdje jedna netočna odluka može uzrokovati smrt ljudske osobe (Burns i sur., 2014, XII). Sustavi

druge generacije pružaju kognitivnu pomoć stavljajući naglasak na ljudsko znanje i sposobnosti rasuđivanja (Ali i sur., 2023, 4).

Sustavi treće generacije UI-e su najnapredniji sustavi koji su razvili zabrinutost zbog neetične primjene (Resnik, Hosseini, 2025, 1512), nedostatka transparentnosti (Alvarado, 2023, 10), nenamjerne pristranosti (van Assen i sur., 2024, 91). Od 2012. godine većina sustava treće generacije djeluje na principu „crne kutije“, a sustavi OUI-e mogli su se razvijati lakše zbog poboljšane računalne tehnologije i novih koncepata za stvaranje objašnjivih mogućnosti (Ali i sur., 2023, 4). Uz to, razvoj sustava OUI-e dodatno je potaknut novim pravnim propisima i regulativom kao što su Opća uredba o zaštiti podataka ili Akt o umjetnoj inteligenciji (Mathew i sur., 2025, 18). Sustavi UI-e koji se primjenjuju u visokoosjetljivim i visokorizičnim područjima znanstvenih istraživanja, kao što su zdravstvo, biologija, ekonomske znanosti, moraju biti u stanju pružiti objašnjenja predviđanja kako bi omogućili znanstvena otkrića i napredak u daljnjim istraživanjima. Rastući interes za OUI-u u trećoj generaciji sustava UI-e dodatno je motiviran činjenicama kako postoje sustavi UI-e koji krše etičke norme (Tartaro, 2025, 262) ili su korisnici iskazali kako nemaju dovoljno povjerenja za upotrebom određenog sustava UI-e (Orbán, Stefkovics, 2025, 4957). Postoje brojni načini na koji se može povećati povjerenje i transparentnost sustava (Mitchell, 2025, 6), ali najčešće dolazi do konvergentnog pristupa u kojem se povjerenje želi zadobiti putem sljedećih dvaju načina: (1) stvaranjem predviđanja koji kao jedan od kriterija imaju koliko dobro čovjek može razumjeti stvoreno predviđanje u danom kontekstu što se zove objašnjivošću ili interpretabilnošću, te putem (2) eksplicitnog objašnjenja predviđanja što se naziva objašnjenjem (Miller, 2019, 1). Objašnjenja se koriste u raznim specifičnim domenama UI-e kao što su opravdanje ponašanja autonomnih agenata (Mercado i sur., 2016, 411), otklanjanje pogrešaka kod sustava UI-e (Kulesza i sur., 2015, 3), objašnjenje medicinskih predviđanja (Fox i sur., 2007, 37) ili objašnjenje predviđanja klasifikatora (Ribeiro i sur., 2016, 1135).

Važno je napomenuti kako je Agencija za napredne istraživačke projekte obrane (engl. DARPA *The Defense Advanced Research Projects Agency*) pokrenula 2017. godine četverogodišnji istraživački program o OUI-i s obzirom kako su performanse neuronskih mreža postale značajno bolje od ljudskih performansi, ali način rada neuronskih mreža nije bilo lako objasniti. Ovaj istraživački program je pokrenut pod vodstvom Davida Gunninga s ciljem da korisnici sustava UI-e mogu bolje razumjeti, vjerovati i učinkovito upravljati inteligentnim sustavima (Gunning i sur., 2021, 1).

Kao što smo već i spomenuli, razvoj OUI-e potaknut je i zbog nedostatka poštivanja etičkih standarda (Coupland i sur., 2025, 2) i zbog nedostatka povjerenja korisnika (Baron, 2025, 2). Razvoj OUI-e je pokrenut pod pretpostavkom kako će korisnici moći na temelju većeg broja informacija o radu transparentnijeg i objašnjivijeg sustava UI-e lakše razumjeti takav sustav i zbog toga moći razviti povjerenje prema primjeni sustava UI-e. No, potrebno je prije svega jasno odijeliti granice između ključnih pojmova kako bismo mogli nastaviti s prikazom nove tematske cjeline koja govori o razvoju povjerenja u primjenu UI-e u zdravstvu.

4.2. Pojmovne granice objašnjive umjetne inteligencije

U prethodnom tekstu smo približili određene terminološke nejasnoće iz područja UI-e koje se mogu pronaći u znanstvenoj literaturi. U ovom potpoglavlju donosimo prvi izvorni znanstveni doprinos rada koji se nalazi u sustavnom razgraničenju ključnih pojmova objašnjivosti, interpretabilnosti i transparentnosti. Važno je napomenuti kako postoje različiti nazivi za OUI-u koji se pojavljuju u literaturi poput: *Augmented Intelligence, Responsible AI, Explicable AI, Supervised AI*, (Bellucci i sur., 2021, 242-244), *interpretability* i *comprehensibility* (Renftle i sur., 2024, 2), te *explainable machine learning* (Ortale i sur. 2025).

Pa ipak, postoji cijeli niz dodatnih sinonima ili sličnoznačnica koje upućuju na objašnjivost, ali nisu objašnjivost, kao što su *Understandable AI*, *interpretable models*, ili *transparent systems*.

Određena terminološka nejasnoća proizlazi iz pogrešne uporabe pojmova kao što su interpretabilnost i transparentnost koji se često neopravdano koriste kao sinonimi (Rudin, 2019, 208; Doran i sur., 2018, 2). Osim toga, cijeli je niz sličnih izvedenica koje se koriste kako bi naglasili određeni aspekt objašnjivosti kojeg smatraju važnijim pa se tako u literaturi nalazi cijeli niz izraza koji označavaju OUI-u: 1) *Augmented Intelligence*, 2) *Responsible AI*, 3) *Explicable AI*, 4) *Supervised AI*, 5) *Interpretable AI*, 6) *Understandable AI*, 7) *Human AI* (Bellucci i sur., 2021, 242-244), 8) *Informed machine learning*, 9) *Intelligible intelligence*. (Roscher i sur., 2020, 42200). U literaturi se također često spominje izraz *human in the loop* što ne označava izraz explainability, ali postoji poveznica u vidu zahtjeva transparentnosti koje bi sustavi UI-e trebali ispunjavati i pružati dovoljno smislena predviđanja kako bi čovjek njima mogao upravljati i donositi konačne odluke (Lash, 2024, 1).

Bartneck i suradnici tako ističu kako se objašnjivost ne smije izjednačiti s transparentnošću. Iako neki pozivaju na maksimalnu transparentnost kada su u pitanju sustavi UI-e, to možda

neće riješiti problem već bi moglo stvoriti nove nejasnoće i izazove. Maksimalna transparentnost sustava UI-e ne može garantirati razumljivost sustava, ali može stvoriti rizik u odnosu na konkurenciju i ometati daljnja ulaganja u razvoj. Zapravo želimo razumjeti način na koji sustav UI-e dolazi do određenog predviđanja, posebno u primjenama koje imaju etičke posljedice (Bartneck i sur., 2021, 39).

Brojni znanstvenici krenuli su razvijati terminologiju za koju smatraju da posebno ističe određeni aspekt relevantan za razvoj OUI-e kao što su razumijevanje, transparentnost, objašnjivost, eksplicitnost, objašnjenje, globalno objašnjenje, lokalno objašnjenje (Schwalbe, Finzel, 2024, 3063-3088; Bruckert i sur., 2020; Páez, 2019, 445; Adadi, Berrada, 2018, 52147).

Stuart Russell i Peter Norvig (Russell, Norvig, 2021, 729) kao i Ajay Thampi jasno razlikuju pojmove objašnjivost i interpretabilnost (Thampi, 2022, 14). Russell i Norvig ističu kako nam objašnjivost pomaže razumjeti zašto određeni izlazni podaci proizlaze iz određenih ulaznih podataka, dok nam intepretabilnost omogućuje shvatiti zašto određeni ulazni podaci dovode do određenih izlaznih podataka i kako bi se izlazni podaci promijenili kada bismo promijenili ulazne podatke (Russell, Norvig, 2021, 729). Thampi se slaže kako se interpretabilnost odnosi na razumijevanje uzroka i posljedice unutar sustava UI-e i dodaje kako interpretabilnost uglavnom mogu uočiti stručnjaci iz područja računarstva koji izrađuju sustave UI-e (Thampi, 2022, 14). Objašnjivost nadilazi interpretabilnost po tome što nam pomaže da shvatimo na način koji je razumljiv prosječnom čovjeku zašto je model došao do predviđanja. Prema Thampiu, interpretabilnost je građevna jedinica objašnjivosti jer interpretabilnost omogućuje objašnjivost (Thampi, 2022, 15).

Osim nedostatka konsenzusa o kriterijima i metodologiji evaluacije objašnjenja, te različitih pristupa koji otežavaju usporedivost rezultata, postoji i problem višeznačnosti.

Računarska zajednica još nije postigla dogovor o specifičnim kriterijma koji bi se trebali koristiti za određivanje objašnjivosti. A kao što smo rekli, iako brojne studije razlikuju interpretabilnost i objašnjivost, ta se dva pojma često koriste pogrešno naizmjenično (Linardatos i sur., 2020, 2). Objašnjivost i interpretabilnost često su karakterizirane nepreciznim definicijama što može dovesti do obmanjujućih zaključaka (Rudin, 2019, 207). Neki autori ističu kako je objašnjivost usko povezana s konceptom interpretabilnosti te iako je „objašnjiva“ ključna riječ u nazivu OUI, u računarskoj zajednici pojam „interpretabilno“ se više koristi od pojma „objašnjivo“ (Adadi, Berrada, 2018, 52141).

Interpretabilnost je stupanj u kojem čovjek može razumjeti i objasniti uzrok predviđanja kojeg je donio sustav UI-e (Miller, 2019, 8). Objašnjivost se odnosi na stupanj transparentnosti rezultata koje proizvodi sustav UI-e, dok interpretabilnost povećava njegovu korisnost pružajući korisnicima razumijevanje temeljnih koncepata sustava UI-e, omogućujući im tako da shvate logiku i algoritme koji upravljaju sustavom UI-e (Linardatos i sur., 2020, 2-3). Model se smatra objašnjivim ako ljudski stručnjak može jasno razumjeti uzročno-posljedičnu vezu između ulaznih podataka i rezultata predviđanja (Linardatos i sur., 2020, 2). Sustav UI-e se smatra interpretabilnim ako zdravstveni stručnjak može jasno razumjeti kako sustav UI-e funkcionira u cjelini (Burkart, Huber, 2021, 247). Sustav UI-e se smatra transparentnim ukoliko omogućuje izravnu inspekciju unutarnjih detalja sustava UI-e (npr. težinu modela linearne regresije) (Marey i sur., 2024, 3). Objašnjivost se odnosi na stupanj u kojem pojedinci mogu razumjeti unutarnji način rada i mehanizme sustava UI-e. Objašnjivost sustava UI-e pozitivno je korelirana s razinom razumijevanja iskazanom tijekom procesa učenja ili donošenja predviđanja u vezi s tim unutarnjim procesima (Černevičienė, Kabašinskas, 2024, 2). Prema Gilpinu i suradnicima, objašnjivost je temeljni zahtjev uz interpretabilnost, koja je sama po sebi neadekvatna (Gilpin i sur., 2018, 80).

Transparentnost s druge strane označava sposobnost da sustav UI-e bude jasno razumljiv i pregledan korisnicima. Neki autori smatraju kako je transparentnost ključna za izgradnju povjerenja u sustave UI-e u zdravstvu te omogućuje bolje razumijevanje i evaluaciju sustava što je važno za sigurnu i učinkovitu upotrebu sustava UI-e u zdravstvu (Eke, Shuib, 2025, 2004), a neki autori transparentnost smatraju nužnim uvjetom za procjenu sustava UI-e pouzdanim (von Eschenbach, 2021, 1608). Prema Dignum, transparentnost je sposobnost opisivanja, pregledavanja i reprodukcije mehanizama putem kojih sustavi UI-e donose odluke i uče se prilagođavati svom okruženju. Osim toga, transparentnost se odnosi na eksplicitnost i otvorenost u pogledu izbora i odluka koje se tiču izvora podataka, razvojnih procesa i dionika (Dignum, 2019, 54).

Osiguravanje transparentnosti, objašnjivosti i interpretabilnosti važno je zbog poštovanja etičkih normi i smanjenja negativnih posljedica u visokorizičnim situacijama, ali još je važnije razlikovati ove koncepte jedne od drugih.

Ovo potpoglavlje donosi izvorni znanstveni doprinos jer prvi puta sustavno i jasno razgraničuje tri ključna, ali u literaturi često miješana i terminološki neujednačena pojma: objašnjivost, interpretabilnost i transparentnost u području umjetne inteligencije, osobito u kontekstu zdravstva. Iako postoje brojni radovi koji spominju različite sinonime i srodne koncepte, dosad

ne postoji rad koji ih pregledno, dosljedno i metodološki utemeljeno uspoređuje, definira i konceptualno razdvaja. Ovaj doprinos je znanstveno značajan jer uklanja višeznačnost koja je duboko ukorijenjena u literaturi, pokazuje zašto je pogrešna zamjena jednog pojma drugim metodološki i etički problematična te uvodi jasno strukturiran okvir koji omogućuje dosljednu uporabu pojmova. Time rad postavlja temelj za preciznije istraživanje, pouzdaniju evaluaciju OUI sustava i sigurniju primjenu UI-e u kliničkoj praksi — što dosad nije postignuto ni jednim postojećim znanstvenim radom.

4.3. Objašnjiva umjetna inteligencija u zdravstvu

Bez objašnjenja u smislu razloga ili opravdanja za određene odluke u pojedinačnim slučajevima, neki znanstvenici smatraju kako je prepuštanje medicinskog odlučivanja sustavima crnih kutija kršenje dubokih moralnih odgovornosti kliničara. Kako William Swartout kaže, kada se liječnik konzultira sa stručnjakom, liječnik može dovesti u pitanje je li neki faktor uzet u obzir ili kakav je učinak određeni nalaz imao na konačni ishod, a od stručnjaka se očekuje da bude u stanju opravdati svoj odgovor i pokazati da su za njegovo dobivanje korišteni zdravi medicinski principi i znanje (Swartout, 1981, 383).

Međutim, sustavi UI-e koji rade na principu „crne kutije“ uobičajeni su u praksi te se smatra kako predstavljaju problem za povjerenje (Dahlin, 2025, 2). Usprkos činjenici da postoje istraživanja koja tvrde kako bi pacijenti više mogli vjerovati sustavima UI-e jer nisu pristrani i nisu skloni činiti ljudske pogreške (Dalton-Brown, 2020, 118) ili makar činjenici kako sustavi UI-e nisu podložni istim predrasudama, iracionalnim tvrdnjama i nemoralnim tendencijama kao ljudi (Taylor, 2025, 1695), rezultati istraživanja provedenog na 1183 ispitanika pokazuju kako pacijenti još uvijek više vjeruju liječniku, zatim liječniku koji koristi UI-u kao podršku i na posljednjem mjestu vjeruju sustavima UI-e (Riedl i sur., 2024, 3). Prema nekoliko autora, nedostatak objašnjivosti sustava UI-e uzrokuje smanjenje povjerenja u trenutne dijagnostičke modele, na primjer u pogledu primjene OUI-e u probiru karcinoma dojke (Bidwai, i sur., 2023, 154). Metode OUI-e mogu pomoći u generiranju predviđanja i pružanju objašnjenja, omogućujući pacijentima da procijene pouzdanost sustava UI-e koji radi na principu „crne kutije“ (Band, i sur., 2023, 2; Hatherley, 2020, 479).

U zdravstvu donošenje dobrih odluka čini velik dio svakodnevnog rada liječnika koji moraju točno dijagnosticirati bolesti na temelju ograničenih dokaza i u ograničenom vremenu te

odrediti najbolju strategiju liječenja među raznim mogućnostima za pacijenta (Grote, Berens, 2020, 205).

Valja imati na umu kako se procjenjuje da se za vrijeme treniranja jedne neuronske mreže proizvodi oko 284 tone ugljičnog dioksida, što je ogromna količina i usporediva s emisijama nekoliko automobila tijekom cijelog života ili velikim brojem letova preko kontinenta (Strubell i sur., 2019, 3645). Velike količine ugljičnog dioksida uzrokuju kompromise u praktičnom smislu, a ti kompromisi mogu dovesti do mnoštva etičkih izazova koji svoje posljedice ne mogu razumjeti u početku.

Uz to, stručnjaci poput Erica Topola upozoravaju kako se izazovi primjene UI-e u zdravstvu, koji se tiču društvenih i etičkih perspektiva, neće moći riješiti samo naprednom tehnologijom ili algoritmima (Topol, 2019A, 17). Stoga je potrebno povećati svijest i razviti mjere za suzbijanje neželjenih učinaka.

Osim kliničke primjene, sustavi OUI-e u zdravstvu doprinose i reviziji sustava UI-e u radiologiji ili dermatologiji putem mapa istaknutosti (DeGrave i sur., 2021, 610) te putem analize troškova koja unapređuje točnost i isplativost sustava UI-e (Erion, 2022, 1384). Sustavi OUI-e mogu pomoći i pri osmišljavanju terapije protiv raka odabirom optimalnih kombinacija lijekova putem stvaranja objašnjenja predviđene učinkovitosti (Janizek i sur., 2023, 811). Iz perspektive liječnika, objašnjivost je ključna u kliničkom kontekstu jer također omogućuje predviđanje potreba za hospitalizacijom (Bergomi i sur., 2025, 2).

Vrijedi istaknuti kako postoje i drugi sustavi OUI-e koji se koriste sa svrhom unapređenja zdravstvene skrbi pri donošenju odluka. Na primjer, Shickel i suradnici su razvili sustav OUI-e pod nazivom DeepSOFA koji na temelju protoka vremena predviđa ozbiljnost bolesti u bilo kojem trenutku tijekom boravka na intenzivnoj njezi (Shickel, 2019, 2). Kaji i suradnici su izradili sustav UI-e koji pruža objašnjenja na temelju mehanizama pažnje kako bi uspješno identificirali ključne varijable za predviđanje sepse ili infarkta miokarda. Dobivene informacije su služile kao potpora za tumačenje predviđanja koje liječnici koriste pri radu u jedinicama intenzivne njege (Kaji, 2019, 13).

Neki autori smatraju kako će upravo principi „crne kutije“ onemogućiti potpuno usvajanje UI-e u zdravstvu, a koje je potrebno pri optimizaciji i pružanju skrbi prilagođene specifičnim potrebama pacijenta (Khan, Ewuoso, 2024, 227).

Takva razmjena doprinosi transparentnosti zdravstvenih postupaka i promiče autonomiju pacijenta kroz proces informiranog pristanka (Khan, Ewuoso, 2024, 228).

No usprkos činjenicama kako OUI može unaprijediti zdravstvenu skrb, te kako zdravstveno osoblje teško razvija povjerenje prema sustavima UI-e koji rade na principu „crne kutije“, otvoreno je pitanje trebaju li svi sustavi UI-e u zdravstvu biti objašnjivi. Geoffrey Hinton je istaknuo kako liječnici i regulatorna tijela ne bi trebali inzistirati na objašnjivosti jer ljudi ionako ne mogu objasniti način na koji rade većinu stvari na poslu (Simonite, 2018).

Kako bismo bili sigurni da je određeno objašnjenje dobro, ono mora odgovarati primjeru iz stvarnog svijeta koje provjeravamo uz pomoć našeg racionalnog uma. No, naši umovi su sposobni dodijeliti objašnjenja za sve vrste fenomena i općenito nesposobni prihvatiti ideju nepredvidljivosti (Taleb, 2010, 135). S opravdanjem se možemo pitati koja je zapravo vrijednost OUI-e ako su ljudska objašnjenja toliko manjkava? Postoji li rizik da će nas automatski generirana objašnjenja za automatske odluke navesti da previše vjerujemo tim odlukama? Što ako objašnjenja ne otkrivaju očite nedostatke, već su zapravo vrijedna samo koliko i *post-hoc* racionalizacije? Pozitivniji pogled na ovaj izazov opravdanja objašnjenja jest da bi UI-a mogla biti korisna za otkrivanje ljudskih pristranosti. Napor da se bolje razumije što predstavlja valjano objašnjenje za odluku mogao bi se koristiti za procjenu ljudskih odluka, ne samo predviđanja UI-e. A neusklađenosti između ljudskih odluka i predviđanja sustava UI-e mogle bi dovesti do zdravog preispitivanja među ljudima (Lee, 2020, 101).

Uz to, važno je istaknuti kako čak i ako razumijemo način na koji sustavi UI-e rade, nije samo po sebi jasno kako ćemo razumjeti pojedina predviđanja sustava UI-e. Objašnjenje formalno može biti sastavljeno od niza logičkih dedukcija u kojemu svaka dedukcija slijedi pravila logike, a niz završava zaključkom. Međutim, objašnjenje koje završava s milijardom, trilijunom ili kvadrilijunom aritmetičkih operacija nije korisno ljudima. Zbog toga Simon tvrdi kako je cilj objašnjenja smanjiti složenost na razumljive komponente, čime se omogućuje bolje razumijevanje i predviđanje takvih pojava (Simon, 1992, 135). Ukoliko pretpostavimo da su uzrok svega što se događa u svijetu samo osnovni zakoni fizike, reakcije između protona i neutrona, onda bismo na temelju odnosa protona i neutrona mogli objasniti uzrok rata u Ukrajini, ali takvo objašnjenje nam ne bi bilo korisno (Searle, 2010, 12). Na sličan način objašnjenje koje dolazi samo od strane neuronskih mreža ne može činiti zadovoljavajuće objašnjenje.

Osim toga, raspon i brzina ljudske obrade podataka neusporedivo su slabije u odnosu na sposobnosti sustava UI-e kao i način odlučivanja. Racionalnost podrazumijeva donošenje odluka temeljenih na logičkim koracima. Idealno objašnjenje za neku odluku bilo bi kronološki prikaz tih koraka. Međutim, ljudska sposobnost za takvo razmišljanje je ograničena, što znači da često koristimo intuiciju više nego sustavno razmišljanje što je poznato kao „Sustav 1" prema Danielu Kahnemanu (Kahneman, 2011, 25). Slično tome, možemo reći kako neuronske mreže ne donose odluke kroz racionalni proces, već djeluju više poput ljudske intuicije. Razumijevanje mehanizama neuronske mreže ne pruža mnogo uvida u racionalna objašnjenja njihovih odluka. Ako inzistiramo na racionalnim objašnjenjima sustava UI-e, inženjeri će razviti sustave UI-e koji su vrlo učinkoviti u konstrukciji *post-fact* racionalizacije, što može dovesti do situacije u kojoj sustavi UI-e postaju neupitni autoriteti.

Važno je istaknuti kako se trenutno u zdravstvenoj skrbi upotrebljava mali broj različitih sustava OUI-e, usprkos prednostima koje objašnjive metode postižu u razvoju povjerenja. Većina primjena se svodi na korištenje *post-hoc* metoda za objašnjenje medicinskih slika, a glavni naglasak je na relativno jednostavnim metodama LIME i SHAP (Alkhanbouli i sur., 2025, 3).

5. Povjerenje u primjenu umjetne inteligencije u zdravstvu

Kroz povijest je povjerenje igralo jednu od ključnih uloga u razvoju društva (Luhmann, 2017, xii). Povjerenje se često definira kao spremnost na ranjivost u prisutnosti drugih (Mayer i sur., 1995, 712) te se često doživljava kao ključni element učinkovitog vodstva i pravilnog raspolaganja (Zhang, Ma, 2025, 2). Kada je naše povjerenje pravilno utemeljeno, povjerenje u samog sebe obično poboljšava svakodnevni život, povjerenje u bliske osobe poboljšava odnose, a povjerenje u ljude općenito olakšava ispunjavanje životnih ciljeva. Slično tome, populacija ljudi s visokim povjerenjem olakšava napredak gospodarstva i razvoj društva što je neophodno za rješavanje novih društvenih izazova (de Fine Licht, Brülde, 2021, 1982).

Važno je istaknuti kako povjerenje može učiniti ranjivim pojedinca koji nekome poklanja povjerenje. Tako povjerenje može biti definirano kao psihološko stanje koje kompromitira namjeru prihvaćanja ranjivosti na temelju pozitivnog očekivanja o namjerama ili ponašanju drugih ljudi (Rousseau i sur., 1998, 395). Posredovanje povjerenja uvijek uključuje pristajanje na određeni rizik s obzirom na nepredvidljivost namjera u međusobnom odnosu (Fehr,

Fischbacher, 2003, 786). Usprkos tome, društveni život bez povjerenja bio bi nepodnošljiv i najvjerojatnije nemoguć jer nam ipak povjerenje omogućava pospiješiti stanje čovjeka kako ga Thomas Hobbes opisuje u Levijatanu kao usamljenog, siromašnog, gadnog i brutalnog (Hobbes, 1651, 78) u ugodnije, učinkovitije i mirnije biće.

Povjerenje nije lako precizno definirati s obzirom kako se radi o višestruko složenom konceptu o čemu svjedoči brojna literatura koja je objavljena o raznim teorijama i psihološkim temeljima koji utječu na formiranje povjerenja u međuljudskim interakcijama (Cao i sur., 2025, 6-7). Osim toga, postoji niz sličnih riječi koji se u literaturi često pripisuju povjerenju ili ga pobliže označuju kao što su: uzajamnost, empatija, reciprocitet, uljudnost, poštovanje, solidarnost, tolerancija i bratstvo (Newton, 2001, 203). Sličnosti između pouzdanosti i povjerenja također unose određenu terminološku nejasnoću. Pouzdanost i povjerenje nisu sinonimi jer ljudi mogu vjerovati stvarima koje nisu pouzdane. Ljudi često vjeruju uglednim medijima poput *The Guardian*, *The Independent* ili *The Telegraph* za koje se kasnije ispostavilo da su izvještavali nepouzdano zbog izazivanja senzacija kod publike (Horn, 2024, 216-223) ili navodnim metodama koje se koriste u ponudi digitalnih kovanica kako bi se provele financijske prijevare (Hornuf i sur., 2022, 1741). Pouzdanost odgovara na pitanja: „Trebamo li vjerovati?“, dok povjerenje odgovara na pitanje: „Vjerujemo li doista?“ Čini se kako Albert Giubilini i suradnici u članku *Expertise, disagreement, and trust in vaccine science and policy. The importance of transparency in a world of experts* tako čine pogrešku određujući doslovno dvosmisleno pojmove „pouzdanost“ i „povjerenje“ kada tvrde kako „pouzdanost zahtijeva od stručnjaka da budu transparentni“ (Giubilini i sur., 2025, 10). Ono što daje povjerenje našoj znanosti jest koliko je ona pouzdana, a znanost će biti pouzdanija što je više provjerljivija. No ova normativna izjava o tome kakva bi znanost trebala biti je pogrešna, i autori su vjerojatno željeli na deskriptivan način izraziti kako smatraju da bi se stručnjaci i znanstvenici trebali ponašati da razvijaju javno povjerenje u znanost (Hyde, 2025, 3).

Postoje i sličnosti između povjerenja i oslanjanja te povjerenja i (samo)pouzdanja. Oslanjanje se često smatra preduvjetom za povjerenje i taj pojam često koriste politolozi prilikom mjerenja kvalitete rada državnih tijela u javnosti te u ostalim istraživanjima objavljenim kroz proteklih 20 godina (de Fine Licht, Brülde, 2021, 1982). Njemački sociolog Niklas Luhmann smatra kako se razlika između povjerenja i (samo)pouzdanja nalazi u percepciji opasnosti i atribuciji razočaranja. Prema Luhmannu, povjerenje pretpostavlja mogućnost rizika i zahtijeva angažman od strane pojedinca. Kao što je to na primjer, u situacijama kada pojedinac kupuje rabljeni automobil koji se ispostavi kao kvaran ili razočaravajući, narušeno je povjerenje. U situacijama

s negativnim ishodom, razočarenje je povezano s prethodnom odlukom o kupnji rabljenog automobila. S druge strane, Luhmann tvrdi kako je (samo)pouzdanje uvjerenje u mogućnost nošenja s izazovima, bez obzira na vanjske okolnosti. Zbog (samo)pouzdanja je pojedinac uvjeren kako njegova očekivanja neće biti iznevjerena. Na primjer, zbog (samo)pouzdanja pojedinac vjeruje kako će političari pokušati izbjeći rat ili kako se automobili neće pokvariti (Luhmann, 1988, 97). Razlika između samopouzdanja i povjerenja se nalazi u percepciji i atribuciji. Ako pojedinac ne razmatra alternative nalazi se u situaciji (samo)pouzdanja. Ako pojedinac odabere jednu radnju umjesto druge unatoč mogućnosti razočaranja, nalazi se u situaciji povjerenja. U slučaju (samo)pouzdanja, pojedinac će na razočaranje reagirati vanjskom atribucijom (neuspjeh pripisan vanjskim faktorima). U slučaju povjerenja, pojedinac će razmotriti unutarnju atribuciju (neuspjeh pripisan unutarnjim faktorima) i na kraju požaliti svoj izbor povjerenja (Luhmann, 1988, 98).

Koncept povjerenja se u početku najčešće proučavao u ekonomiji i filozofiji kroz perspektivu društvenih odnosa i materijalnih dobara, jer se povjerenje smatralo međuljudskim odnosom temeljenim na iskustvima poput posuđivanja novca i dobivanja povrata te poslovnih odnosa (Uslaner, 2018, 5). Osobe koje ne poklanjaju povjerenje jedni drugima ne mogu surađivati s drugima samo u okviru formalnih pravila i propisa o kojima moraju raspraviti, oko kojih se moraju dogovoriti, o kojima se ponekad mora voditi parnica i koji se ponekad moraju provoditi i prisilnim sredstvima. Rašireno nepovjerenje u društvu stoga može rezultirati svojevrsnim dankom ovisnosti o pretjeranim formalnim pravilima i propisima, „poreza“ kojeg društva s visokim povjerenjem ne moraju plaćati (Fukuyama, 1995, 27). Nedostatak povjerenja je u prošlosti doveo do negativnih ekonomskih posljedica i popratnih društvenih implikacija. Francis Fukuyama u knjizi *Trust: the social virtues and the creation of prosperity* navodi tri jasna primjera u kojima nedostatak međusobnog povjerenja dovodi do negativnih ishoda i lošijih životnih uvjeta za cijelu zajednicu. U prvom primjeru, Fukuyama navodi grad na jugu Italije u kojem je Edward Banfield primjetio za vrijeme 1950-ih kako bogati građani nisu imali dovoljno povjerenja za udružiti se s ostalim građanima s ciljem izgradnje prijekopotrebne infrastrukture grada, poput škole i bolnice. Zbog nedostatka povjerenja svi građani su živjeli slabijom kvalitetom života. U drugom primjeru, Fukuyama navodi kako je nedostatak povjerenja u procjeni zaposlenika od njihovih nadređenih u Francuskoj rezultiralo uvođenjem formalnih pravila koji reguliraju radne odnose. Tako nedostatak povjerenja rezultira birokratizacijom sustava koji ograničava solidarnost među radnicima i otežava mogućnost inovacije. Posljedično, nedostatak povjerenja unutar radne zajednice uvodi dodatne

institucionalne mehanizme kontrole. U trećem primjeru, iz Sjedinjenih Američkih Država, nedostatak povjerenja unutar etničkih zajednica Afroamerikanaca rezultira nemogućnošću ujedinjavanja u ekonomsku zajednicu koja olakšava ekonomski napredak i zajedničke projekte. S druge strane, zajednica Koreanaca s visokom razinom povjerenja i obiteljskih veza uspješno provodi poslovne planove i ostvaruje ekonomski probitak (Fukuyama, 1995, 12).

Fukuyama povjerenje definira kao očekivanje koje nastaje unutar zajednice redovitog, poštenog i kooperativnog ponašanja, temeljenog na zajednički podijeljenim normama od strane drugih članova te zajednice (Fukuyama, 1995, 26). Znanstvena literatura o povjerenju postala je sve opsežnija od 1970-ih godina i kasnije, a povjerenje se često povezivalo s društvenim kapitalom. Tako je Robert Putnam u knjizi *Bowling alone: The collapse and revival of American community* istaknuo važnost koju društveni kapital ima u društvu, posebice u kontekstu osnaživanja homogenih skupina društva te umrežavanja različitih skupina društva (Putnam, 2000, 23). Putnam razlikuje više vrsta kapitala: fizički, ljudski, društveni. Fizički kapital se, prema Putnamu, odnosi na fizičke objekte, ljudski kapital na svojstva pojedinaca, dok se društveni kapital odnosi na veze među pojedincima. Na taj je način društveni kapital blisko vezan s građanskim vrlinama. Društveni kapital tako skreće pozornost na činjenicu kako je građanska vrlina najsnažnija kada je ugrađena u gustu mrežu recipročnih društvenih odnosa (Putnam, 2000, 19). Prema Putnamu, povjerenje među ljudima i njihovo sudjelovanje u društvenim aktivnostima nisu odvojene pojave nego dva različita aspekta istog fenomena kojeg naziva socijalnim kapitalom, a opisuje ga putem mreže odnosa i normi koje olakšavaju suradnju i djelovanje u društvu (Putnam, 2000, 73).

Povjerenje je kao koncept istraživano u brojnim područjima kao što su psihologija (Baier, 1986, 231), političke znanosti (Uslaner, 2018, 3), računalne znanosti (Eke, Shuib, 2025, 2000), ekonomija (Leblang i sur., 2022, 1), sociologija (Rotter, 1967, 651), ekologija (Dietz i sur., 2003, 1907). Suvremena istraživanja o povjerenju su dovela do spoznaje kako transparentnost igra važnu ulogu u komuniciranju povjerenja, posebice kada znanstvenici, novinari ili političari govore u javnom prostoru (Varda i sur., 2025, 470). Usprkos toj važnoj ulozi, transparentnost ipak nije dovoljna kako bismo izgradili povjerenje nego su potrebne dodatne strategije koje minimiziraju rizik da će pojedinci izgubiti povjerenje u stručnjake. Za promicanje povjerenja je uz poštovanje transparentnosti potrebno pokušati s drugim strategijama kao što su aktivno uključivanje u dijalog s javnošću te javno tumačenje rezultata i načina na koji znanost funkcionira (Intemann, 2024, 14).

Nedostatak povjerenja uz mogućnost transparentnosti zbog toga možemo ilustrirati kroz sljedeća dva primjera. Pretpostavimo prvi slučaj u kojem je banka odbila zahtjev za kreditom osobi zbog predviđanja do kojeg je došao sustav UI-e. Pristup vođen načelima OUI-e ima za cilj pružiti toj osobi objašnjenje koje osvjetljava zašto je zahtjev odbijen i omogućuje osobi da ponovno podnese zahtjev za kredit prema točno određenim specifikacijama prema kojima bi zahtjev bio odobren. Sustavi UI-e koji nemaju mogućnost objašnjenja i ne pružaju transparentan uvid u svoj način rada kod podnositelja zahtjeva izaziva negativne posljedice u vidu zabrinutosti te podnositelji zahtjeva posljedično razvijaju nepovjerenje prema takvim institucijama. Do razvoja nepovjerenja ili zabrinutosti ne mora doći ukoliko pruženo predviđanje sustava UI-e sadrži adekvatno objašnjenje za odbijanje zahtjeva kojeg podnositelji zahtjeva očekuju kako bi svoj zahtjev u budućnosti mogli prilagoditi kriterijima specifične banke (Barocas i sur., 2020, 83).

Zamislimo drugi malo detaljniji primjer u kojem klijent traži zahtjev za zajam kod banke. Banke u obzir pri odobrenju ili odbijanju zajma računaju kreditnu sposobnost klijenta kako bi smanjila rizik u poslovanju banke (Chang i sur., 2024 A, 2) ili otklonile mogućnost potencijalnih financijskih prevara (Goyal i sur., 2025, 3). Kreditna sposobnost se najčešće očituje kroz kreditni rejting i kreditnu povijest do kojih banka dolazi putem složenih računalnih operacija i uz podršku sustava UI-e na temelju različitog niza faktora kao što su dob, prihodi, trajanje radnog odnosa, broj članova obitelji klijenta i brojni drugi (Chang i sur., 2024 B, 29). Ukoliko bi klijent zatražio transparentnu analizu njegove kreditne sposobnosti, banka bi mu isporučila datoteku od nekoliko megabajta u kojoj se nalazi opis svih računalnih operacija koje je sustav UI-e napravio kako bi došao do konačnog predviđanja o odbijanju ili prihvaćanju zahtjeva klijenta. Klijentu takva datoteka ne bi značila puno s obzirom da se radi o pukom opisu računalnih operacija na temelju kojeg klijent ne može izgraditi povjerenje u predviđanje banke zbog čega transparentnost sama po sebi nije dovoljna za izgradnju povjerenja (Mitchell, 2025, 2). Pa ipak, autori napominju kako za uspostavljanje povjerenja nije dovoljno samo to da predviđanja budu transparentna (Søgaard, 2025, 314) ili samo pouzdana (Hatherley, 2020, 479), usprkos tome što se tvrdi kako je za uspostavljanje povjerenja ključno da predviđanja budu transparentna (Zerilli i sur., 2022, 1). Potrebno je da predviđanje ispuni određene kriterije kako bi se opseg povjerenja mogao primijeniti na tehnologiju, sustave UI-e i institucije o čemu ćemo govoriti u daljnjem tekstu.

Još jedno značajnije područje u kojem kontekstu se povjerenje istražuje vezano je uz pitanje autoriteta i povjerenja u autoritete (Trench, 2025, 323) ili institucije (Joxhe, 2023, 1). Vodeći

tehnološki lideri danas u kontekstu rasprave o postizanju superinteligencije tvrde kako su upravo oni autoriteti kojima je potrebno vjerovati. Stuart Russell tako u knjizi *Human Compatabile* izvještava o tri primjera u kojima tehnološki lideri sebi pripisuju autoritet kojemu treba vjerovati jer prema njima, sva zabrinutost oko negativnih posljedica razvoja UI-e proizlazi iz neznanja (Russell, 2019, 152). Oren Etzioni, izvršni direktor Allen instituta za umjetnu inteligenciju tako smatra kako svaka nova tehnologija izaziva strah jer ne znamo kako će utjecati na nas. Još od razvoja straha prema mehaničkim tkalačkim stanovima, pa do današnjeg straha od ubojitih robota naše reakcije proizalze iz neizvjesnosti. Kada smo nesigurni, naš um popunjava praznine strahom i maštom, smatra Etzioni (Etzioni, 2014). U drugom primjeru, Erik Sofge stavlja znanstvenike iz područja UI-e koji „znaju bolje“ nasuprot Bill Gatesu i Elonu Musku tvrdeći kako se „pravi znanstvenici“ iz područja UI-e ne brinu da će im se superinteligencija prikrasti, sada ili u budućnosti. Suprotno jezivim pričama, koje na primjer priča Elon Musk, „pravi znanstvenici“ ne poduzimaju panične korake za zaštitu od UI-e, jer oni „znaju bolje“, prema Sofgeu (Sofge, 2015). U trećem primjeru, David Kenny, potpredsjednik IBM-a, u pismu američkom Kongresu navodi kako tek nakon što se stvarno bavite UI-om i kada ju primjenite u stvarnom svijetu poslovanja i društva, tek onda shvaćate kako ta tehnologija ne podržava širenje straha (Russell, 2019, 153). Sva tri primjera pozivajući se na autoritet imaju jednu zajedničku tvrdnju kako istinski znanstvenici iz područja UI-e odbacuju rizike UI-e što nije točno s obzirom da postoji niz autora koji istražuju različite oblike rizika koji dolaze od strane UI-e (Kasirzadeh, 2025, 1975; Sunyaev i sur., 2025, 2; Zanotti i sur., 2024, 4-7; Dung, 2025A, 1177-1178; Douglas, 2025, 4142; Tang i sur., 2025, 3-4).

Iako je povjerenje često proučavana tema, rijetko kada je uspješno precizno definirana varijabla zbog toga što dostupna mjerenja povjerenja nisu dovoljno precizna ni validirana (Nowak i sur., 2025, 380), a povjerenje i nepovjerenje nisu međusobno dovoljno diferencirani (Nowak i sur., 2025, 375). Činjenica kako se o odnosu povjerenja i nepovjerenja raspravlja već gotovo 40 godina od utjecajnog teksta „Trust and Antitrust“ autorice Annette Baier, koja povjerenje uvodi kao moralnu obvezu, nam govori kako je ova tema od izrazite važnosti (Baier, 1986, 247).

Povjerenje u znanost raste kada znanstvenici poštuju znanstvene norme, poput zajedništva, univerzalizma, ravnodušnosti, organiziranog skepticizma, koje je opisao Robert Merton (Merton, 1943, 270), a znanstvenici putem transparentnosti signaliziraju da upravo to i čine (Jamieson i sur., 2019, 19235). Na taj način, ali i u brojnim drugim primjerima transparentnost pomaže utvrditi povjerenje javnosti u znanost. Byron Hyde navodi kako transparentno komuniciranje pozitivnih vijesti osnažuje povjerenje javnosti. S druge strane, transparentno

komuniciraje negativnih vijesti koje su ispod očekivanja javnosti negativno utječu na povjerenje. Hyde navodi primjer vezan uz sukob interesa, sigurnost cjepiva te prešućivanje raznih neuroloških nuspojava cjepiva protiv COVID-19 kod različitih skupina ljudi, ovisno o dobi, spolu, povijesti bolesti i postojećem imunitetu (Hosseini, Askari, 2023, 2), kao i pokrenute tužbe od strane nekoliko desetaka pacijenata protiv tvrtke AstraZeneca zbog krvnih ugrušaka uzrokovanih cjepivom te tvrtke (Dyer, 2023, 1). Nuspojave koje se mogu pojaviti u fazi testiranja eksperimentalnog lijeka ne bi se smatrale negativnom vijesti jer su takvi rizici u fazi testiranja očekivani. Tom logikom, pozitivne vijesti za javnost su one koje potvrđuju pouzdanost, a negativne vijesti su one koje negiraju pouzdanost što je suprotno s pouzdanošću u znanosti. Hyde ovim primjerom želi naglasiti kako postoje specifični izazovi u kojima javnost očekuje da znanost bude savršena što nije moguće. U tim situacijama znanstvenici imaju mogućnost transparentno komunicirati nesigurnosti, sukobe interesa i greške čim će doći do smanjenja povjerenja. No s druge strane, znanstvenici mogu izabrati strategiju održavanja povjerenja javnosti zbog čega mogu selektivno prikazivati ili čak lagati o podacima i nesigurnostima preuveličavajući sigurnost rezultata istraživanja, prikrivajući sukobe interesa ili oblikujući podatke tako da izgledaju jasnije i pouzdanije. Na taj način laž paradoksalno kratkoročno može povećati povjerenje u znanost, ali dugoročno može izazvati katastrofalne posljedice za znanost i zdravstvo (Hyde, 2025, 8).

Povjerenje u zdravstvu je istraživano na brojne načine i prepoznato je kao jedno od temeljnih dobara, uz empatiju i suosjećanje (Kerasidou, 2020, 245). Povjerenje u kompetentnost, razumijevanje i želju zdravstvenog osoblja da pomogne pacijentu pomaže u razvoju povjerenja kod pacijenta (Charon, 2001, 1899; Halpern, 2003, 672-673). Istraživanja su pokazala kako povjerenje pacijenta u zdravstveno osoblje rezultira u višestrukim koristima te uključuje poboljšano zadovoljstvo pacijenata, povećano pridržavanje preporuka liječenja (Kelley i sur., 2014, 6) te poboljšane zdravstvene ishode (Joffe i sur., 2003, 106).

Uz to, dosadašnja istraživanja o povjerenju provedena su puno šire izvan zdravstvenog konteksta te su pokazala kako snažne pozitivne emocije koje ostvarujemo s osobama od povjerenja čine prijateljske odnose nezamjenjivima (Hruschka i sur., 2015, 261), zbog čega se uzrujamo ukoliko osobe od povjerenja umru ili se teško razbole (Dunbar, 2021, 68). Detaljna teorijska analiza koncepta povjerenja nalazi se u članku Karoline Reinhardt koja upućuje na brojne smjernice i detaljno prikazuje kako se povjerenje u literaturi često odnosi na nešto što treba unaprijediti, izgraditi, njegovati, uspostaviti, poboljšati, održavati i podržavati (Reinhardt, 2023, 736).

Specifičan izazov povjerenja tiče se adekvatne kalibracije povjerenja što je postupak kojim čovjek prilagođava svoju razinu povjerenja u skladu s pouzdanošću sustava UI-e (Tomsett i sur., 2020, 2). S ciljem iskorištavanja što većeg potencijala sustava UI-e koji pružaju podršku u procesima donošenja odluka, čovjek mora odgovarajuće kalibrirati svoju razinu povjerenja u sustave UI-e (Kochenderfer i sur., 2015, 296). Odgovarajuća kalibracija povjerenja trebala bi biti prilagođena mogućnostima sustava UI-e, dok bi pogrešno kalibrirano povjerenje trebalo rezultirati narušenom timskom dinamikom (Tomsett i sur., 2020, 3). Osim toga, premalo povjerenje može rezultirati averzijom prema sustavima UI-e (Dietvorst i sur., 2015, 2), a previše povjerenja može rezultirati pretjeranim oslanjanjem na sustave UI-e (Holland i sur., 2024, 604). Iako je potrebno provesti dodatna istraživanja u ovom području, dosadašnja istraživanja upućuju na to kako je veće povjerenje u sustav UI-e povezano s većom vjerojatnošću da će pojedinac koristiti sustav UI-e (Rittenberg i sur., 2024, 2580). Dodatna istraživanja pokazala su kako antropomorfizacija agenata UI-e nije značajno utjecala na kalibraciju povjerenja, a u nekim slučajevima pokazalo se kako antropomorfni agenti mogu dovesti do povećanih očekivanja i većeg otpora na povratne informacije (Reuter i sur., 2025, 408).

Prikazali smo važnost povjerenja, razne situacije u kojima može doći do narušavanja povjerenja, te specifične izazove povjerenja vezane uz transparentnost i kalibraciju povjerenja. Povjerenje se na poseban način proučava s obzirom na razvoj UI-e u svakodnevnici i brojni radovi istražuju povjerenje u primjenu UI-e u raznim područjima (Afroogh i sur., 2024, 2). U kontekstu proučavanja povjerenja u primjenu UI-e, povjerenje se gradi kada određena institucija[8] koja daje povjerenje, to jest davatelj povjerenja, može predvidjeti ponašanje korisnika[9], to jest, onog kome se povjerenje daje kako bi znao odgovara li sustav UI-e željama onoga kome se povjerenje daje (Jacovi i sur., 2021, 629). Povjerenje u sustave UI-e može biti otežano nedostatkom objašnjivosti u sustavima UI-e koji rade prema principu „crne kutije“, a nedovoljna objašnjivost i transparentnost mogu biti jedan od glavnih razloga zašto su uspješna implementacija i integracija sustava UI-e u kliničku praksu rijetke (Yang i sur., 2022, 29). Stoga će oni koji pružaju sustave UI-e, poput pojedinaca ili organizacija, ostvariti puni potencijal UI-e samo ako se uspostavi povjerenje u njezin razvoj, primjenu i korištenje (Thiebes i sur., 2021, 447).

[8] engl. *Trustor*
[9] engl. Trustee

Primjena i korištenje sustava UI-e je blisko povezana uz problem kontrole UI-e o kojoj su do sada pisali Stuart Russell, Nick Bostrom, ali i drugi znanstvenici (Russell, 2022, 19; Russell, 2019, 143-144; Bostrom, 2014, 155; Nyholm, 2023A, 1231). Zbog sve veće sposobnosti u postizanju ciljeva sustava UI-e javlja se zabrinutost kako će sustavi UI-e postati toliko sposobni da u jednom trenutku nećemo moći zadržati kontrolu nad postizanjem ciljeva i uspješno upravljati UI (Russell, 2019, 144). Bostrom smatra kako bismo u jednom trenutku mogli doći do superinteligentnih sustava UI-e koji će postati superučinkoviti u postizanju bilo koji ciljeva kojima oni teže (Bostrom, 2014, 115) što bi moglo dovesti do egzistencijalnih rizika (Ord, 2021, 35). Zabrinutost za kontrolu nad UI-om postojala je i u vrijeme Alana Turinga koji je 1951. godine smatrao kako bi u nekom trenutku strojevi mogli preuzeti kontrolu s obzirom da su metode strojnog učenja puno snažnije od čovjekovih kognitivnih sposobnosti (Turing, 2004, 475). Norbert Wiener je problem kontrole formulirao na sličan način na koji se danas izražava. Wiener je 1960. godine smatrao da moramo biti sasvim sigurni kako stroj ima baš onu svrhu koja je nama potrebna da bismo došli do ostvarenja ciljeva putem određene mehaničke agencije na čiji rad ne možemo učinkovito utjecati (Wiener, 1960, 1358).

Istaknuto rješenje problema kontrole koje je možda i vodeće tiče se usklađivanja vrijednosti o kojemu su pisali brojni znanstvenici (Schwerzmann, Campolo, 2025, 5192; Hellrigel-Holderbaum, Dung, 2025, 4; Russell, 2019, 137; Christian, 2020, 13). Usklađivanje vrijednosti je ideja kako sustave UI-e možemo naučiti specifičnim ljudskim vrijednostima i uputiti ih da te vrijednosti uzmu u obzir prilikom stvaranja predviđanja. Stuart Russell je u kontekstu usklađivanja vrijednosti pisao navodeći problem kralja Mide kao slučaj koji demonstrira kako može doći do katastrofalnih posljedica ukoliko vrijednosti sustava UI-e nisu usklađene s ljudskim vrijednostima. Kralj Mida želio je da se sve čega se dotakne pretvori u zlato što mu se činilo kao sjajna ideja. Međutim, čim su se hrana, piće i njegova obitelj pretvorili u zlato kada ih je dotaknuo, shvatio je kako ono što se događa nije zaista ono što on želi (Russell, 2019, 137). Analogno ovaj primjer vrijedi i za primjenu UI-e. Ukoliko na pogrešan način odredimo ostvarivanje ciljeva koje bi sustav UI-e trebali postići ili ako sustavi UI-e pogrešno protumače naše ciljeve, to može dovesti do katastrofalnih posljedica. Kontrola sustava UI-e ne mora uvijek biti isključivo pozitivna stvar, ali svakako nas potiče na razmišljanje o djelovanju koje pripisujemo sustavima UI-e i kako se ono odnosi na naše ljudsko djelovanje (Nyholm, 2023A, 1232). Više o usklađivanju vrijednosti ćemo pisati u daljnjem tekstu potpoglavlja.

U kontekstu analize djelovanja sustava UI-e, znanstvenici otvoreno raspravljaju o dodjeljivanju statusa agenta i moralnog agenta sustavima UI-e.[10] Iako ni po pitanju prepoznavanja statusa agenta u sustavima UI-e ne postoji jednoglasno mišljenje, postoje znanstvenici koji smatraju kako sustavima UI-e možemo dodijeliti status agenta (Floridi, Sanders, 2004, 375; Hildebrandt, 2016, 316; Soltanzadeh, 2025, 1725), ali i znanstvenici suprotnog mišljenja koji tvrde kako sustave UI-e ne možemo kategorizirati kao agente niti kao moralne agente (Conradie, Nagel, 2024, 9), zbog čega pripisivanje povjerenja sutavu UI-e smatraju kategoričkom pogreškom (Conradie, Nagel, 2024, 5). Neki od najčešćih argumenata zbog kojih znanstvenici tvrde kako sustave UI-e ne može smatrati agentima i moralnim agentima tiču se normativnog koncepta vezanog uz moralno djelovanje (Sullins, 2006, 24). Tako, na primjer, Zafar smatra kako sustavi UI-e ne ispunjavaju nužne uvjete koji se često navode kao zadovoljavajući okvir zadobivanja statusa moralnog agenta kod čovjeka, kao što su: slobodna volja, autonomija, odgovornost, svijest (Zafar, 2025, 2607). Ostali autori navode kako sustavi UI-e ne ispunjavaju status moralnog agenta s obzirom da ne posjeduju namjeru (Sepinwall, 2024, 27) ili moralno razumijevanje (Hills, 2009, 99). Na temelju navedene argumentacije, postoji još protivljenje ideji kako UI ne može biti odgovorna niti joj se može vjerovati. Prema tim autorima, da bi netko bio nositelj odgovornosti, potrebno je da je taj entitet prvo agent, i to moralni agent (Matthias, 2004, 181–182; Himmelreich, 2019, 734–735; Dignum, 2019, 57). S obzirom kako odgovornost jedino može biti pripisana agentima, sustavi UI-e ne mogu biti niti odgovorni niti im se može vjerovati (Dorsch, Deroy, 2024, 18). Floridi i Sanders se slažu kako umjetni moralni agenti ne mogu biti odgovorni za svoje djelovanje iako oni jesu izvor moralnog djelovanja (Floridi, Sanders, 2004, 376). Rasprava o pripisivanju različitih vrsta agencija time postaje kompleksna te se najčešće odvija između ljudskih agenata i agenata UI-e ili robotskih agenata (Floridi, 2025A, 4). U trenutku kada imamo agente koji su sposobni ostvariti određeni cilj i biti odgovorni za ostvarenje tog cilja, izuzetno je važno precizno odrediti o kojim ciljevima se radi i na koji način se ti ciljevi mogu ostvariti. Ponekad ciljevi mogu biti moralno upitni, kao što je to slučaj u poznatnom misaonom eksperimentu Nicka Bostroma o maksimizatoru spajalica u kojem sustav UI-e podredi sve ostvarenju cilja o umnožavanju spajalica što izazove apokalipsu (Bostrom, 2014, 152-153). Ponekad i način na koji se pojedini cilj želi ostvariti može biti moralno upitan ukoliko način ostvarenja nije usklađen s ljudskim vrijednostima. Na primjer,

[10] Važno je istaknuti kako postoje razlike u vrsti djelovanja između agenta i moralnog agenta. Agenti se mogu kategorizirati u više različitih razreda ovisno o složenosti djelovanja od nižih prema višim stupnjevima djelovanja. Entiteti mogu utjecati na svoju okolinu putem različitih kanala, a utjecajniji kanali dovode do viših stupnjeva djelovanja.Vidi: Soltanzadeh, 2025, 1723.

sustav UI-e može imati moralno opravdan zadatak smanjiti gužvu u prometu, ali to može ostvarivati na moralno upitan način, na primjer, da odluči kako je najbolje da ljudi uopće više na voze automobile blokiranjem pristupa gorivu, onesposobljavajući automobile ili nanoseći fizičke ozljede ljudima.

Jedna od najčešćih i najširih definicija agenta glasi kako je agent sve što se može smatrati da opaža svoju okolinu putem senzora i djeluje na tu okolinu putem aktuatora (Russell, Norvig, 2021, 54). Pri tome se smatra kako ljudski agent ima oči, uši i druge organe koji mu koriste kao senzori, dok ruke, noge, govorni aparat predstavljaju aktuatore. Umjetni agent može imati kamere kao senzore i razne motore kao aktuatore. Agent UI-e obrađuje sadržaj informacija, datoteka i ljudskih podataka (putem tipkovnice, miša, zaslona osjetljivog na dodir ili glasa) kao senzorne ulaze i djeluje na okolinu izradom informacija, datoteka, prikazom informacija ili generiranjem zvuka. Okolina u kojoj agenti djeluju može biti sve u svemiru iako je u praksi to samo onaj dio svemira čije nam je stanje važno prilikom dizajniranja agenta (Russell, Norvig, 2021, 54). Više o statusu agenta, moralnog agenta i epistemološkog autoriteta UI-e ćemo reći u nadolazećem potpoglavlju.

Važno je napomenuti kako je s obzirom na činjenicu da sposobnosti sustava UI-e koji mogu nadmašiti sposobnosti liječnika u raznim područjima zdravstva, pitanje povjerenja se posljednih godina posebno razvija u zdravstvu vezano uz mogućnosti informiranog pristanka, vrijednosti koje unosimo u sustave UI-e, potencijalnu zamjenu liječnika sa sustavima UI-e, sustave UI-e kao epistemološkog autoriteta te posljedice prevelikog oslanjanja na tehnologiju (Tuncdogan i sur., 2025, 1). Izgradnja povjerenja u sustave UI-e u kontekstu zdravstvu podrazumijeva ostvarivanje visoke točnosti, objašnjivosti, ponovljivosti i reproducibilnosti, od čega je najteže osigurati objašnjivost. Koncept ponovljivosti znači kako bi neka predviđanja trebala pružiti slične rezultate na istom setu podataka s istim postupkom treniranja podataka, a koncept reproducibilnosti pretpostavlja kako bi sustav UI-e trebao moći pružiti pouzdana predviđanja čak i s drugim setom podataka (Rajpurkar i sur, 2022, 35). Osim toga, kada govorimo o povjerenju u UI-u u zdravstvu, važno je napomenuti kako nepotrebno visoko povjernje u UI-u koja može nadmašiti ljudske sposobnosti može imati katastrofalne posljedice, Zbog toga određeni znanstvenici podržavaju koncept optimalnog povjerenja u kojem i ljudi i sustav UI-e imaju određenu razinu skepticizma prema odlukama onog drugog s obzirom kako oboje mogu donijeti pogrešnu odluku (Asan, 2020, 4).

5.1. Informirani pristanak

Proces izgradnje sustava UI-e za primjenu u zdravstvu obilježen je s nekoliko kompleksnih koraka, a tijekom treniranja i dizajna modela, znanstvenici donose niz tehničkih izbora koji utječu na završni oblik sustava UI-e. No ti izbori, iako tehnički, vođeni su određenim vrijednostima i preferencijama znanstvenika, koji su ponekad skriveni, ali oblikuju sustav UI-e (Ratti, Graves, 2022, 807). Brojni znanstvenici se slažu kako nije dovoljno da sustavi UI-e pružaju pouzdana i točna predviđanja već je potrebno i objasniti kako je određeni sustav UI-e izgrađen te objasniti način na koji dolazi do svojih predviđanja (Kyrimi i sur., 2025, 3883; Liu i sur., 2025, 3; Alkhanbouli i sur., 2025, 2; Li i sur., 2024A, 4; Borole, Rajan, 2024, 1; Perni i sur., 2023, 1890; Holzinger, 2021, 437). Jer, čak i ako su sustavi UI-e pouzdani, još uvijek postoje etičke zabrinutosti vezane uz paternalizam, autonomiju pacijenata i informirani pristanak s obzirom kako dio medicinske prakse postaje nevidljiv zbog principa „crne kutije“ (Theunissen, Browning, 2022, 8).

Međutim, predviđanja brojnih sustava UI-e u zdravstvu nije lako objasniti s obzirom kako je način rada brojnih sustava UI-e karakteriziran fenomenom „crne kutije“ (DeGrave i sur., 2025, 294). Prema različitim etičkim kodeksima liječnika, kao što su Kodeks medicinske etike Američkog liječničkog udruženja pod oznakom 1.1.3 koji se tiče prava pacijenata ili Kodeks medicinske etike i deontologije Hrvatske liječničke komore u članku 2., stavka 8., savjetovanje pacijenta o različitim mogućnostima liječenja sastavni je dio etičke dužnosti liječnika prilikom pružanja informacije kako bi pacijent mogao donijeti informiranu odluku vezanu uz njegovo liječenje (American Medical Association, 2016A; Hrvatska liječnička komora, 2006). Takav pristup pacijentima pruža određeni osjećaj razumijevanja, prihvaćanja i kontrole nad situacijom, osiguravajući poštovanje dostojanstva i autonomije.

No, prije nego bi liječnik mogao savjetovati različite mogućnosti liječenja, morao bi razumjeti izazove vezane uz te mogućnosti i razumjeti razloge za preporuku određene opcije liječenja na „temelju objektivne profesionalne prosudbe liječenika“, kako to, na primjer, stoji u Kodeksu medicinske etike Američkog liječničkog udruženja. Ukoliko bi liječnik umjesto primjerenog savjetovanja koristio sustav UI-e koji radi prema principu „crne kutije“ ne bi mogao pružiti objašnjenje zašto postoje različite mogućnosti liječenja ili zašto su možda neke druge mogućnosti liječenja isključene. To bi pacijentu uskratilo razumijevanje konačnog plana liječenja, a oduzelo bi i mogućnost informiranog izbora budući da liječnik ne bi mogao ponuditi objašnjenje za propisno liječenje pacijentu (Grote, Berens, 2020, 208; Bjerring, Busch, 2021, 363; Chan, 2023, 288). U kontekstu odabira liječenja, Rosalind McDougall je istaknula kako bi

slaba točka sustava UI-e u zdravstvu mogla biti postavljanje prioriteta koji vode rangiranju različitih mogućnosti liječenja dajući prioritet maksimiziranju životnog vijeka po cijeni veće patnje. Takvo prioritiziranje ugrožava autonomiju pacijenta u procesu donošenja odluka (McDougall, 2019, 158). Ali liječnik bi na temelju preporučenih mogućnosti liječenja koji dolaze od sustava UI-e mogao otprilike reći koje nuspojave ili prednosti dolaze uz svako liječenje.

Informirani pristanak kao temelj zdravstvene skrbi, osigurava etičke odluke o liječenju i skrb usmjerenu na pacijenta na način da zdravstveni djelatnik informira pacijenta o rizicima, koristima i alternativama određenog medicinskog postupka ili intervencije (Shah i sur., 2024). Informirani pristanak se u literaturi često javlja u kontekstu sukoba između principa autonomije i dobročinstva (Shatz, 1986, 257; McCoy, 2008, 785; Vaughn, 2020, 106; Milligan, Jones, 2016, 26) pozivajući se na opise načela autonomije (Beauchamp, Childress, 2001, 57) i načela dobročinstva (Beauchamp, Childress, 2001, 166) kako su ih u literaturi popularizirali Beauchamp i Childress (Beauchamp, Childress, 2001, 77). Dok načelo autonomije osigurava informiranje i povjerljivost podataka, načelo dobročinstva često može doći u sukob s načelom autonomije kada liječnik smatra da zna što je najbolje za pacijenta. Taj sukob postaje najizraženiji kada liječnik pacijentu ograničava slobodu izbora i nameće vlastite poglede vezane uz dijagnozu, prognozu ili liječenje bolesti i tako nameće određenu vrstu paternalizma (Varkey, 2021, 19). Odnos povjerenja i informiranog pristanka u kontekstu UI-e je izuzetno važna tema, ali rijetko je do sada spomenut u smjernicama za razvoj UI-e (Reinhardt, 2023, 737).

U zdravstvenoj skrbi, dobivanje informiranog pristanka se smatra potrebnim prije liječenja kako bi se osiguralo da pacijent razumije potencijalne rizike i koristi zdravstvenog postupka u dovoljnoj mjeri (Smolenski, 2024, 115). Informirani pristanak nije rezerviran samo za područje zdravstva, nego se primjenjuje i u vojsci (Gross, 2019, 298), u pravu (Grant, 2021, 1; Cohen, Slottje, 2024, 167), pri digitalnim tehnologijama (Chomanski, Lauwaert, 2024, 2879), u akademskim istraživanjima (Skulmowski, 2025, 31). Koncept informiranog pristanka počeo se pojavljivati kao odgovor na nekoliko značajnih pravnih slučajeva, poput slučaja Schloendorff protiv Society of New York Hospitals iz 1914. godine u kojem je sudac Benjamin Cardozo presudio kako svako ljudsko biće odrasle dobi i zdravog razuma ima pravo odlučiti što će se učiniti s njegovim tijelom (Lombardo, 2025, 90). Uz to, sudac je dodao kako svaki kirurg koji izvodi operaciju bez pacijentovog pristanka čini prekršaj i mora biti odgovoran za potencijalnu štetu nastalu operacijom (Schloendorff v. Society of New York Hospital, 1914). U kasnijem razvoju događaja, zahtjev za informiranim pristankom postaje snažniji s obzirom na posljedice

zdravstvenih praksi iz nacističkog doba te je motiviran s ciljem sprječavanja šteta koje su nastale za vrijeme nacizma u budućnosti (Smolenski, 2024, 116). Tijekom poznatog „Suđenja liječnicima“ 1946. godine uspostavljena su načela Nürnberškog kodeksa koji čini osnovu modernih etičkih kriterija za ekpserimentiranje na ljudima, od kojih je informirani pristanak najvažniji. U ovom suđenju je 20 uglednih njemačkih liječnika i 3 javna službenika javno suđeno za ratne zločine i zločine protiv čovječnosti, u kontekstu neopisivih medicinskih eksperimenata na zatvorenicima koncentracijskih logora (Chelouche, 2013, 72). Iako se Nürnberški kodeks usredotočuje na medicinsko eksperimentiranje, inzistiranje na pristanku u istraživačkom kontekstu ubrzo se razvilo u inzistiranje na pristanku u svim zdravstvenim područjima koji uključuje i medicinski postupak. Ali formalnom uspostavljanju informiranog pristanka je prethodio slučaj *Salgo protiv Leland Stanford Jr. Univ. Bd. Trustees* u kojem je pacijent Martin Salgo bio paraliziran nakon operacije, ali mu nikada nije rečeno kako je to mogući rizik njegovog zdravstvenog postupka. Stoga je sud utvrdio kako je neotkrivanje rizika i alternativa samo po sebi razlog za pokretanje pravnog postupka koji ide dalje od slučaja nanošenja tjelesnih ozljeda (Mazur, 1988, 370). Američko liječničko udruženje objavilo je svoj stav o praksi informiranog pristanka tek 1981. te je potvrdilo kako informirani pristanak nije smo formalnost nego osnova zdravstvene struke (American Medical Association, 2016B).

Tako tijekom velike većine povijesti zdravstvene skrbi informirani pristanak nije igrao značajnu ulogu u skrbi za pacijente, a nevoljko promicanje potpunog otkrivanja informacija od pacijenta bio je dominantan stav zdravstvene struke (Green, MacKenzie, 2007, 117). Iako se Hipokratovo načelo „Primarno, ne našteti“ smatralo temeljnom obvezom liječnika još od antičkog doba, ono nije sa sobom nosilo obvezu otkrivanja informacija pacijentima. Naprotiv, često se ovo Hipokratovo načelo tumačilo kao ideja kako zdravstveni djelatnik nije obvezan podijeliti informacije jer bi to moglo uzrokovati rizik od štetnog ishoda (Faden, Beauchamp, 1986, 4). John Gregory i Benjamin Rush, liječnici iz 18. stoljeća su slično tumačili kako je primarna vrijednost informiranja pacijenata osigurati poštivanje preporuka liječnika. Ti liječnici su poticali druge kolege liječnike da popuste pacijentima u stvarima od malog značaja, ali da zadrže neprikosnoven autoritet nad pacijentima u pitanjima koje su bitne za život (Faden, Beauchamp, 1986, 9). Prema stavu kojeg su iznijeli Gregory i Rush, iako liječnici mogu dopustiti pacijentu koji se nalazi u niskom riziku da odluči hoće li ići ili ne na kontrolni pregled, ne bi trebali biti jdnako popustljivi ako pacijent ima dijagnozu koja ga potencijalno životno ugrožava kao što je karcinom ili mu je potrebna transplantacija organa koja spašava život. Budući kako bi takve intervencije bile potrebne za spašavanje pacijentova života, Rush bi se

zalagao za zamjenu pacijentove procjene liječnikovom, ako se pacijent i liječnik slažu kako bi trebalo nastaviti s liječenjem (Faden i sur., 1986, 23).

Informirani pristanak pacijenata koji se podvrgavaju određenom zdravstvenom postupku nije važan samo iz etičkih i pravnih razloga (Faden i sur., 1986, 24), već je presudan i za osiguravanje kvalitete zdravstvene skrbi: pacijentovo razumijevanje omogućuje suradnju, poboljšava rezultate i zadovoljstvo te pomaže u sprječavanju pogrešaka (Al-Jardani i sur., 2024, 3).

Danas poznajemo kako je informirani pristanak oblikovan i odnosom između pacijenta i liječnika, što je posebno važno s obzirom kako ne postoji potpuna jednakost između uloge pacijenta i uloge liječnika. Uloga liječnika od izuzetne je važnosti i moralna odgovornost liječnika poprilično dobro opisana u literaturi od strane Edmunda Pellegrina i Davida Thomasmae kroz četiri principa:

1) Liječnik i pacijent moraju slobodno donositi informirane odluke i djelovati u potpunosti kao moralni agenti;
2) Liječnici imaju veću odgovornost zbog nejednakosti informacija i moći;
3) Liječnici moraju biti osobe osobnog moralnog integriteta;
4) Liječnici moraju poštovati i razumjeti moralnu dvosmislenost, ali ne smiju napustiti potragu za onim što je ispravno i dobro u svakoj odluci (Pellegrino, Thomasma, 1987, 43-44).

Uloga liječnika je uvijek vezana uz veću poziciju stručnog znanja, moći i autoriteta, dok se pacijent nalazi u podređenijoj poziciji. Zbog toga se u praksi događa da informirani pristanak oblikuje i usmjerava liječnika jer on posjeduje stručno znanje i moć što daje dodatnu odgovornost samoj ulozi liječnika (Aramini, 2009, 209).

Na kraju, važno je napomenuti kako bioetičari brane informirani pristanak kao institut koji štiti povjerenje u zdravstvene djelatnike, zdravstvene institucije i zdravstvo općenito (Eyal, 2014, 437). O'Neill smatra kako je informirani pristanak općenito važan jer između ostalog može dati značajan doprinos obnovi povjerenja (O'Neill, 2002, 145). O'Neill prati misao Torbjörna Tännsjöa, Jennifer Jackson i drugih bioetičara koji su isticali potrebu za izgradnjom i održavanjem povjerenja kao razlogom za zahtijevanje informiranog pristanka u kliničkoj skrbi (Eyal, 2014, 437). Nancy Kass i suradnici smatraju kako je održavanje javnog povjerenja apsolutno ključno za istraživanje, a u drugim radovima pokazano je kako nepoštovanje informiranog pristanka rezultira negativnim ishodima po povjerenje (Eyal, 2014, 437).

Povjerenje pacijenata ponekad se izgubi nepovratno ukoliko otkriju da nisu poštovana etička pravila informiranog pristanka i da su prevareni. U takvim situacijama pacijenti mogu izgubiti povjerenje u liječnike, ali i u prave lijekove koji im mogu trebati u budućnosti (Bok, 1999, 73). U sljedećem potpoglavlju ćemo istražiti na koji način UI utječe na razvoj povjerenja u kontekstu informiranog pristanka.

5.1.1. Informirani pristanak u kontekstu primjene umjetne inteligencije u zdravstvu

Važnost informiranog pristanka nije nestala primjenom UI-e u zdravstvu. Dapače, brojni autori smatraju kako rasprava filozofa i etičara o prihvaćanju digitalnih tehnologija u zdravstvu s pojavom i velikim napretkom UI-e postaje sve aktualnija (Hurley i sur., 2025, 1; Pruski, 2024, 475; Park, 2024, 2). Sustavi UI-e pružaju velika obećanja za primjenu u zdravstvu, ali moralne implikacije nisu dovoljno ispitane što također vrijedi i za izazove informiranog pristanka. Iako rasprava o sukobu između načela dobročinstva i autonomije nije nova, primjena UI-e u zdravstvu dodaje novost, jer čak i kada pacijent želi koristiti svoju autonomiju kako bi primio skrb putem sustava UI-e, čini se kako ga upravo predanost autonomiji zbog nedostatka transparentnosti, putem informiranog pristanka, sprječava kako bi izabrao upotrebu sustava UI-e (Director, 2025, 1).

Više autora smatra kako se uzrok etičkih izazova informiranog pristanka u ovom kontekstu nalazi u nedostatku transparentnosti i objašnjivosti sustava UI-e te nedostatku informacija o različitim aspektima skrbi koju pružaju medicinski uređaji koji rade uz podršku UI-e (Petersen i sur., 2022, 58380; Avanzo i sur., 2021, 8; Xie i sur., 2021, 1131). Uz to, brojni sustavi UI-e ne mogu objasniti svoja predviđanja što se posebno odnosi na najpopularnije pristupe neuronskih mreža koji se trenutno koriste. Također, postoje znanstvenici koji tvrde kako je autonomija pacijenata narušena upotrebom sustava UI-e u zdravstvu što može narušiti povjerenje pacijenata u njihovu sposobnost sudjelovanja u donošenju odluka o zdravstvenoj skrbi ili odbijanju medicinskog postupka (Mooghali i sur., 2024, 6; Morley i sur., 2020, 11). Osnovna argumentacija sugerira kako primjena UI-e bez mogućnosti objašnjenja predviđanja onemogućuje informirani pristanak s obzirom kako liječnik ne može propisno obavijestiti pacijenta o potencijalnim rizicima zdravstvenog zahvata ili intervencije (Chan, 2023, 285). Uz to, očekuje se kako će sustavi UI-e u budućnosti nadmašiti liječnike u točnosti što će pacijentima u budućnosti pružiti razborit razlog za odabrati liječenje sustavima UI-e umjesto liječnika. No s obzirom na izazove „crne kutije“, sustavi UI-e ne mogu pacijentu objasniti zašto

su došli do određenog predviđanja što je potrebno kako bi pacijent mogao pružiti informirani pristanak (Director, 2025, 1). Stoga se čini kako bi liječenje isključivo putem sustava UI-e moglo biti zabranjeno s obzirom da ne ispunjava zahtjeve informiranog pristanka (Kundu, 2021, 1328) o čemu ćemo više govoriti u nastavku.

U kontekstu primjene UI-e u zdravstvu i informiranog pristanka, informirani pristanak se proučavao u kontekstu medicinskog prava (Cohen, 2020, 1426) i građanskog prava (Harvey, Gowda, 2022, 237) s obzirom na potrebu za informiranjem pacijenta o primjeni UI-e u zdravstvu. Cohen smatra kako američka sudska praksa vjerojatno ne bi zahtjevala obvezno informiranje o primjeni sustava UI-e u zdravstvu s obzirom kako je do trenutka izrade njegovog rada značajna većina sudova odbacila ideju kako neotkrivanje iskustva ili kvalifiakcije liječnika krši dužnost informiranog pristanka na temelju teorije kako su samo informacije o zdravstvenom postupku bitne (Cohen, 2020, 1435). Prema Cohenu, iznimke se nalaze u slučajevima kada pacijent pita o sudjelovanju sustava UI-e u donošenju odluka u zdravstvenom postupku, kada sustav UI-e ne radi na transparentan način, kada se sustavu UI-e daje prevelika uloga u konačnom donošenju odluka ili kada se sustav UI-e koristi za smanjenje troškova, a ne za poboljšanje zdravlja pacijenta (Cohen, 2020, 1429). Valja istaknuti kako je zbog sve većeg utjecaja UI-e na procese donošenja odluka u sektorima sa snažnim utjecajem, kao što je zdravstvo, na području Sjedinjenih Američkih Država u posljednjih nekoliko godina došlo do značajnog razvoja legislative u području UI-e. S tim ciljem je administracija predsjednika Joe Bidena 2023. godine izdala Izvršnu naredbu o sigurnom, zaštićenom i pouzdanom razvoju i korištenju UI-e (The White House, 2023) potaknuta Nacrtom za Povelju o pravima umjetne inteligencije koju je izdao Ured za znanost i tehnološku politiku SAD-a 2022. godine. Nacrt se sastoji od 5 načela i prateće prakse koja pomaže pri dizajnu, korištenju i primjeni sustava UI-e.

Unutar teritorija Europske unije, Opća uredba o zaštiti podataka, ili popularnije GDPR, te Povelja o temeljnim pravima Europske unije se smatraju izvorom prava koji podržava pravo pacijenta na informirani pristanak. GDPR u članku 22. zahtjeva ljudski nadzor za primjenu UI-e u zdravstvu te u članku 25. propisuje kako programeri moraju prilikom dizajniranja sustava UI-e voditi računa o zahtjevu za zaštitom podataka i privatnosti (Europski parlament i Vijeće Europske unije, 2016). Povelja o temeljnim pravima Europske unije putem članka 2. prava na zaštitu života te članka 7. prava na privatni život osigurava kako se rad u zdravstvenim ustanovama obavlja samo u skladu s najsuvremenijim standardom zdravstvene skrbi i informiranim pristankom (Europski parlament, Vijeće Europske unije i Europska komisija, 2016). Tako Europska unija dodatno regulirala izazove informiranog pristanka nastale u

kontekstu primjene sustava UI-e u zdravstvu s naglaskom na odgovornu i transparentnu upotrebu sustava UI-e zbog veće zaštite prava pacijenata (Stöger, Schneeberger, 2021, 35). Astromskė i suradnici smatraju kako GDPR potvrđuje pravo pacijenta na smisleno objašnjenje o logici uključenoj u stvaranje predviđanja sustava UI-e (Astromskė i sur., 2021, 515). Zanimljivo je i istaknuti kako neki znanstvenici primjećuju kako se odredbe GDPR-a o automatiziranom stvaranju predviđanja primjenju samo onda kada je odluka temeljena isključivo na sustavu UI-e. To znači kako u situacijama kada se sustav UI-e koristi kao alat za podršku ne postoji zakonska obveza informiranja pacijenta o njegovoj upotrebi (Schneeberger i sur., 2020, 211; Hacker i sur., 2020, 418).

Znanstvenici se generalno slažu kako bi pacijenti morali biti informirani o primjeni UI-e u njihovoj zdravstvenoj skrbi (Park, 2024, 2), a posebice kada je sustav UI-e uključen u proces donošenja odluka (Müller i sur., 2021, 121; Miguel i sur., 2020, 1143). Kiener je u svom članku naveo 3 glavna rizika koji dolaze s upotrebom UI-e u zdravstvu o kojima bi liječnik trebao obavijestiti pacijenta, a to su: rizik od kibernetičkog napada (Kiener, 2021, 707-708), rizik sustavne pristranosti sustava UI-e (Kiener, 2021, 709) te rizik neusklađenosti sustava UI-e s cjelokupnim medicinskim i kontekstualnim okvirom u kojem se pacijent nalazi (Kiener, 2021, 711). U suprotnom liječnik ili onemogućuje informirani pristanak ili krši obvezu upozoriti pacijenta na potencijalno štetne posljedice (Kiener, 2021, 705). Frank Ursin i suradnici su identificirali 8 važnih elemenata koji proizlaze iz tehničkih specifikacija UI-e i koje je važno istaknuti pacijentu iz etičkih razloga prilikom dijagnosticiranja dijabetičke retinopatije uz podršku sustava UI-e. U kontekstu informiranog pristanka, pet elemenata se odnosi na informiranje pacijenta, dva elemena potiču razumijevanje informacija, a jedan element osigurava dobrovoljni pristanak. Dobrovoljni pristanak se može formalno osigurati upućivanjem pacijenta na oftamološki pregled. Međutim, takav pristup bi mogao umanjiti očekivane pozitivne učinke dijagnoze uspostavljene sustavom UI-e, čiji je primarni cilj olakšati i ubrzati postupak dijagnosticiranja za liječnike. Dodatno razumijevanje se može potaknuti opisima ulaznih i izlaznih podataka sustava UI-e kao i objašnjenjem načina na koji je treniran sustav UI-e. Informacije koje je važno podijeliti s pacijentom uključuju tri ključna rizika: neusklađenost sustava UI-e, potencijalna pristranost sustava UI-e, te mogućnost kibernetičkog napada (Ursin i sur., 2021, 3). Sličan pristup poštivanja etičkih normi ističu i Kiselva i suradnici koji naglašavaju kako liječnici trebaju obavijestiti pacijente o upotrebi sustava UI-e pri dijagnozi i liječenju, te bi morali pružiti informacije o namjeni, razini sigurnosti te kvaliteti prije primjene sustava UI-e u zdravstvu (Kiseleva i sur., 2022, 12). Na taj način, integracija

tehničkih karakteristika UI-e s etičkim zahtjevima informiranog pristanka omogućuje transparentan i odgovoran postupak dijagnosticiranja, čime se istovremeno štite prava pacijenata i olakšava rad liječnika.

No čak i ako liječnik može navesti nuspojave ili prednosti određenog liječenja, bez razumijevanja specifičnog obrazloženja koje dolazi od strane sustava UI-e pacijentu može biti teško odabrati između liječenja, a pacijent možda neće moći dati informirani pristanak na način koji uravnotežuje njegove osobne prioritete. Na taj način pacijent je lišen određenog razumijevanja i autonomije pa ne bi mogao donijeti informiranu odluku. Stoga bi u budućnosti sustavi UI-e mogli biti opremljeni mogućnostima objašnjenja kako bi mogli pružiti neke dodatne informacije u procesu odlučivanja. No izazov je odlučiti koje točno informacije su potrebne pacijentu, a koje je potrebno uskratiti (Grote, Berens, 2020, 208)? Ovaj izazov postaje još specifičniji s obzirom na činjenicu kako su za različite pacijente relevantni različiti podaci, ali i činjenicu kako svaki pacijent patnju i bol doživljavaju subjektivno. Iako su potrebna dodatna istraživanja u ovom području, neki pacijenti bi za visoko rizične situacije željeli znati očekivani životni vijek i očekivanu bol ili nelagodu pri liječenju, dok bi drugi pacijenti željeli znati informacije poput vjerojatnosti neuroloških poteškoća ili moždane magle (Chan, 2023, 288). Presudno je zaštiti jezgru informiranog pristanka (autonomiju i dostojanstvo pacijenta) kako se informirani pristanak bez pružanja odgovarajućih informacija pretvorio u puko formalistički i pravni oblik zaštite (Grote, Berens, 2020, 208).

Do sada je već dokazano kako pacijenti i pružatelji zdravstvene skrbi mogu imati veće povjerenje i donositi informirane odluke kada razumiju obrazloženje zdravstvenih dijagnoza (Asan, 2020, 5). No s obzirom na brojne komplikacije do kojih dolazi prilikom uspostavljanja OUI-e u zdravstvu, nameće se pitanje trebamo li odustati od inzistiranja na informiranom pristanku ili bismo trebali odustati od upotrebe sustava UI-e koji rade prema principu „crne kutije“?

5.1.2. Informirani pristanak i „crna kutija“

Čini se kako je upotreba sustava UI-e koji rade prema principu „crne kutije“ u suprotnosti s nekoliko različitih etičkih i epistemoloških normi. Pretpostavimo na trenutak, zbog potreba razvoja argumenta, kako će sustavi UI-e nadmašiti sposobnosti liječnika u svakom smislu te će pružati pouzdane dijagnoze i predlagati liječenje koje daje bolje zdravstvene ishode za pacijente nego što to čine liječnici. Iako se možda radi o nedohvatljivoj pretpostavci, dopustit ćemo ju u

ovom trenutku kako bismo pokušali analizirati činjenicu da su sustavi UI-e koji djeluju prema principu „crne kutije“ problematični za informirani pristanak, a ne kako bismo testirali hipotezu o točnosti sustava UI-e.

Izazovi algoritamske pristranosti i pravednosti jedno su od područja koje je potrebno istražiti prilikom analize sustava UI-e koji su potencijalno problematični za informirani pristanak. S obzirom kako uspostavljanje dijagnoze i preporuke liječenja nisu toliko blisko vezani uz pitanje raspodjele resursa između pojedinaca, izazov pravednosti je manje relevantan za ovo pitanje. No izazov pristranosti je izuzetno relevantan jer sustav UI-e može biti pristran na način da iskrivljuje preporuke za određene skupine pacijenata. Na primjer, ako sustav UI-e smatra kako pacijenti crne boje kože žive u prosjeku kraće ili imaju veću toleranciju na bol od drugih etničkih skupina, mogao bi predložiti oblike liječenja koji vode do lošijih zdravstvenih ishoda za pacijente crne boje kože (Director, 2025, 5).

Prisutnost pristranosti je izuzetno važan izazov u području primjene UI-e u zdravstvu, a neki autori čak tvrde kako je nemoguće stvoriti sustav UI-e bez pristranosti (Poslon, 2025, 286). Različite vrste pristranosti mogu ugroziti informirani pristanak zbog nepredvidljivog načina na koji pristranosti mogu utjecati na donošenje odluka u zdravstvu. Dok informirani pristanak želi ojačati autonomni način donošenja odluka pacijenta, pristranost može dovesti u pitanje informirani pristanak uzrokujući nepravdu ili pristran razvoj sustava UI-e (Sylvestre i sur., 2021, 1).

Pristranosti se općenito svrstavaju u 3 glavne kategorije: vođene podacima, algoritamske i ljudske pristranosti (Norori i sur., 2021, 1). Do pristranosti koje su vođene podacima dolazi kada se neuravnoteženi ili pogrešno predstavljeni podaci unose kao podaci za učenje sustavima UI-e koji bi mogli potpuno ignorirati pogrešno predstavljane skupove podataka kao na primjer skupove podataka koji ne sadrže dovoljno podataka o ranjivim skupinama (Norori i sur., 2021, 3). Algoritamske pristranosti uzrokovane su sustavom UI-e koji je treniran na pristranim podacima pa dolazi do pojačavanja uzorka iz dominantnog skupa podataka s kojim je treniran. U najjednostavnijem slučaju, kada se sustav UI-e trenira za klasifikaciju skupa podataka koji se sastoji od 80% nepristranih i 20% pristranih podataka, sustav UI-e može postići maksimum od 80% točnih predviđanja (Norori i sur., 2021, 3). Ljudske pristranosti su povezane ljudskom pogreškom i najsuptilnije su jer je rezultat dugotrajnih društvenih predrasuda (Norori i sur., 2021, 4).

No u slučajevima sustava UI-e koji rade prema principima „crne kutije“ puno je teže identificirati i spriječiti pojavu pristranosti. To su dobro naglasili Obermeyer i suradnici u analizi komercijalnog sustava UI-e koji je prilikom predviđanja troškova liječenja pokazao očitu rasnu pristranost. Sustav UI-e je zbog nejednakog pristupa pacijentima crne boje kože predvidio kako je potreban manji trošak zdravstvene skrbi pacijentima crne boje kože nego pacijentima bijele boje kože, iako su pacijenti crne boje kože bili bolesniji (Obermeyer i sur., 2019, 3). Na taj način pristranost sustava UI-e uzrokuje nepravedne medicinske odluke te dodatne podjele neutemeljene u stvarnosti.

Poseban oprez u odnosu na pristranosti je potreban pri razvoju i evaluaciji sustava UI-e, implementaciji sustava UI-e te objavljivanju znanstvenih radova u kojima su korišteni sustavi UI-e kao podrška (Cross i sur., 2024, 5).

Naivan pristup programera pri razvoju i evaluaciji sustava UI-e bio bi pretpostaviti inherentnu objektivnost predviđanja sustava UI-e. Jedan od koraka razvoja sustava UI-e u zdravstvu uključuje odabir i obuku sustava UI-e i naknadnu evaluaciju sustava na neovisnim skupinama pacijenata (Vokinger i sur., 2021, 2-3). No poznato je kako u ovim koracima dolazi do nasljeđivanja pristranosti u podacima za obuku koji mogu uvesti i vlastite pristranosti što programer nikako ne bi smio zanemariti (Cross i sur., 2024, 10).

Implementacija sustava UI-e u zdravstvu također može uključivati pristranosti čak i ukoliko su sustavi UI-e dobili regulatorno odobrenje ili se čini kako pružaju potpuno nepristrana predviđanja. Ukoliko se pacijenti iz stvarnog svijeta razlikuju od pacijenata u skupu podataka za obuku i validaciju po demografskom sastavu, performanse sustava UI-e mogu se pogoršati i to različito za različite skupine pacijenata (Vokinger i sur., 2021, 2). Sličan trend može se prepoznati kada se sustavi UI-e obučeni u jednoj zdravstvenoj ustanovi, kao što je sveučilišna bolnica koja obično dokumentira osobe s višim prihodima, mlađe dobi i bijele rase, primjenjuju na podatke u drugoj zdravstvenoj ustanovi, bolnici u zajednici kojoj prvenstveno gravitiraju populacije pacijenata s niskim prihodima i pripadnici manjinskih skupina (Gianfrancesco i sur., 2018, 3). Ova vrsta pristranosti naziva se pristranost odabira uzorka te je proučena u literaturi i može biti uzrokom štetnih odluka u stvarnom kliničkom okruženju (Brewer, Carlson, 2024, 383; Nazer i sur., 2023, 8; Chauhan i sur., 2025, 1). Klinički primjer ove vrste pristranosti dobro se očitava u sustavu UI-e pod nazivom *Epic Sepsis Model* (dalje: ESM), koji je unatoč dobrim rezultatima predviđanja u početnim izvješćima i širokoj kliničkoj primjeni na nacionalnoj razini u stotinama bolnica SAD-a, kroz postupak vanjske validacije pokazao slabe rezultate u predviđanju početka sepse. ESM je identificirao 183 od 2552 pacijenta sa sepsom (7%) koji

nisu pravovremeno primili antibiotike što ističe nisku osjetljivost ESM-a u usporedbi s kliničkom praksom (Wong i sur., 2021, 1065). Osim toga, ESM nije identificirao 1709 pacijenata (67%) za koje se naknadno utvrdilo da imaju sepsu, stvarajući tako velik broj lažno negativnih predviđanja za pacijente koji su razvili sepsu. ESM nije identificirao 1709 pacijenata sa sepsom usprkos generiranju upozorenja za pacijente koji su ostvarili rezultate od 6 ili više na skali za stvaranje upozorenja na sepsu. U istraživanju kojeg su proveli Wong i suradnici takvo upozorenje je stvoreno za 6971 pacijenta od ukupno 38 455 pacijenta koje je ESM obrađivao uzrokujući umor od upozorenja (Wong i sur., 2021, 1068). Umor od upozorenja je ozbiljan izazov u zdravstvu koji zbog prevelikog broja lažnih upozorenja može dovesti do toga da zdravstveno osoblje manje ozbiljno shvaća upozorenja ili ih počne ignorirati (Xu i sur., 2025, 2) što u konačnici dovodi do smanjenjog povjerenja u primjenu sustava UI-e u zdravstvu (Casey i sur., 2018, 36). Činjenica kako je ESM pokazao puno lošije rezultate predviđanja u implementaciji nego u početnim izvješćima naglašava rizik od kliničke upotrebe sustava UI-e prije nego se stvarne performanse temeljito potvrde.

Pristranosti koji se javljaju prilikom objavljivanja znanstvenih radova o primjeni sustava UI-e također mogu stvoriti dodatne izvore pristranosti u sustavima. Tako Cross i suradnici ističu kako obrazac objavljivanja, to jest, način na koji se znanstveni radovi o primjeni sustava UI-e u zdravstvu stvaraju, objavljuju i distribuiraju može uzrokovati pristranost na globalnoj razini (Cross i sur., 2024, 11). Istraživanje objavljeno 2022. godine tako je pokazalo kako više od polovice sustava UI-e koristi javne podatke koji su stvoreni u SAD-u ili Kini što odražava naprednu tenhološku infrastrukturu tih zemalja. U istraživanju je potvrđeno kako su znatnu prednost ostvarili znanstveni radovi koji prikazuju pozitivne rezultate sustava UI-e u zdravstvu naglašavajući superiornost performansi sustava UI-e te pružajući na taj način nepotpun pregled stvarnog stanja primjene sustava UI-e u zdravstvu (Duyx i sur., 2017, 92). Također, većina autora je primarno bilo iz Kine i SAD-a, prvi i zadnji autori su bili podatkovni znanstvenici, a ne kliničari te su radiologija i patologija bile dvije najzastupljenije medicinske domene (Celi i sur., 2022, 2). Drugi primjer navodi istraživanje koje je provedeno za vrijeme pandemije COVID-19 kada su znanstvenici pokušali izgraditi učinkovite sustave UI-e za dijagnosticiranje bolesti uzrokovane virusom COVID-19 na temelju rendgenskih snimaka prsnog koša i računalne tomografije. Međutim niti jedan od 320 radova koji su uključivali podršku sustava UI-e nije pokazao zadovoljavajuće rezultate, a 55 od 62 rada u konačnoj analizi su pokazali visok rizik od pristranosti (Roberts i sur., 2021, 201).

Svi ovi čimbenici zajedno mogu pristrano utjecati na razvoj i prioritete nadolazećih publikacija o sustavima UI-e jer podatkovni znanstvenik može imati različite ciljeve od kliničara u pogledu razvoja i implementacije sustava UI-e u zdravstvu (npr. podatkovni znanstvenik može kao cilj imati optimizaciju performansi sustava UI-e dok bi kliničar kao cilj mogao imati maksimiziranje kliničke korisnosti). Te razlike mogu dovesti do pretjeranog naglašavanja važnosti razvoja sustava UI-e za samo određene primjene zanemarujući ostale primjene koje mogu biti jednako ili više utjecajne u kliničkim uvjetima (Cross i sur., 2024, 11-12).

U konačnici, cijeli je niz strategija kako bi se ublažile posljedice pristranosti u zdravstvu (Hasanzadeh i sur., 2025, 8; Mihan i sur., 2024, e752-e753; Sasseville i sur., 2025, 6). Važno je istaknuti kako se rješenja za ublažavanje pristranosti trebaju provoditi s oprezom što uključuje prikupljanje velikih i raznolikih skupova podataka, metode statističkog uklanjanja pristranosti, temeljitu evaluaciju modela, naglasak na objašnjivosti sustava UI-e te standardizirano izvještavanje o pristranostima (Cross i sur., 2024, 1).

Osim pristranosti, u literaturi se pojavljuje i argument nedostatka pravovaljane komunikacije liječnika i pacijenta uslijed nedostatka objašnjivosti predviđanja sustava UI-e. Nedostatak pravovaljane komunikacije liječnika s pacijentom može uzrokovati da pacijent počne propitivati treba li vjerovati liječniku ili ne (Kundu, 2021, 1328).

Važno je istaknuti kako se pravovaljani informirani pristanak sastoji od informacija i podijeljenog znanja između liječnika i pacijenta, te plana liječenja koji se temelji na zajedničkim odlukama. Potrebno je da pacijent pročita i potpiše pisani obrazac informiranog pristanka, ali osim toga informirani pristanak uključuje i usmenu komunikaciju koja pomaže liječnicima upostaviti snažniji odnos s pacijentom (Glaser i sur., 2020, 1). Agozzino i suradnici ističu dvije važne komponente informiranog pristanka: informacije o rizicima, koristima i alternativama kao i pristanak za podvrgavanje predloženom medicinskom zahvatu (Agozzino i sur., 2019, 2). Cilj je informiranog pristanka, između ostalog, implementirati načelo autonomije prema kojem se mora poštovati pravo pacijenta da odluči sam hoće li prihvatiti ili odbiti određeno liječenje ili medicinski zahvat (Durand i sur., 2015, 9). Postupak dobivanja pristanka mora osigurati da pacijent razumije prirodu svog stanja, rizike i koristi predloženog liječenja i njegovih alternativa te da na njega dobrovoljno pristaje. Složene odluke poput operacije ili drugog invazivnog postupka zahtijevaju raspravu o nesigurnostima (Brezis i sur., 2008, 352). Iako ne postoji jasna definicija što predstavlja adekvatne informacije, obično se definiraju kao ono što bi razumna osoba u pacijentovoj situaciji trebala da donese odgovarajuću odluku (Lemaire, 2006, 2). Pa iako je informirani pristanak ustaljena praksa, često ne ispunjava svoju

svrhu (Brezis i sur., 2008, 352). Brojni znanstvenici ističu kako je odgovarajuće razumijevanje pruženih informacija od strane pacijenata glavno ograničenje informiranog pristanka (Pietrzykowski, Smilowska, 2021, 1; Glaser i sur., 2020, 2; Brezis i sur., 2008, 356). Dokaz tomu je istraživanje provedeno do 2006. godine u kojemu se pokazalo kako je samo 50% sudionika razumjelo sve komponente informiranog pristanka u kirurškim i kliničkim ispitivanjima (Falagas i sur., 2009, 420). Zabilježeno je i nerazumijevanje određenih komponenti informiranog pristanka kao što su dobrovoljno sudjelovanje (Pietrzykowski, Smilowska, 2021, 3), nasumičan odabir (Durand-Zaleski i sur., 2008, 3) i upotreba placeba (Mandava i sur., 2012, 21). Nedostatak odgovarajućeg razumijevanja posebno je čest kod ranjivih skupina kao onih koji se suočavaju s jezičnim barijerama (Matiasek, Wynia, 2008, 127), ograničenim obrazovanjem (Ankuda i sur., 2014, 328) ili ograničenom zdravstvenom pismenošću (Bickmore i sur., 2009, 1; Sudore i sur., 2006, 867).

U praksi, traženje informiranog pristanka često može biti formalni čin u kojem se dobiva potpis pacijenta, pri čemu liječnici vjeruju kako je važna obveza ispunjena bez obzira na to da li je pacijentu pružena odgovarajuća informacija o medicinskom zahvatu do kojeg će doći (Turillazzi, Neri, 2015, 451). Osim toga, postoji malo smjernica vezanih uz savjete potrebne za snalaženje u složenosti komunikacije rizika i koristi. Razvoj takvih smjernica usmjeren na praksu trebao bi se temeljiti na dokazima o načinu komunikacije rizika i koristi (Kahrass i sur., 2021, 71).

Standardi prakse informiranog pristanka se nastavljaju razvijati i razvoj može biti pod utjecajem kulturnog konteksta (Glaser i sur., 2020, 2). Posljednih godina postoji trend razvoja standarda informiranog pristanka koji su usmjereni na pacijenta (Salzburg Global Seminar, 2011; Spatz i sur., 2016, 2). Promjene koje će utjecati na razvoj informiranog pristanka ovisit će o trajanju procesa informiranog pristanka i objašnjavajućim vještinama liječnika (Tam i sur., 2015, 191).

Međutim, kada je u određeni zdravstveni postupak uključen sustav UI-e koji radi prema principu „crne kutije“ prezentacija informacija može biti otežana i na temelju potencijalnih strahova, zbunjenosti ili pretjeranog samopouzdanja pacijenata i liječnika. Christian Bjerring i Jacob Busch ističu kako u nastojanju uspostavljanja zajedničkog uma između pacijenta i liječnika ključne komponente postaju razmjena informacija, objašnjenja i razumijevanja (Bjerring, Busch, 2021, 360). Da bi proces informiranog pristanka tekao na odgovarajući način liječnici moraju biti dovoljno upućeni kako bi pacijentima mogli objasniti način na koji sustav UI-e radi što je nemoguće s obzirom na netransparentan način rada „crne kutije“. U takvim

situacijama rezultat procesa je sustav koji je ljudima uglavnom nerazumljiv izvan svojih najosnovnijih ulaza i izlaza (Schiff, Borenstein, 2019, 140).

Iako su zabilježene situacije u kojima su pacijenti odbili objašnjenje, ovdje želimo ispitati može li pristanak biti informiran u slučaju u kojem pacijent ne bi mogao pristupiti objašnjenju dijagnoze i plana liječenja. Jordan Waden se slaže s nizom znanstvenika koji tvrde kako neprozirnost otežava zdravstvenim djelatnicima način rada sustava UI-e te onemogućuje pacijentu razumijevanje predviđanja. Waden ističe kako se u takvoj situaciji pacijent može smatrati dovoljno kompetentnim donijeti dobrovoljnu odluku, ali bez odgovarajućih informacija ne može dati informirani pristanak. Prema Wadenu, u najboljem slučaju u okolnostima prisutnosti principa „crne kutije“ pacijent može pristati na određeni zdravstveni postupak, ali takav pristanak nema istu pravnu i proceduralnu težinu kao informirani pristanak (Wadden, 2021, 96).

Samuel Director u pristupu ovom problemu predlaže svojevrsno redefiniranje koncepta informiranog pristanka razlikujući informacije višeg i prvog reda. Director smatra kako bioetičari i znanstvenici koji su proučavali pitanje informiranog pristanka implicitno razumiju kako su za pravovaljani informirani pristanak pacijenta dovoljne informacije prvog reda, iako takva distinkcija nije eksplicitno formirana unutar literature o informiranom pristanku. Director informacije prvog reda naziva one informacije koje su predmet liječnikovog procesa kliničkog rasuđivanja. Na primjer, broj bijelih krvnih stanica je povišen, što upućuje na stanje X za čije rješavanje će liječnik propisati zdravstveni postupak lijekom Y jer je dokazano kako lijek Y ublažava stanje X zbog razloga A, B i C. S druge strane, informacije višeg reda su, smatra Director, one informacije koje govore o samom liječniku ili liječenju. Na primjer, kada liječnik bez objašnjenja tvrdi kako u svim slučajevima X koje je prije liječio 99% ih je nastavilo živjeti zdravim životom ili da će lijek Y vjerojatno izliječiti pacijenta, ali nemamo uzročno objašnjenje zašto je to slučaj (Director, 2025, 13). Director smatra kako bi informacije višeg reda mogle biti dovoljne za pravovaljani informirani pristanak jer pacijenti danas ionako pristanak daju na temelju povjerenja i statističke pouzdanosti, a ne na temelju razumijevanja zdravstvenog postupka. Director na taj način predlaže redefiniciju informiranog pristanka u kontekstu primjene sustava UI-e i predlaže svojevrsno kršenje dosadašnjih etičkih standarda o pružanju objašnjenja pacijentu zdravstvenog postupka pod uvjetom da pacijent ima dovoljno povjerenja u svog liječnika ili način liječenja (Director, 2025, 2). Iako Director naglašava kako bi najbolja opcija bila kada bismo mogli imati vrlo točnu i objašnjivu umjetnu inteligenciju u zdravstvu, dok to ne postignemo, smatra kako je najbolje rješenje naglasak staviti na povjerenje u sustave

UI-e (Director, 2025, 1). Ukoliko bismo informirani pristanak smatrali pravovaljanim na temelju informacija višeg reda, značilo bi kako je naše povjerenje u sustave UI-e kao epistemološke autoritete dovoljno veliko da im povjeravamo cijenu vlastita života. Ipak, brojni primjeri su nam do sada pokazali kako povjerenje u sustave UI-e može biti narušeno te ćemo tezu o epistemološkim autoritetima UI-e temeljito analizirati u nadolazećem potpoglavlju.

Sličan pristup redefinicije informiranog pristanka na temelju povjerenja predlaže i Brian Pickering (Pickering, 2021, 1). Pickering smatra kako bi na temelju sljedećih četiri zapažanja trebalo razmotriti redefiniciju informiranog pristanka: prvo, informirani pristanak ima različite implikacije u različitim situacijama kao što su klinička ispitivanja, zaštita podataka, medicinske intervencije, online usluge. Drugo, u literaturi se pojavljuje više slučajeva u kojima prepotstavke informiranog pristanka ne vrijede, a uvjeti korištenja obično su nametnuti pa pristanak ne može biti informiran ni slobodan. Treće, postoje znantno složeniji scenariji od jednostavne razmjene podataka između liječnika i pacijenta ili istraživača i sudionika u istraživanju. Četvrto, objašnjivost se ponekad uspostavlja na način u kojem svi dionici doprinose zajedničkom razumijevanju određenog konteksta ili informacije (Pickering, 2021, 15). Informirani pristanak kojeg Pickering predlaže temelji se na pouzdanosti sustava UI-e i povjerenje u stručnjake za razliku od klasičnog informiranog pristanka koji se temelji na potpunoj informiranosti i razumijevanju što predstavlja novost u teoretskom pristupu informiranom pristanku.

U ovom dijelu je još vrijedno spomenuti kako, sustavi UI-e koji rade prema principu „crne kutije“ predstavljaju etičke i epistemološke izazove koje smo iznijeli u prethodnim poglavljima, smatramo kako se njihova primjena u zdravstvu ne mora odbaciti u potpunosti. Važno je naglasak u proučavanju staviti na interakciju ljudi i sustava UI-e, a svijest o epistemološkim ograničenjima sustavima UI-e uvjet je za učvršćivanje odgovorne upotrebe i interakcije s takvim sustavima (Durán, Jongsma, 2021, 334). Osim toga, postoji uvjet unutar kojeg bi moglo biti moralno opravdano koristiti sustave UI-e koji rade prema principu „crne kutije“, kao što je kopilot upotreba UI-e (Chan, 2023, 288).

Ovaj uvjet je vezan uz slučajeve u kojima sustav UI-e stvara dijagnozu ili plan liječenja, ali nakon što se ti rezultati predstave liječnik može neovisno potvrditi njihovu točnost, što se naziva „kopilot upotreba sustava UI-e“. Američka Nacionalna akademija znanosti, inženjerstva i medicine procjenjuje kako 5% svih odraslih osoba u SAD-u koji traže godišnju ambulantnu skrb doživi dijagnostičku pogrešku (Chan, 2023, 289).

Beltramin i suradnici ističu sličnu perspektivu u kojoj sustavi UI-e mogu raditi rame uz rame liječnicima pružajući podršku kao alat, ali konačna odluka ostaje na liječnicima. Takvi primjere imamo kod sustava UI-e koji rade prema principu „crne kutije“ za automatsko bilježenje simptoma tijekom medicinskih konzultacija (Rajkomar i sur., 2019, 836). U takvim okolnostima sustav UI-e ne mora biti objašnjiv jer ne ometa dijagnostički postupak i prikuplja samo ograničenu količinu podataka (Beltramin i sur., 2022, 251).

5.1.3. Informirani pristanak nasuprot „crnoj kutiji“

Jedan od najboljih prigovora koji postoji vezan uz argument kako liječnici trebaju izbjegavati upotrebu sustava UI-e koji rade prema principu „crne kutije“ jest da liječnici ponekad daju preporuke za liječenje iako ne razumiju zašto su određeni zdravstveni postupci učinkoviti. U tom smislu, nerazumijevanje i nedostatak objašnjivosti već je prisutan u medicinskoj praksi pa bi zbog toga bilo nedosljedno *apriori* zabraniti upotrebu sustava UI-e u zdravstvu samo zato što rade prema principu „crne kutije“. No Alex London pronalazi određene pukotine u shvaćanju i razumijevanju medicine kao znanosti u kojoj mogućnosti objašnjenja, predviđanja i uspostavljanja kauzalnih nizova nije usporediva s drugim znanostima. London kao primjer uzima građevinsko inženjerstvo koje, kao dio tehničke znanosti, jako precizno može utvrditi i objasniti zašto određeni most izdržava određenu količinu opterećenja, u kojim slučajevima može doći do narušavanja statike i urušavanja mosta. Zbog velike društvene važnosti projekata koji su omogućeni zahvaljujući zakonitostima građevinskog inženjerstva i tehničkim sposobnostima izvedbe, danas znanstvenici koji rade u tom području uživaju velik ugled, ali i nose veliku odgovornost. Na sličan način se medicina povijesno shvaćala gotovo kao tehnika i umijeće liječenja, a ne samo kao teorijska znanosti. Ako građevinski inženjer može raditi prema strogim standardima o kojima ovisi sigurnost ljudi, onda bi i liječnici analogno trebali moći ispunjavati iste standarde, objašnjenja i metode liječnika moraju biti jednako pouzdani i utemeljeni jer i oni upravljaju ljudskim životima. No London ističe kako građevinsko inženjerstvo uvelike može uskladiti teoriju i praksu jer su sposobnosti da se jasno, dosljedno i pouzdano objasne pojave i predvide ponašanja sustava znatno veća u građevinskom inženjerstvu nego u medicini (London, 2019, 17). Drugim riječima, znanje koje se primjenjuje u građevinskom inženjerstvu najčešće je dovoljno cjelovito u odnosu na praktične potrebe i dovoljno je dobro da daje točne rezultate. S druge strane, u zdravstvu se češće može dogoditi da su uzročno-posljedični odnosi nedovoljno istraženi i kauzalni nizovi nepoznati zbog čega se ta dva područja ne mogu adekvatno usporediti.

Stoga ne bismo uvijek trebali očekivati kako liječnici mogu objasniti zašto su određeni medicinski postupci učinkoviti. London ističe kako su liječnici redovito propisivali aspirin gotovo stoljeće bez razumijevanja mehanizma kako aspirin djeluje. Litij se također desetljećima koristi kao stabilizator raspoloženja za liječenje mentalnih bolesti, ali do danas istraživači ne mogu objasniti zašto djeluje (London, 2019, 17). Dakle, zbog činjenice kako sustavi UI-e djeluju prema principu „crne kutije“ ne bi trebali biti razlog za izbjegavanje primjene sustava UI-e koji djeluju prema principu „crne kutije“ u zdravstvu. Daljnji argument moguće je razviti u smjeru toga kako sustav UI-e koristi velike skupove podataka za davanje preporuka za liječenje pa je predviđanje takvog sustava temeljeno na nečemu barem jednako pouzdano kao i podaci koji se koriste u nasumičnim kliničkim ispitivanjima litija. Stoga je važno da liječnik i dalje samostalno donosi konačnu odluku, oslanjajući se na podatke koji su jednako pouzdani i provjereni kao podaci iz kliničkih ispitivanja.

Ipak, usporedba s kliničkim ispitivanjima je neprimjerena jer liječnici koji propisuju litij i dalje razumiju osnovne čimbenike koji utječu na zdravstveno stanje pacijenta (kao što su: povijest bolesti, simptomi, osjetljivost na lijekove, opće zdravstveno stanje) i uvelike utječu na odluku o propisivanju litija. Pri razmatranju propisivanja litija, osim što liječnik zna da njihov pacijent pati od nestabilnosti raspoloženja i drugih činjenica o pacijentu, mogu se prisjetiti i onoga što su čuli o kliničkim ispitivanjima litija. Mogu se prisjetiti presudnih čimbenika u ispitivanjima koji bi ih naveli da propišu litij određenom pacijentu, čimbenika poput dobi ili spola ispitanika, kao i stope uspješnosti liječenja (koje bi podatke iz ispitivanja koji podržavaju mogle čak i potražiti ako je potrebno). Dakle, liječnici ipak razumiju osnovne čimbenike koji ih navode da propišu litij pacijentu čiji čimbenici odgovaraju profilu pacijenta. Nasuprot tome, kada liječnici koriste UI-u koja radi prema principu „crne kutije“ predviđanje sustava UI-e može ukazati kako će određeno liječenje biti učinkovito, ali bez objašnjenja jer je način rada na koji je sustav UI-e došao do tog predviđanja neproziran. U takvoj situaciji liječnici imaju vrlo malo informacija koje bi im pomogle razumjeti zbog čega je određeno liječenje preporučeno, primjerice koji čimbenici vezani uz pacijenta najviše utječu na nastanak takvog predviđanja (Chan, 2023, 290). Iako je riječ o pouzdanoj statističkoj operaciji, predviđanje sustava UI-e koje tvrdi „Ovo liječenje će biti učinkovito“ i dalje ne ulijeva povjerenje kao što bi liječniku povjerenje ulilo „Ovo liječenje će biti učinkovito na temelju uvida u laboratorijske nalaze krvi.“ Uvid u konkretne čimbenike (kao što su: povijest bolesti, laboratorijski nalazi krvi) pomoglo bi liječniku u razumijevanju razloga zbog kojih je određeno liječenje potrebno. No, uvid u

konkretne čimbenike predstavlja uvid u unutarnji rad ponekad milijuna ponderiranih faktora[11] između umjetnih neurona koje treba interpretirati na izuzetno složen način (Gorban i sur., 2020, 391). Dakle, iako liječnici ne znaju ponderirane faktore na temelju kojih je sustav UI-e došao do predviđanja, razumiju ponderirane faktore u svojoj odluci. Iako ponderirani čimbenici sami po sebi ne pružaju dovoljno objašnjenje razloga za konačnu odluku sustava, nego su potrebni dodatni pristupi poput kontračinjeničnih objašnjenja (Fernández-Loría i sur., 2022, 1655), ponderirani faktori mogu pomoći u prvim koracima prema definiranju propisnog objašnjenja (Amann i sur., 2020, 7). No, ti ponderirani faktori mogu biti poprilično kompleksni s obzirom kako sustavi UI-e postaju sve složeniji tako da ih ne razumiju niti programeri koji su izrađivali sustav UI-e (Beisbart, 2025, 9).

London ističe kako, ponekad čak i tehnike za identifikaciju značajki ili skupa značajki kojima sustav UI-e daje značajnu težinu pri procjeni određenog slučaja, uspostavljeni odnosi između tih značajki i izlaznih klasifikacija mogu biti neizravni i krkhi. Zbog toga male promjene u naizgled nepovezanom aspektu podataka može rezultirati značajno drugačijim ponderiranjem značajki, a različiti ulazni podaci mogu rezultirati potpuno različitim izlaznim podacima (London, 2019, 17).

Čak i ukoliko bi sustav UI-e liječnicima ukazao na informacije poput etničke i dobne skupine povezane sa skupovima podataka koje je sustav UI-e koristio (npr. iz kliničkih ispitivanja) bilo bi nejasno jesu li ti čimbenici ponderirani faktori u sustavu UI-e. Slično tome, ako je sustav UI-e i programiran da uzme u obzir sve demografske podatke pacijenta, sustav UI-e ne bi mogao reći liječniku koji su od tih podataka imali presudnu težinu. Zbog toga ovaj niz argumenata prikazuje kako je usporedba s propisivanjem litija neprikladna (Chan, 2023, 291). Dakle, legitimno propisivanje litija ne može se koristiti za opravdanje izravne upotrebe sustava UI-e od strane liječnika za preporuku liječenja.

5.2. Posredovanje povjerenja sustavima umjetne inteligencije

Do sada smo prikazali primjere u kojima sustavi UI-e mogu posredovati povjerenje, ali isto tako mogu i narušiti povjerenje u primjenu UI-e u zdravstvu. Kako sustavi UI-e sve brže postaju dio svakodnevnog života čovjeka, prikladno je pokušati istražiti način posredovanja povjerenja sustavima UI-e. Osim što su važni ciljevi koje želimo ostvariti uz podršku sustava UI-e, važan

[11] U literaturi se javlja izraz „težinski faktor“ ili samo „ponderi“.

je i način na koji ćemo taj cilj ostvariti. Kada sustavi UI-e mogu znatno utjecati na život čovjeka, osim tehničke i logičke ispravnosti, izrazito je važno voditi računa o usklađenosti sustava UI-e s ljudskim vrijednostima. Na primjer, sustav nadzora koji se ostvaruje uz podršku UI-e može biti potpuno pouzdan, ali i izuzetno opasan ako ga koristi autoritarna vlada za suzbijanje političkih disidenta. Takvi sustavi mogu posredovati povjerenje samo ukoliko su prihvatljivi onima koji nadziru, ali i onima koji su nadzirani. U takvim slučajevima, usklađivanje vrijednosti sustava UI-e mora imati za cilj osigurati da svi na koje UI može utjecati imaju dobar razlog prihvatiti njezine rezultate. A kako se UI sve više uključuje u donošenje odluka, posebno o moralno upitnim pitanjima, usklađivanje vrijednosti postaje sve važniji i izazovniji aspekt sigurnosti UI-e. Jedna od prvih osoba koja je tvrdila kako je usklađivanje vrijednosti za posredovanje povjerenja iznimno važno bila je ranije spomenuta Annette Baier. Ona je smatrala kako povjerenje uključuje moralnu odgovornost pojedinca da djeljuje u najboljem interesu određene institucije. To implicira kako povjerenje zahtijeva određenu usklađenost vrijednosti i interesa između pojedinca i uključene intstitucije (Baier, 1986, 255). Na primjer, ukoliko pojedinac vjeruje kako će se institucija pobrinuti za njegov ili njezin interes, očekuje se kako će pojedinac razumijeti ili dijeliti te interese što predstavlja oblik usklađivanja vrijednosti.

Literatura iz područja psihologije i neuroznanosti otkriva dio složenih uloga koje ljudski mozak vrši prilikom posredovanja povjerenja, pri čemu se često koristi „igra povjerenja“ kako bi se ispitao način na koji ljudi donose odluke o povjerenju u situacijama rizika i nagrade (Bellucci i sur., 2019, 2). Neurološke korelacije u odnosu na povjerenje su posebno zanimljive kada se u obzir uzme kako su ljudi jedina vrsta koja povjerenje i suradnju mogu ostvariti s genetski nepovezanim strancima, s pojedincima koje više nikada neće sresti ili čak i u situacijama u kojima izostaje društveni ugled ili korist (Smith, 1998, 11; de Quervain i sur., 2004, 1255). Provedena su razna neurološka istraživanja o posredovanju povjerenja koja su vezana uz tri glavne oznake povjerenja. Prvo, uz pouzdana očekivanja koja uključuju iščekivanje pozitivnih nagrada, drugo, predviđanje ponašanja u budućnosti, te, treće, neizvjesnost o potencijalnim učincima povjerenja koja je vezana uz poziciju ranjivosti (Dimoka, 2010, 377). Istraživanja vezana uz iščekivanje pozitivnih nagrada provedeno je temeljem promatranja moždane aktivnosti za vrijeme razmjene dobara s ciljem utvrđivanja aktivacije regija mozga i promjene moždane aktivnosti kroz vrijeme. Dokazano je kako dio strijatuma (nakupina neurona u potkorteksnoj baznoj gangliji prednjeg mozga) ima važnu ulogu u procesu učenja iz društvenih odnosa, predviđanju tko je pouzdan te aktivaciji osjećaja nagrade kada se povjerenje isplati (King-Casas i sur., 2005, 80-81). Drugo, istraživanja vezana uz predviđanje ponašanja u

budućnosti provedena su analizama moždane aktivnosti i snimke magnetske rezonance pokazala su kako je paracingularni korteks (prednji dio moždane kore ili prefrontalni korteks) odgovoran za razvoj povjerenja jer pomoću tog dijela mozga ljudi zaključuju o namjerama druge osobe kako bi se predvidjelo buduće ponašanje druge osobe (Krueger i sur., 2007, 20084). Treće, orbitofrontalni korteks pokazuje veću aktivnost pri višim razinama nesigurnosti (Huettel i sur., 2005, 3307) te orbitofrontalni korteks može razlikovati više razina nesigurnosti (Hsu i sur. 2005, 1682). Na temelju istraživanja o nesigurnosti možemo očekivati kako će viši stupanj povjerenja biti povezan s nižom aktivacijom u orbitofrontalnom korteksu. Time istraživanja iz područja psihologije i neuroznanosti pokazuju kako na razvoj povjerenja utječu složeni kognitivni procesi do kojih dolazi kroz učenje iz društvenih iskustava putem međudjelovanja više moždanih područja.

S druge strane, neki filozofi tvrde kako je povjerenje jedinstvena vrsta odnosa koji se uspostavlja između dvije ljudske osobe i uključuje reakciju ili namjeru (Jones, 2012, 63). Zbog toga što neki filozofi na UI-u gledaju isključivo kao na skup tehnika razvoja softvera, koje nemaju status moralnog agenta niti slobodnu volju, oni smatraju kako im se ne može u potpunosti vjerovati (Ryan, 2020, 10; David, 2012, 15; Nickel i sur., 2010, 443). Detaljnije, kroz analizu razumskog, afektivnog i normativnog utemeljenja povjerenja Mark Ryan zaključuje kako ljudi prema tehnologiji i sustavima UI-e ne mogu razvijati odnos povjerenja nego razvijaju odnos oslanjanja. Prema Ryanu, razumsko utemeljenje povjerenja ne zahtijeva da UI razvija emocije prema osobi od povjerenja (kao što to zahtijeva afektivno utemeljenje povjerenja) niti da UI-a bude odgovorna za svoje postupke (kao što to zahtijeva normativno utemeljenje povjerenja). Osim toga, afektivno utemeljenje povjerenja zahtijeva da osoba djeluje iz dobre volje kako bi došlo do razvoja povjerenja prema osobi od povjerenja. UI nema slobodnu volju pa stoga ne može djelovati iz dobre volje niti ispuniti zahtjev afektivnog utemeljenja povjerenja. Normativno utemeljenje povjerenja zahtijeva da osoba od povjerenja snosi odgovornost za svoje postupke, a UI ne može preuzeti odgovornost za svoje postupke (Ryan, 2020, 17).

Anders Søgaard tvrdi kako ideja da se povjerenje ne može odnositi na tehnologiju ili sustave UI-e podliježe logičkoj falaciji *petitio principii* (Søgaard, 2025, 313). Søgaard se usredotočuje na argument autorice Karen Jones jer upravo njezin argument o utemeljenju povjerenja smatra klasičnim predstavnikom onih autora koji stavljaju granicu između ljudi i strojeva.[12] Karen

[12] Termin „strojevi“ ovdje označava termine „tehnologiju“ i „sustave UI-e“

Jones na sljedeći način objašnjava utemeljenje povjerenja i uspostavljanje granice između čovjeka i strojeva:

a) povjerenje zahtijeva kompetenciju;
b) s obzirom kako se ljudi mogu suzdržati od korištenja vlastitih kompetencija, povjerenje zahtijeva sposobnost izravnog djelovanja;
c) strojevi nisu u stanju djelovati izravno;
d) stoga strojevi ne mogu biti predmetom povjerenja (Jones, 2012, 63).

Søgaard upozorava kako je premisa pod b) umetnuta s ciljem da se strojeve isključi iz mogućnosti primjene povjerenja te da se primjena povjerenja ograniči isključivo na ljude. Budući da samo ljudi mogu odlučiti ne reagirati, dok se strojevi ponašaju prema unaprijed određenim pravilima i nemaju volju kojom bi mogli obustaviti djelovanje, ta premisa uvodi antropomorfizam u argument i omogućuje jedino ljudima da svjesno ili nesvjesno koriste vlastite kompetencije. Zaključak tako postaje premisa, a dedukcija koja vodi do zaključka da strojevi ne mogu biti predmetom povjerenja postaje primjer kružnog zaključivanja (Woods, Walton, 1982, 78). Søgaard stoga predlaže korekciju utemeljenja povjerenja i briše sposobnost izravnog djelovanja kao jednog od uvjeta utemeljenja povjerenja te proširuje opseg povjerenja i na tehnologiju, sustave UI-e i institucije (Søgaard, 2025, 317).

Osim toga, sa Søgaardom se slažu i ostali autori koji tvrde kako postoje sličnosti između povjerenja u ljude i povjerenja u automatizaciju, tehnologiju i sustave UI-e (Li i sur., 2024B, 3). Na primjer, Lewandowsky i suradnici se slažu kako postoji sličan mehanizam narušavanja povjerenja kod ljudi i automatiziranih sustava u situacijama kada dolazi do neočekivane greške (Lewandowsky i sur., 2000, 121). Neurološka istraživanja potvrdila su kako se neurološki mehanizmi koji se aktiviraju pri interpersonalnim igrama povjerenja koriste i u procjenama ponuda na web stranici eBay temeljenim na povjerenju (Hoff, Bashir, 2015, 440). Provedeno je i sveobuhvatno istraživanje metoda za mjerenje povjerenja u automatizaciju te je zbog raznolikosti potrebna standardizacija i razvoj integriranih metoda koje obuhvaćaju kognitivnu, emocionalnu i socijalnu dimenziju povjerenja (Kohn i sur., 2021, 17).

Epistemološki gledano, ukoliko imamo dobar razlog vjerovati kako su predviđanja sustava UI-e vjerojatno moralno ispravna, onda imamo dobar razlog prihvatiti ih. Radi se o istoj vrsti razloga koje imamo za prihvaćanje prosudbi i odluka ljudskih stručnjaka, kao što je to u slučajevima kod liječnika, čak i kada su odluke liječnika u suprotnosti s našim nestručnim stavom. Sve dok imamo dobar razlog vjerovati kako takvi stručnjaci znaju bolje od nas, imamo

dobar epistemološki razlog prikloniti se njihovim sudovima. Slično tome, ako imamo dobar razlog misliti kako sustav UI-e „zna“ bolje od nas u vezi moralnih pitanja, imamo dobar moralno-epistemološki razlog prikloniti se sustavu UI-e kao moralnom autoritetu (Schuster, Kilov, 2025, 3).

Razni primjeri nas uče kako nedostatak usklađivanja između čovjeka i sustava UI-e mogu završiti tragično i umjesto posredovati povjerenje, zapravo uzrokovati povećanje tehnoloških rizika.

Jedan od prvih primjera koji je mogao rezultirati nesrećom velikih razmjera dogodio se 26. rujna 1983. godine kada je potpukovnik Sovjetskog saveza Stanislav Petrov odlučio kako je sustav ranog upozoravanja za nuklearno oružje OKO pogrešno signalizirao nadolazeći nuklearni napad od strane SAD-a. Senzor koji je bio zadužen za očitovanje temperaturnog signala pogrešno je signalizirao nuklearni napad sa 5 balističkih raketa jer je odsjaj sunčeve svjetlosti na oblacima zamijenio za temperaturnu oznaku koji je bio identičan temperaturnoj oznaci motora interkontinentalne balističke rakete (Tamburrini, 2023, 156). Petrov nije obavijestio nadređene o potencijalnoj opasnosti od napada nuklearnim oružjem, iako je to bila njegova vojna obveza, ali je time spriječio potencijalni napad Sovjetskog saveza nuklearnim oružjem što bi uzrokovalo nesreću velikih razmjera što je moglo biti uzrokovano tehničkom pogreškom. U kasnijem objašnjenju zbog izostanka reakcije Petrov je rekao kako nije očekivao da se radi o stvarnoj prijetnji jer nije očekivao kako bi netko započeo nuklearni rat sa 5 balističkih raketa što je zdravorazumsko razmišljanje (Tamburrini, 2023, 156). Sustavi UI-e još uvijek nemaju sposobnosti donositi predviđanja na takav način i često otkrivaju krhkost u promjenjivim kontekstualnim situacijama koje izlaze izvan dosega uskih skupova pretpostavki i graničnih uvjeta (Cummings, 2021, 7).

Drugi primjer je vezan uz prometnu nesreću 2016. godine sa smrtonosnim ishodom koja je uzrokovana pogrešnim prepoznavanjem autonomnog vozila. Na autocesti u Kaliforniji 2016. godine autonomno vozilo tvrtke Tesla nije zakočilo nego se zabilo u bijeli kamion s obzirom da nije moglo razlikovati bijeli kamion od pozadinske svjetline neba (Stilgoe, 2018, 26).

U trećem primjeru prikazat ćemo na koji način sustavi UI-e mogu postati toliko moćni da preoblikuju stvarnost koju su u početku sustavi UI-e opisivali te bi se u konačnici stvarnost na neki način prilagodila opisu kojeg je sustav UI-e predvidio. Hannah Arendt je takvu situaciju opisala u knjizi „Podrijetlo totalitarizma“ govoreći o biheviorizmu kada je izjavila kako smatra da problem s modernim teorijama biheviorizma nije u tome što su pogrešne već u tome što bi

mogle postati istinite (Arendt, 1951, 345). Poznat je primjer iz 2018. godine u kojem je autonomno vozilo druge kategorije u vlasništvu tvrtke Uber u kojoj se nalazio taksist uzrokovalo smrtan slučaj zbog neuspjelog senzorskog prepoznavanja pješaka na cesti (Giannini, Kwik, 2023, 73). Prema izvješću Nacionalnog odbora za sigurnost prometa, automobil i senzori nisu uspjeli prepoznati da li je osoba koja je prelazila cestu pješak, vozilo ili biciklist. Iako su senzori automobila uočili pješakinju koja je gurala bicikl preko ceste 6 sekundi prije nego je došlo do sudara, odgovarajuća reakcija je izostala zbog toga što autonomno vozilo nije bilo trenirano za situacije u kojima pješak prelazi cestu gurajući bicikl (National Transportation Safety Bord, 2019). Na taj način, ukoliko svako autonomno vozilo ne bi imalo integrirane dodatne zaštitne mehanizme za prepoznavanje pješaka koji cestu prelaze gurajući bicikl, uvijek bi dolazilo do sudara. Na taj način, sustavi UI-e mogli bi predvidjeti kako ne postoji niti jedan pješak koji će cestu prelaziti gurajući bicikl, a pješaci više ne bi prelazili cestu gurajući bicikl zbog straha od sudara s autonomnim vozilima i stvarnost se na taj način prilagođava predviđanju UI-e o nepostojanju pješaka koji cestu prelazi gurajući bicikl.

No postoje i delikatniji primjeri koji jasnije naglašavaju izazov usklađivanja. Jasno je kako ne želimo niti jednu nesreću uzrokovanu pogrešnim predviđanjima i nemogućnošću prepoznavanja osoba u bilo kojoj situaciji koja bi mogla biti životno ugrožavajuća. U kontekstu autonomnih vozila cilj je primjenom nadzora sustava UI-e smanjiti broj prometnih nesreća na statistički što je manju moguću stopu, a po mogućnosti manju nego što se događa kada vozilima upravljaju ljudi. Ali u drugim domenama kao što je jezik, statistički parametri više nisu toliko jasni u postizanju određenog cilja. Opravdano se stoga pitati ukoliko se ne možemo dogovoriti oko osnovnih pitanja morala, kako naše ciljeve i standarde možemo pripisivati sustavima UI-e? Ovo pitanje je moguće razdvojiti na dva manja. Prvo pitanje se odnosi na to kako možemo napraviti nešto bolje od ljudske moralnosti. Znanstvenici poput Blaise Aguere Arcasa smatraju kako ljudi trenutno nisu osobito dobar izvog moralnog autoriteta i kako ljudske vrijednosti nisu dovoljno dobre (Christian, 2020, 247). Paul Cristiano smatra kako se ljudske vrijednosti i moralni autoritet mogu razviti uz pomoć sustava UI-e koje zove Humans Consulting HCH. Sustav UI-e u takvim situacijama služi kao resurs koji osnažuje ljude kako bi jasnije razumijeli vlastite moralne odluke, ali i ponudili određena predviđanja na pitanja kao što su: da li je određena medicinska procedura moralna? S obzirom da sustav UI-e postavlja detaljna pitanja o našim moralnim odlukama, sustav UI-e može posvijestiti moralne vrijednosti u praktičnim izazovima i pomoću predviđanja ponuditi određene posljedice odluka kroz koje je još jasnije do kojih vrijednosti nam je zaista stalo (Christian, 2020, 249). Drugi dio pitanja odnosi se na

pitanje: Kako sustavi UI-e mogu utjeloviti ljudske vrijednosti kada su ljudske vrijednosti heterogene i ljudi se ne slažu uvijek oko odgovora na etička pitanja ili postoji višemodalnih mogućnosti? Na primjer, u slučajevima kada autonomno vozilo zaobilazi objekt i statistički bi pola vozila objekt zaobišlo s lijeve strane, a druga polovica bi zaobišla s desne strane, rješenje ne može biti da autonomno vozilo samo nastavi voziti po sredini. Odgovor na ovo pitanje u kontekstu transplantacije bubrega ćemo pokušati dati u posljednjem poglavlju.

Važno je istaknuti kako različiti dionici percipiraju povjerenje u sustave UI-e različito: programeri se fokusiraju na performanse, korisnici na osobno povjerenje, a regulatori na usklađenost i izvješćivanje.

Programer	Korisnik	Regulatori
Povećano razumijevanje Poboljšanje performansi Stvaranje boljeg algoritma	Povećano povjerenje Pristranost i povjerenje Razumijevanje posljedica Izvješće i analize	Povećano povjerenje Pristranost i povjerenje Usklađenost s regulativom Izvješća

Tablica 3. Različito razumjevanje povjerenja sustava UI-e u odnosu na dionike između čovjeka i stroja

5.2.1. Problem usklađivanja vrijednosti u zdravstvu

S rastom sposobnosti donošenja samostalnih odluka UI-e jedno od najvažnijih izazova istraživanja postaje ponovna procjena sigurnog i odgovornog dizajna UI-e. Virginia Dignum ističe kako UI mora biti u stanju uzeti u obzir društvene vrijednosti, moralna pitanja, uravnotežiti relativnu važnost vrijednosti koje dijele različiti dionici i u multikulturalnim okruženjima te objasniti svoje predviđanje (Dignum, 2017, 1).

Istraživanja u području računarstva su do sada uspjela prikupiti ograničena znanja o integraciji ljudskih vrijednosti u sustave UI-e (Han i sur, 2022, 1385). Jedna od najstarijih definicija ljudskih vrijednosti dolazi od Clydea Kluckhohna koji je vrijednost definirao kao eksplicitnu ili implicitnu koncepciju, karakterističnu za pojedinca ili skupinu, koja se odnosi na ono što se smatra poželjnim i utječe na izbor između dostupnih načina, sredstava i ciljeva djelovanja

(Kluckhohn, 1951, 395). Ova definicija je usvojena u istraživanjima o vrijednostima u kontekstu razvoja računalnih programa (Zeng i sur., 2025, 3; Orlowski i sur., 2025, 2). Vrijednost se kasnije odnosila na ono što osoba ili skupina ljudi smatra važnim u životu (Friedman i sur., 2003, 2) što je u skladu s definicijom Stuarta Russella koji je vrijednost definirao kao ono što bi ljudi željeli imati u životu (Russell, 2019, 235). Sotala je ljudske vrijednosti definirao kao mentalne reprezentacije koje kodiraju vrijednosnu funkciju mozga (u smislu učenja s potkrepljenjem) prožetim kontekstualno osjetljivim osjećajem (Sotala, 2016, 113). Muehlhauser i Helm su došli do zaključka kako su ljudske vrijednosti dinamične, složene i teško ih je specificirati te kao rješenje predlažu teoriju preferencija kao obećavajući pristup za postizanje usklađenosti sa sustavima UI-e (Muehlhauser, Helm, 2012, 13).

Znanstveno proučavanje ljudskih vrijednosti ima dugu tradiciju u filozofiji, etici, psihologiji i sociologiji. Izvorno su vrijednosti shvaćene kao filozofski koncepti koji su neraskidivo povezani s kreposnim životom i moralom (Perry, 1926, 15). Doprinos razumijevanju ljudskih vrijednosti dao je Milton Rokeach koji je vrijednost definirao kao trajno uvjerenje da je određeni način ponašanja ili krajnje stanje postojanja osobno ili društveno poželjnije od suprotnog načina ponašanja ili krajnjeg stanja postojanja (Rokeach, 1973, 5). U posljednjim desetljećima ljudske vrijednosti su proučavane u kontekstu različitih znanstvenih izazova kao što su moralna prosudba (Weber, 1993, 435), procjena osobnosti (Heaven, 1993, 307) i proces psihoterapije (Kelly, 1990, 171). Rokeach je u popularnom istraživanju vrijednosti došao do zaključka kako je potrebno razlikovati dvije vrste vrijednosti: instrumentalne i terminalne. Instrumentalne vrijednosti se odnose na načine ponašanja i odražavaju karakteristike ponašanja koje se smatraju društveno poželjnima, a terminalne se odnose na krajnja stanja postojanja ili krajnje načine života koji su idealizirani. Drugim riječima, terminalne vrijednosti predstavljaju krajnje ciljeve u životu pojedinca, dok instrumentalne vrijednosti predstavljaju način kako pojedinac smatra da bi trebalo doći do ostvarenja pojedinog cilja. Primjeri instrumentalnih vrijednosti su: ambicioznost, otvorenost, vedrina, hrabrost, nježnost, pristojnost, a primjeri terminalnih vrijednosti su: svjetski mir, obiteljska sigurnost, sloboda, mudrost, spasenje, osjećaj postignuća, društveno priznanje (Rokeach, 1973, 363). Novija istraživanja vrijednosti svrstavaju u intrizične i ekstrinzične kategorije do kojih je došlo kroz analizu Frommovih promišljanja o vrijednostima kao sukobu između imanja i bivanja (Kasser, Ryan, 1996, 281). Intrinzične vrijednosti pogoduju ispunjavanju osnovnih psiholoških potreba, većoj integraciji osobnosti te uključuju: emocionalnu intimnost, prijateljstvo, samoprihvaćanje, osobni rast, doprinos zajednici (Sheldon, Kasser, 2001, 494). Ekstrinzične vrijednosti su obično motivirane

željom za materijalnom nagradom ili izgradnjom poštovanja, uključuju novac, financijski uspjeh, popularnost, fizički izgled (Kasser, Ryan, 1996, 280).

Vrijednosti su danas dio općeg referentnog okvira za život što znači da su relevantne u svakodnevnom ponašanju i donošenju odluka, a ponekad i ključne (De Giorgis i sur., 2022, 3). Vrijednosti su od neizmjerne važnosti kada procjenjujemo ponašanje i motivaciju pojedinca jer vrijednosti omogućuju da razumijemo ponašanje, motivaciju (Cieciuch i sur., 2015, 43) i aspekte ličnosti pojedinaca i smisao života (Siwek i sur., 2017, 561). Vrijednosti su povezane sa znanjem, zdravim razumom, izražavanjem osobnih stavova i slobodom prosuđivanja. S vremenom se na osnovi tih postavki razvila filozofska disciplina aksiologija kao teorija vrijednosti koja proučava ono što treba biti jer je tako dobro, odnosno ono što je dobro ili bolje u odnosu na alternativu (Van Fraassen, 1973, 6). Aksiologiju je dobro razlikovati od deontologije, ili teorije dužnosti, koja kao filozofska disciplina proučava ono što bi trebalo biti jer to proizlazi iz njegove dužnosti (Van Fraassen, 1973, 5).

Posljednih godina se ulažu napori s ciljem kako bi UI-a bila sposobna za donošenje dobrih odluka u skladu s ljudskim vrijednostima, a pojedini znanstvenici poput Garya Marcusa smatraju kako će u bliskoj budućnosti doći trenutak u kojem će svi sustavi UI-e imati ugrađene etičke sustave (Etzioni, Etzioni, 2016, 149). Ipak, vrijedi napomenuti kako trenutno razumijevanje ljudskih vrijednosti pruža slabu osnovu za razvoj UI-e (Han i sur, 2022, 1386). Trenutno postoji zabrinutost u znanstvenoj zajednici kako sustavi UI-e mogu stvarati neetična predviđanja zbog nedovoljne usklađenosti UI-e s ljudskim vrijednostima (Etzioni, Etzioni, 2016, 155). Neetična predviđanja bi mogla nanijeti štetu i ugroziti povjerenje u sustave UI-e. Znanstvenici smatraju kako se povjerenje korisnika mora temeljiti na UI koja je pouzdana, kompetentna, transparentna i usklađena s ljudskim vrijednostima (Dang, Li, 2025, 2). U svrhu uspostavljanja povjerenja prema predviđanjima koji dolaze od sustava UI-e, liječnici bi trebali surađivati s programerima kako bi naučili koristiti sutav UI-e, razumjeti kako funkcionira i naučiti kako objasniti predviđanja (Iniesta, 2025, 390; Gundersen, Bærøe, 2022, 3). Ukoliko liječnici razumiju kako sustavi UI-e primjenjuju zdravstvena predviđanja, moći će bolje procijeniti ishode i uključiti takve informacije u svoje odluke ili upozoriti na netočna predviđanja osiguravajući da se uvijek ponašaju u skladu s informiranim pristankom.

U literaturi se sve više stavlja naglasak na pacijenta koji nije samo biološki organizam nego ljudsko biće specifičnih potreba i vrijednosti (Ramsey, 2002, 268) te bi zdravstvena skrb trebala poštovati potrebe, želje i vrijednosti pacijenta te osigurati da vrijednosti pacijenta vode sve kliničke odluke (Abid i sur., 2024, 144; Hoque, 2024, 1; Institute of Medicine, 2001, 6).

Luciano Floridi smatra kako se budućnost UI-e nalazi u razvoju jedinstvenih sposobnosti djelovanja agenata UI-e uz osiguravanje usklađenosti s ljudskim vrijednostima i društvenim potrebama (Floridi, 2025A, 25). Ljudska dimenzija zdravstvene skrbi uključuje odnos između liječnika i pacijenta koji je prožet prepoznatljivim vrijednostima i dužnostima. U znanstvenoj zajednici se općenito smatra kako odnos pacijenta i liječnika zahtijeva pristup usmjeren na pacijenta koji poštuje autonomiju pacijenta te promiče informiran izbor koji je u skladu s vrijednostima pacijenta (Emanuel, Emanuel, 1992, 2226). Ostale vrijednosti uključuju poštivanje privatnosti, povjerljivosti, pravednosti i skrbi kao i promicanje dobročinstva i izbjegavanje štete te bi sustavi UI-e morali biti usklađeni s navedenim vrijednostima (Quinn i sur., 2021, 892).

Jedno od najvažnijih pitanja problema usklađivanja vrijednosti glasi: s kojim ili čijim vrijednostima bi se sustavi UI-e trebali uskladiti? Ukoliko promatramo tri najčešća scenarija, mogli bismo reći kako jedna vizija naglasak stavlja na utilitarističku viziju UI-e, drugi pristup je Kantovskog karaktera, a treći je obilježen vrlinama. Utilitaristička vizija smatra kako bi dugoročno tehnologije trebale biti dizajnirane kako bi stvorile najveću sreću za najveći broj ljudi. Kantovska etika sugerira kako bi principi koji upravljaju UI-u trebali biti samo oni za koje bismo racionalno mogli htjeti da budu univerzalni zakon, kao što su principi pravednosti ili dobročinstva. Etika vrline smjera na razvoj sustava UI-e utemeljen na vrlinama i razvoju karakternih osobina.

Podsjetit ćemo ukratko kako sustavi UI-e mogu učiti na tri glavna načina: nadzirano učenje, nenadzirano učenje i učenju s potkrepljenjem. Iako sva tri načina učenja imaju određene razlike, sva tri načina sustavi UI-e mogu učiti određene vrijednosti i na temelju njih sudjelovati u svakodnevnom životu čovjeka i zamjenjivati dio ljudske prosudbe. U knjizi *The Alignment Problem: How Can Machines Learn Human Values* Brian Christian navodi primjere državnih i saveznih zakona SAD-a koji nalažu određivanje jamčevine i uvjetnog dopusta putem sustava UI-e te automobila i kamiona koji na cestama voze sve autonomnije uz podršku sustava UI-e. Zbog toga je izrazito važno osigurati da sustavi UI-e djeluju u skladu s ljudskim normama i vrijednostima, razumiju naš način razmišljanja te rade ono što mi želimo (Christian, 2020, 12-13).

Također, u ranijem tekstu smo naveli izvore koji tvrde kako sustavi UI-e predviđanja mogu donositi objektivnije, neutralnije, racionalnije i bez predrasuda za razliku od ljudi (Araujo i sur, 2020, 613; Kahneman i sur, 2021, 275; O'Neil, 2016, 116). No isto tako, činjenica je kako neki sustavi UI-e mogu biti vođeni vrijednostima jer sadrže eksplicitne vrijednosne sudove (Kraemer

i sur., 2011, 251). Vrijednosni sud možemo shvatiti kao bilo koju tvrdnju koja izražava stav o tome kako bi stvari trebale ili ne bi trebale biti, ili što je dobro ili loše, ili što je poželjno ili nepoželjno. Postoje dakako sustavi UI-e koji ne izražavaju vrijednosne sudove nego prate određena logička pravila, kao što je to na primjer kod sustava UI-e koji izračunavaju leksikografski poredak konačnog skupa nizova za što koriste abecedni red. Takvi sustavi UI-e imaju samo jedno moguće točno rješenje za koje nije potrebna aproksimacija ili interpretacija (Kraemer i sur., 2011, 252). No sustavi UI-e koji se koriste za podršku u donošenju odluka rangiraju određena predviđanja na temelju unaprijed definiranih kriterija te čine određene vrijednosne sudove. Primjer rezultata predviđanja sustava UI-e u toj situaciji glasio bi: „Mogućnost A je najbolja opcija“ ili „Mogućnost A je bolja od mogućnosti B s obzirom na kriterij C“. Matteo Turilli se slaže kako je potrebno za takve sustave posebno voditi računa o skupovima etičkih načela s kojima su usklađeni sustavi UI-e (Turilli, 2007, 49).

U literaturi su istaknuta dva najčešća pristupa usklađivanja vrijednosti: „odozgo prema dolje“ (Allen i sur., 2005, 149-151) i „odozdo prema gore“ (Gabriel, 2020, 416). Prema pristupu „odozgo prema dolje“ vrijednosti bi trebale biti eksplicitno programirane u sustave UI-e prije implementacije kako bi sustav UI-e slijedio skup unaprijed određenih normi i vrijednosti, odnosno moralnu teoriju. Takvi sustavi UI-e bi u moralnim izborima uvijek pružili predviđanje u skladu s onim izborom za kojeg su programirani, na primjer: „Ubij dvije osobe kako bi spasio njih devet!“, „Najprije pruži zdravstvenu skrb onima koji su najugroženiji!“, „Nasumično odaberi pojedinca kada moraš birati između jednako pogođenih strana!“ Pristup „odozdo prema gore“ nalaže učenje sustava UI-e iz podataka na kojima se obučava, te potencijalno iz interakcije s korisnicima. No, s obzirom kako postoji malo podataka za obuku o tome kako ljudi reagiraju na moralne dileme u stvarnom životu, prema ovom pristupu možda ćemo morati siljediti podatke anketa i eksperimentalne podatke o tome što ljudi misle da bi trebalo učiniti u određenoj moralnoj dilemi (Kneer, Viehoff, 2025, 2672).

U članku *Is there an ethics of algorithms?* Felicitasa Kraemer i suradnici navode primjer segmentacije snimke magnetske rezonance poprečnog presjeka ljudskog srca. Na temelju snimke se procjenjuje točan volumen krvi tijekom različitih faza otkucaja srca. Razlika između krvi i tkiva srčanog mišića označava se sivim vrijednostima na snimci magnetske rezonance, a procjena volumena krvi započinje određivanjem koji je dio slike obojen svjetlije sivo i brojanjem piksela u tom svijetlosivom području što se naziva segmentacija. Proces segmentacije se provodi pomoću numeričkog praga P na način da su pikseli s vrijednošću svjetline iznad P označeni kao „svijetli“, a pikseli s vrijednoošću svjetline ispod P označeni kao

„tamni“. Granice između svijetlih i tamnih područja nisu oštre i ne postoji apriori ispravna vrijednost koja označava određeni prag niti principi iz kojih se mogu izvesti budući da je šum na snimci magnetske rezonance uvijek rezultat procesa mjerenja. Zbog toga programeri obično biraju vrijednost koja se čini razumnom za pragove, a izbor te vrijednosti utječe na procijenjeni volumen krvi. Procjena o prisutnosti krvi kasnije će utjecati na daljnje vrijednosti i samu dijagnozu tako da će u graničnim slučajevima dijagnoza ovisiti o vrijednostima praga P. Za brojne slučajeve u kojima će rezultati volumena krvi biti izvan graničnog slučaja i upućivati na jako zdrave ili jako patološke konfiguracije, neće biti puno nesigurnosti prilikom dijagnoze. Ali većina stanja pacijenata koji se nalaze u području graničnog slučaja ovisit će o vrijednosti praga kojeg definiraju programeri. Kako bi kvaliteta dijagnoze bila odgovarajuća, potrebno je u graničnim slučajevima poraditi na eliminaciji lažno pozitivnih i lažno negativnih predviđanja koji mogu u velikoj mjeri bili određeni vrijednostima pragova ili drugih varijabli (Kraemer i sur., 2011, 255).

Jedan od mogućih načina kako upravljati lažno pozitivnim i lažno negativnim predviđanjima jest pomoću epistemološke interpretacije načela opreza koje tvrdi kako je bolje dobiti lažno pozitivan nego lažno negativan rezultat predviđanja jer će to spriječiti liječnika i pacijente da lažno vjeruju kako je netko zdrav dok je zapravo bolestan. Što bolest može donijeti ozbiljnije posljedice po zdravlje i život pacijenta, to je važnije otkriti je li prihvatljivije riskirati pogrešnu identifikaciju kod pacijenta pomoću lažno pozitivnog rezultata (Peterson, 2007, 7). Masovni probir raka dojke kod žena je tako, na primjer, kritiziran zbog velikog broja lažno pozitivnih rezultata predviđanja jer je pogrešno dijagnosticiranje raka dojke izuzetno uznemirujuće za žene i njihove obitelji (Gigerenzer i sur., 2009, 1219).

Kraemer i suradnici napominju kako se znanstvenici općenito slažu kako je važnije izbjegavati lažno pozitivne nego lažno negativne rezulate predviđanja jer znanstveno znanje ima tendenciju kumulativnosti. To bi značilo kako skupu već postojećih uvjerenja dodajemo nova uvjerenja o tome kakav je svijet, a opravdanje za nova uvjerenja obično ovisi na manje ili više složene načine o starim uvjerenjima. No, kada je riječ o sustavima UI-e koji se primjenjuju u zdravstvu, trebali bismo primijeniti rigorozniji sustav epistemoloških vrijednosti jer posljedice lažnog vjerovanja koje se predstavljaju kao sigurne mogu biti katastrofalne. Ukoliko sustav UI-e liječnika navodi da bolesnog pacijenta ne liječi, tom pacijentu se nanosi šteta time što se ne liječi (Kraemer i sur., 2011, 257). Pri razmatranju etičkih aspekata sustava UI-e u zdravstvu ne postoji najjsnija argumentacijska linija kako bi liječnici i donositelji odluka trebali preferirati sustave UI-e koji ih navode da vjeruju u ono što je najvjerojatnije istina. Možda bi baš suprotno

trebalo snažno argumentirati kako bi sustavi UI-e u zdravstvu trebali biti osmišljeni tako da je vjerojatnije kako će predviđanja biti lažno pozitivna nego lažno negativna. Pri tome računa treba voditi da povećani broj lažno pozitivnih predviđanja ne dovodi do previše nepotrebnih i potencijalno opasnih zdravstvenih postupaka. Konačna odgovornost u odlučivanju prikladnosti liječničke intervencija treba biti na liječniku (Kraemer i sur., 2011, 258).

Kraemer i suradnici ističu kako se odabir praga i upravljanje lažno pozitivnim i lažno negativnim rezultatima predviđanja ne odnose na donošenje isključivo tehničkih odluka nego takve odluke zadiru duboko u sustave normativne filozofske etike s kojima se liječnici u konačnici suočavaju (Kraemer i sur., 2011, 259). Liječnik koji je sklon deontološkoj filozofskoj etici usmjeren je na zaštitu fizičkog i mentalnog integriteta pojedinog pacijenta pa će se stoga vjerojatno odlučiti za implementaciju liberalnijeg sustava UI-e za analizu snimke teških bolesti s ciljem dobivanja više lažno pozitivnih i manje lažno negativnih rezultata. U Kantovskom duhu, cilj je takvom liječniku izbjeći nanošenje štete pacijentu, a do štete bi moglo doći ako se teška bolest ne bi mogla otkriti na zdravstvenim snimkama. Liječnik koji bi tražio statističku korist za što veću količinu pacijenata tražit će strože sustave UI-e s manje lažno pozitivnih predviđanja čak i ako to znači više lažno negativnih predviđanja. Takav liječnik skloniji je utilitarističkom sustavu vrijednosti koji kao kriterij maksimizira korist za što veći broj ljudi (Kraemer i sur., 2011, 257). Liječnik koji se vodi etikom vrline bi mogao kalibrirati svoju odluku u odnosu na 4 osnovne vrline kao što su: pravednost, poštenje, odgovornost i briga. Ukoliko, na primjer, u obzir prilikom odabira praga liječnik nastoji minimizirati diskriminaciju i osigurati pravedan zdravstveni postupak svih pacijenata kao temelj kriterija za donošenje odluke može uzeti pravednost kao vrlinu (Hagendorff, 2022, 13).

Koji god sustav vrijednosti ili kriterij liječnik uzme, važno je napomenuti kako izbor granične vrijednosti za određeni sustav UI-e u zdravstvu ide ruku pod ruku s nekim normativnim pretpostavkama koje je odredio programer. Do izazova dolazi ukoliko nije moguće predvidjeti jesu li pretpostavke programera u skladu s pretpostavkama korisnika kao što su liječnici, zdravstveno osoblje i pacijent. Ukoliko etičke pretpostavke programera ostanu implicitne i ne podudaraju se s vrijednostima korisnika, nastaje ozbiljan problem u kojem korisnik donosi odluke na temelju rezultata predviđanja sustava UI-e bez svijesti o tome kako se njegove osnovne etičke vrijednosti sukobljavaju s vrijednostima programera. Više ćemo o ovoj temi pisati u posljednjem poglavlju.

Ranije u tekstu smo spomenuli kako se problem usklađivanja pojavljuje kada sustav UI-e nije usklađen s ljudskim vrijednostima (Christian, 2020, 13) te kako bi sustavi UI-e trebali moći

razlikovati što je u skladu s ljudskim vrijednostima, no teško je definirati kriterije prema kojima sustav UI-e može razlikovati vrijednosti (Bostrom, 2014, 235). Hristova i suradnici ističu kako je problem usklađivanja dvosmislen jer je problem usklađenosti normativan i ograničen, a istovremeno složen i ponekad kontradiktoran. Problem usklađenosti je normativan i ograničen zbog usmjeravanja na strogo slijeđenje skup pravila ili vrijednosti koje smo unaprijed odredili pa u praksi takav teoretski određen skupa pravila ne može uzeti u obzir stvarne složene situacije. Problem usklađenosti je složen i kontradiktoran jer tehnologija i sustavi UI-e ne mogu u potpunosti odražavati ljudske preferencije i subjektivnosti što izaziva kontradiktorne tvrdnje (Hristova i sur., 2025, 1339).

U literaturi postoji širok pregled uloge vrijednosnih sudova u znanosti (Hollnaicher, 2025, 111; Brown, 2024, 111; Scriven, 1974, 219). Rasprava se posebice vodila oko mogućnosti objektivnog i neutralnog položaja znanstvenika prilikom prihvaćanja ili opovrgavanja znanstvenih hipoteza. Richard Rudner je smatrao kako su empirijske hipoteze u znanosti uvijek privremene i neizvjesne te je zagovarao stav kako znanstvenik ne može čekati dok dokazi potpuno potvrde ili opovrgnu neku znanstvenu hipotezu. Zbog toga, prema Rudneru, znanstvenik mora postaviti određeni prag pouzdanosti u kojem definira koliko potvrde dokaza je dovoljno kako bi prihvatio ili opovrgnuo hipotezu (Rudner, 1953, 2). Taj prag često nije moguće odrediti čistom legikom nego implicitno uključuje uvid znanstvenika i vrijednosne sudove na temelju kojih znanstvenik procjenjuje koja razina pogreške može biti dozvoljena što je zapravo vrijednosna odluka o prihvatljivosti rizika. Ukoliko greška prihvaćanja ili odbacivanja hipoteze može imati životno ugrožavajuće posljedice (kao što je to slučaj u zdravstvu), prag pouzdanosti naše potvrde dokaza mora biti visok (Rudner, 1953, 3). Rudner smatra kako su vrijednosna neutralnost i ideja da znanstvenik može potpuno izuzeti vrijednosti iz kvalitativnog odlučivanja u praksi neodržive jer je prag pouzdanosti već prihvaćanje vrijednosnog značenja.

Cilj je usklađivanja vrijednosti UI-e pronaći načela i normativne etičke teorije kako bi se radnje i predviđanja sustava UI-e uskladile s ljudskim vrijednostima (Russell, 2019, 137; Kneer, Viehoff, 2025, 2671). Do neusklađivanja vrijednosti dolazi kada sustav UI-e ne uspijeva uskladiti ostvarivanje ciljeva s važnim moralnim vrijednostima ili etičkim ograničenjima zbog čega uzrokuje štetu. Brojni autori su izrazili zabrinutost vezanu uz neusklađivanje vrijednosti s obzirom na sve snažnije mogućnosti i važniju ulogu sustava UI-e u društvu zbog čega neusklađenost može biti uzrokom katastrofalnih posljedica po čovječanstvo (Bostrom, 2014, 235; Dung, 2025B, 1202), a neki autori naglašavaju kako bi neusklađenost mogla dovesti do

izumiranja ili trajnog onesposobljavanja čovječanstva (Ord, 2021, 138; Russell, 2019, 18). Na taj način, upravljanje sustavima UI-e postaje povezano s idejom korisne UI-e koja bi trebala biti sila dobra u svijetu. Ukoliko nije moguće stvoriti sustav UI-e koji pouzdano može ostvarivati određene ciljeve, postoji ozbiljan rizik kako takvi sustavi UI-e neće biti korisni ljudima (Dung, 2023, 2).

Osiguravanje usklađivanja vrijednosti od posebne je važnosti jer je pridržavanje standarda ljudske moralne procjene od posebne važnosti i motivirajuće za ljude (Gert, 2013, 2). Osim toga, u budućnosti će moralne vrijednosti biti sve važnije s obzirom kako će sustavi UI-e moći izvršavati zadatke s većom razinom autonomije i brzinom zbog čega ljudima neće biti moguće procjenjivati da li je svaki zadatak izvršen na odgovoran i etičan način (Allen i sur., 2005, 149). Drugim riječima, problem usklađivanja ćemo riješiti kada shvatimo kako izgraditi sustav UI-e tako da pokušavaju učiniti ono što tvorci sustava UI-e žele. Usklađivanje vrijednosti je posebice važno pri izazovima kada sustavi UI-e pružaju podršku za raspodjelu određenih dobara ograničenih resursa. U literaturi su takvi izazovi opisani u tzv. „problemu s kolicima“[13] koji opisuju žrtvene dileme u kojima je moguće intervencijom spasiti određeni broj ljudi iz problema s kolicima (Kamm, 2019, 11). Postoje i drugi primjeri moralnih izazova koji su aktualizirani u zdravstvu zbog organizacije zdravstvene skrbi i potreba odabira između pomaganja malom broju teško bolesnih pacijenata ili mnogo većem broju umjereno bolesnih pacijenata (Voorhoeve, 2018, 125). Slični primjeri moralnih izazova u zdravstvu analizirani su u kontekstu maksimiziranja ukupne dobrobiti koju možemo postići ili davanja pravednih šansi ljudima za preživljavanje (Kamm, 2020, 297; Bourguignon, Mossé, 2025, 5).

Izazov usklađivanja vrijednosti ima dvostruki karakter: normativni, koji određuje koja moralna načela je potrebno programirati, i tehnički, koji se bavi pitanjem kako se etička načela na pouzdan način mogu programirati u sustavima UI-e. Normativni zahtjev ima za cilj izdvojiti najčešće postojeće etičke sustave kao potencijalne prijedloge vrijednosti, a tehnički zahtjev nalaže programerima edukaciju o etici UI-e kako bi postali svjesni rizika i potencijalne štete koju sustavi UI-e mogu prouzročiti ljudima kako bi bili produktivniji u traženju strategija za bolje usklađivanje sustava UI-e s ljudskim vrijednostima (Iniesta, 2025, 393). Kroz analizu tehničke i normativne perspektive usklađivanja vrijednosti Iason Gabriel ističe prednosti učenja s potkrepljenjem jer agent UI-e tako uči pokušavajući maksimizirati nagradu koju prima iz okoline (Gabriel, 2020, 414). Sutton i Barto opisuju učenja s potkrepljenjem kroz usporedbu

[13] engl. *Trolley Problem*

dobrih i loših signala koji dolaze kao oblik nagrade agentu UI-e s iskustvima užitka i boli koji se pojavljuju u biološkom sustavu (Sutton, Barto, 2018, 7). Agent UI-e pokušava kroz procese pokušaja, pogrešaka i usavršavanja doći do sve boljih performansi koji su prvi puta opisani (Legg, Hutter, 2007, 394).

Do sada smo već prikazali primjere u kojima je na tehnički neispravan način programiran sustav UI-e autonomnih vozila tvrtke Tesla (Stilgoe, 2018, 26) ili tvrtke Uber (Giannini, Kwik, 2023, 73). Osim navedenih primjera, postoje i brojni drugi primjeri sustava UI-e u zdravstvu, kao što su generativni sustavi UI-e koji se bave otkrićem lijekova (Jin i sur., 2020, 55) i podložni su tehničkom izazovu koji se u literaturi pojavljuje pod nazivom „hakiranje nagrada“. Radi se o specifičnom izazovu koji se pojavljuje kod snažnijih agenata UI-e koji koriste učenje s potkrepljenjem te pronalaze domišljate načine za postizanje svog cilja ili nagrade iako nije bilo zamišljeno da agenti dolaze na taj način do cilja ili nagrade. Na primjer, ukoliko robotu zadamo cilj: „Sakupi što više smeća.“, robot može početi bacati ili uništavati stvari koje su nam korisne s ciljem dolaska do što veće količine smeća koje se nalaze u teško dostupnim mjestima, kao što je to iza dvosjeda, u kutovima sobe i slično. Upravo zbog tih novih načina ostvarivanja cilja dolazi do odstupanja u namjeravanim radnjama sustava UI-e zbog čega je potrebno razvijati strategije za sprječavanje „hakiranje nagrada“ (Yoshizawa i sur., 2025, 7). Postoje i drugi tehnički izazovi koji su uz okvir za operacionalizaciju i odgovoran razvoj sustava UI-e koji uključuju ljudske vrijednosti donijeli Solanki i suradnici (Solanki i sur., 2023, 228-234). Navedeni autori ističu kako se vrijednosna ograničenja trebaju postaviti kao vodeći kriterij gdje je god to izvedivo kako bi se osiguralo da se sustavi UI-e razvijaju na odgovoran način (Solanki i sur., 2023, 231). Prepoznavajući izazove formalnog kodiranja moralnih teorija u sustavima UI-e, neki znanstvenici su pokušali pronaći način kako bi mogli zaobići potrebu za izravnim specificiranjem moralnih načela u sustavima UI-e. Između ostalog, u tu svrhu je razvijeno inverzivno učenje s potkrepljenjem kao vrsta strojnog učenja koja za razliku od učenja s potkrepljenjem nema unaprijed određenu funkciju nagrade koju agent želi maksimizirati. Na taj način zaobilazi se jedan od najčešćih tehničkih izazova „hakiranja nagrada“ i cilj inverznog učenja s potkrepljenjem jest naučiti zašto određeni stručnjak nešto radi. Na primjer, sustavi koji uče na temelju inverzivnog učenja s potkrepljenjem pokušavaju iz reakcija iskusnog vozača zaključiti zašto vozač usporava na raskrižju ili ubrzava na ravnoj cesti. Iako sustavi UI-e ne znaju motivaciju vozača za određenom radnjom (sigurnost, brži dolazak na destinaciju, udobnost), pokušavaju na temelju radnji rekonstruirati vrijednosti koje motiviraju vozača na određenu radnju.

Određeni tehnički karakter izazova usklađivanja vrijednosti može se nazrijeti ukoliko su naše namjere za korištenje sustava UI-e iracionalne ili su korisnici sustava UI-e pogrešno informirani. Osim toga, korisnici sustava UI-e mogu željeti ostvariti ciljeve koji su amoralni ili štetni. Zbog toga moramo posegnuti za normativnim karakterom izazova usklađivanja vrijednosti.

Normativni karakter pitanja usklađivanja vrijednosti odnosi se na pitanja poput: Tko je stručnjak iz područja etike od kojeg bi UI trebala učiti? Iz kojih podataka i na koji način bi UI trebala učiti moralne vrijednosti? Trebaju li ti podaci uključivati ponašanja svih ljudi ili bi trebali isključiti ponašanja onih koji se očito ponašaju neetično, kao što je slučaj kod sociopatskih oblika ponašanja, ili nerazumno, kao što je slučaj kod fundamentalističkih oblika ponašanja? Konačno koje kriterije treba koristiti za određivanje najmoralnijeg agenta i da li je moguće rangirati entitete na ovaj način (Gabriel, 2020, 416)?

Pri odabiru sustava vrijednosti, na raspolaganju imamo minimalno 3 različite strategije. Jedna je odabir globalnog javnog morala i ljudskih prava, drugi je odabir sustava vrijednosti vođen misaonim eksperimentom „vela neznanja“ kojeg je razvio John Rawls, a treći je odabir vrijednosnog sustava unutar određene normativne teorije.

Brojni teoretičari se slažu kako globalni javni moral postoji, a načela pronalaze u doktrini univerzalnih ljudskih prava (Donnelly, 2007, 290; Cohen, 2010, 20; Ignatieff, 2001, 113). Usprkos postojanja pluralizma vrijednosti, ovi teoretičari smatraju da postoje vrijednosti oko kojih se u praksi svi slažu, a to je princip pravde koji je utemeljen prema međunarodnom pravu i podrazumijeva da svaki pojedinac ima pravo na vodu, hranu, obrazovanje, zaštitu od fizičkog nasilja i slično. No, s obzirom da različite skupine ljudi na ljudska prava gledaju različito, potrebno je detaljnije preispitati univerzalnost ljudskih prava, jer određene države i regije svijeta imaju puno više moći za određivanje standarda ljudskih prava. Osim toga, neki ljudi preferiraju ljudska prava jer vjeruju kako je život svet, drugi na ljudska prava gledaju kao proizvod ugovora između države i građanina, a treći u ljudskim pravima vide način promoviranja blagostanja i minimiziranja štete (Gabriel, 2020, 427).

„Veo neznanja“ kao misaoni eksperiment Johna Rawlsa popularan je pristup koji naglašava nepristran odabir načela i donošenje odluka koje nije pristrano prema bilo kojoj društvenoj skupini. Rawlsov pristup algoritamskoj pravednosti popularizan je kroz njegovo temeljno djelo *A Theory of Justice* iz 1999. godine u kojem je temeljno opisao „veo neznanja“ (Rawls, 1999, 118). No ovakav pristup ima određene poteškoće s obzirom da se pravednost često izražava u

odnosu prema različitim društvenim skupinama o kojima mi ne znamo ništa zbog prirode „vela neznanja“. Iako ovakav pristup može biti koristan za općenitu analizu moralne pravednosti, ovaj pristup je kritiziran pri upotrebi koja služi za detaljne odluke u kontekstu UI-e (Franke, 2021, 1814).

Treći pristup u kojem se sustavi UI-e usklađuju s normativnim teorijama, vrijednostima te izračunavaju funkcije promocije i degradacije je također korišten u brojnim istraživanjima (Bench-Capon, Atkinson, 2009, 45; Luo i sur., 2017, 203; Lopez-Sanchez i sur., 2017, 1613; Serramia i sur., 2018, 1294). Funkcije promocije i degradacije vrijednosti pokazuju koliko određena norma podržava ili umanjuje vrijednost koja nam je važna, pri čemu se funkcija promocije može koristiti za izračunavanje usklađenosti normativnog sustava s tim vrijednostima (Serramia i sur., 2023, 763). Ovaj pristup ćemo koristiti prilikom normativne analize sustava UI-e koji se koriste za izradu liste čekanja za transplantaciju bubrega. Pri tome ćemo se voditi idejom kako je važno identificirati pravedne principe za usklađivanje koji dobivaju adekvatnu podršku unatoč širokim razlikama u moralnim uvjerenjima ljudi (Gabriel, 2020, 411).

5.3. Umjetna inteligencija kao epistemološki autoritet

Suvremena društva su doživjela nezabilježen razvoj znanja do čega je došlo zahvaljujući podjeli epistemološkog rada, vezanog uz stjecanje znanja, obradu i primjenu znanja, te specijalizaciju rada (Hauswald, 2025, 716). U nekim područjima se društva hiperspecijaliziraju do mjere u kojoj se profesionalni standardi oblikuju u vlastiti rječnik u višegodišnjem ili višedesetljetnom procesu pa nitko izvan te profesije ne može razumjeti te standarde niti izazove ili ciljeve profesije (Millgram, 2015, 2). Posljedica takvog razvoja znanja jest da danas više ne postoje polihistori kao što su to bili Leonardo da Vinci u 16. stoljeću ili Thomas Jefferson u 19. stoljeću (Kapsetaki, 2024, 440), a udio cjelokupnog dostupnog znanja kojim jedna osoba može ovladati postaje sve manji. Zbog toga sve više povjerenja moramo ulagati u epistemološke autoritete, odnosno ljudske stručnjake u određenim područjima koji mogu pružiti istinita uvjerenja, znanje i razumijevanje.

Različiti misaoni eksperimenati su se provodili s ciljem dubljeg razumijevanja unutarnjih mentalnih stanja strojeva i sustava UI-e. Jedno od temeljnih pitanja koje je pomoglo u razvoju brojnih misaonih eksperimenata glasi: Može li stroj misliti (Gonçalves, 2023B, 2549; Hoffmann, 2010, 203; Dennett, 2004, 295; Wang, 2025, 1; Turing, 1950, 442)? U pozadini

odgovora na ovo naizgled jednostavno pitanje nalazi se kompleksan odgovor o statusu epistemološke aktivnosti. S obzirom kako pripisivanje epistemološke aktivnosti sustavima UI-e nije očito, važno je ispitati status epistemološke aktivnosti kojeg možemo dodijeliti sustavu UI-e (Coeckelbergh, 2025, 3). Jedan od prvih misaonih eksperimenata koji se razvio u odgovoru na to pitanje predložio je Alan Turing u obliku Turingovog testa o kojem smo ranije pisali. Turingov cilj je bio utvrditi može li računalo zavarati čovjeka na način da čovjek povjeruje kako komunicira s drugim čovjekom umjesto s računalom (Turing, 1950, 442). Razni znanstvenici pokušali su upisati različite interpretacije pitanja o mislećem stroju. Tako je na primjer, Noam Chomski izjavio kako je pitanje „Može li stroj misliti“ zapravo pitanje jezika te je i sam Turing primjetio kako je pitanje previše besmisleno za znanstvenu raspravu (Chomsky, 1995, 9). Marvin Minsky je Turingov test smatrao šalom jer je vjerovao kako je Turing predložio vlastiti test kao jedan od načina procjene stroja, ali je smatrao kako Turingov test nikada nije trebao služiti za procjenu inteligencije stroja (Gonçalves, 2023B, 2500).

Pojava humanoidnog robota Sofije 2017. godine privukla je pažnju javnosti i znanstvene zajednice jer je riječ o prvom robotu kojem je Saudijska Arabija dodijelila počasno državljanstvo. Osim što joj je dodijeljeno počasno državljanstvo Saudijske Arabije, Sofija je gostovala u raznim televizijskim emisijama, poput *The Tonight Show* s Jimmyjem Fallonom, i sastala se sa svjetskim vođama, kao što je to bila njemačka kancelarka Angela Merkel (Nyholm, 2020, 2). Znanstvena zajednica nastavila je proučavati odnos čovjeka i digitalnih tehnologija kako bi se utvrdile ključne razlike u vrstama djelovanja koje ljudi i roboti mogu obavljati. Vrste djelovanja posebno su važne prilikom moralnog rasuđivanja i donošenja odluka zbog utvrđivanja odgovornosti za djelovanje, posebice pri situacijama visokog rizika u zdravstvu. Malezijska tvrtka za robotiku razvila je po uzoru na Sofiju prvog malezijskog robota ADAM u 2019. godini i očekuje se kako će u nadolazećom razdoblju biti sve više ovakvih robota (Nyholm, 2020, 199).

Razvoj velikih jezičnih modela, poput ChatGPT-a, i sve veća primjena drugih sustava UI-e u brojnim područjima, kao što je zdravstvo, navelo je brojne znanstvenike na dodatna istraživanja o mogućnosti sustava UI-e kao epistemološkog autoriteta (Misselhorn, 2022, 31; Schwerzmann, Campolo, 2025, 8; Kerasidou, Kerasidou, 2025, 3). Kao što smo već pokazali, znanstvena literatura nije dovoljno precizna u izričaju kada su u pitanju etički i epistemološki izazovi umjetne inteligencije. Spomenuli smo tri glavne vrste UI-e na početku rada: slaba ili uska umjetna inteligencija, jaka ili opća umjetna inteligencija i superinteligencija, koje je dobro razlikovati. Status UI-e kao epistemološkog autoriteta u nedavnim publikacijama o „umjetnim

moralnim agentima“[14] najčešće je raspravljan u kontekstu velikih jezičnih modela, a budući kako superinteligencija trenutno nije stvarno moguća, većina rasprava o statusu UI-e vezana je uz slabu i jaku UI-u što nažalost nije dovoljno jasno naznačeno u suvremenoj literaturi (Misselhorn, 2022, 31; Firt, 2025, 175).

Sustavi UI-e ne dolaze do znanja kao ljudi putem ljudskih procesa s težnjom opravdanog istinitog vjerovanja. Sustavi UI-e proizvode probabilistička predviđanja do kojih dolaze iz skupova podataka algoritamskim zaključivanjem te statističkom korelacijom i čitljivom klasifikacijom čime se znanje čini izračunljivim (Shin, 2025, 3). Valja imati na umu kako takva vrsta predviđanja ne predstavlja temeljnu ontološku datost nego društveno-tehnološki konstrukt (Maalsen, 2023, 206). Usprkos tome što način razmišljanja niti djelovanje čovjeka i sustava UI-e nisu identični, široko je prihvaćeno razumijevanje kako sustavi UI-e nose određeni moralni značaj (Johnson, 2006, 195). Znanstvenici često upadaju u iskušenje načiniti neopravdan prijelaz između prepoznavanja moralnog značaja sustava UI-e i prepoznavanja sustava UI-e kao moralnog agenta. Osim toga, Zafar ističe kako znanstvenici u analizi epistemološkog statusa i moralnog djelovanja UI-e rade četiri česte pogreške: lažne analogije, obmanjujući redukcionizam, ignoriranje logičke međuovisnosti pojmova i ignoriranje normativnosti. Kao rezultat te četiri pogreške rasprava o moralnom djelovanju UI-e nailazi na ozbiljne probleme prilikom razgraničenja sposobnosti UI-e te dovodi do zbunjujućih pretpostavki, kao na primjer, kako su sustavi UI-e autonomna i svjesna bića sa složenim skupovima uvjerenja, što je pogrešno (Zafar, 2025, 2614).

Ne slažu se svi oko pitanja poput: Kakvu točno moralnu odgovornost[15] možemo pripisati sustavima UI-e? Mogu li sustavi UI-e razmišljati ili djelovati na moralan način? Mogu li sustavi UI-e imati status moralne i pravne osobnosti? Zbog potrage za tim i sličnim pitanjima, ali i zbog neusklađenosti znanstvenih pojmova u literaturi su se kao predmet rasprave pojavili takozvani minimalni koncepti moralne agencije koji u kontekstu moralnog djelovanja UI-e obično povezuju djelovanje sustava UI-e s deflacijskim značenjima autonomije, odgovornosti, inteligencije i srodnih pojmova koji proizlaze iz konteksta povezanih s ljudskim djelovanjem (Zafar, 2025, 2605). Moralna agencija ovdje se podrazumijeva u kontekstu klasičnog pogleda na ljudsko moralno djelovanje prema kojem moralni agenti moraju ispunjavati uvjete

[14] U literaturi se koristi oznaka „AMA“ što na engleskom jeziku označava „artificial moral agents“.

[15] Iako je to poznato, valja istaknuti važnu terminološku distinkciju u literaturi u engleskom govornom području. Naime, razlikuju se moralna odgovornost (engl. *responsiblity*), poslovna odgovornost ili odgovornost vezana uz dužnost koju netko obavlja (engl. *accountability*) te pravna odgovornost (engl. *liability*). Vidi: Smith, 2021, 535. Zbog ograničenja ovog rada zadržat ćemo se samo na analizi moralne odgovornosti.

racionalnosti, slobodne volje, autonomije i svijesti (Behdadi, Munthe, 2020, 197). Istraživanje moralne agencije kod sustava UI-e aktualizirano je kod autonomnih vozila (Bonnefon i sur., 2016, 1574) i vojnih dronova (Sparrow, 2007, 63) kako bi se ispitale mogućnosti u kojima sustavi UI-e donose moralne odluke koje potencijalno mogu utjecati na situacije života i smrti. U takvim visokorizičnim situacijama, znanstvenici su željeli ispitati hoće li prevladati interes pojedinca ili javno dobro. Iako se možda visokorizične situacije života i smrti čine malo vjerojatnim, čak i događaji male vjerojatnosti mogu se ostvariti ukoliko u jednom trenutku milijuni autonomnih vozila krenu voziti na cesti. No kako se te situacije nikada ne bi dogodile, programiranje autonomnih vozila i dalje mora uključivati pravila odlučivanja o tome što činiti u takvim hipotetskim situacijama. Stoga se ove vrste odluka koje se tiču raspodjele štete moraju donijeti mnogo prije nego što autonomna vozila postanu globalno zastupljena (Bonnefon i sur., 2016, 1574).

Osim moralnog agenta, u literaturi se često spominje sintagma moralni pacijent kao primatelj moralnog djelovanja ili kao entitet na kojeg se može djelovati dobro i loše. Prema klasičnom pogledu, svi entiteti koji se kvalificiraju kao moralni agenti također se kvalificiraju kao moralni pacijenti, ali ne vrijedi i obrnuto (Floridi, Sanders, 2004, 350). Ova distinkcija se u znanstvenu literaturu uvodi jer klasična podjela na moralne agente kao protagoniste moralnog djelovanja, bića za koja smatramo da mogu donositi moralne odluke i biti odgovorni za njih, nije dovoljna kako bi obuhvatila sva bića prema kojima imamo moralne obveze. Tako moralni pacijenti zaslužuju određeni moralni tretman, ali nemaju dužnost moralnog tretmana prema drugima (Kelley, 2025, 1089).

Jedna od najznačajnijih klasifikacija umjetnih moralnih agenata dolazi od Jamesa Moora koji je predložio hijerarhijsku klasifikaciju temeljenu na postojanju određenih sposobnosti relevantnih u kontekstu moralnog djelovanja. Moor razlikuje 4 vrste moralnih agenata: moralni agent utjecaja, implicitni moralni agent, eksplicitni moralni agent i potpuni moralni agent (Moor, 2006, 19-21). Moralni agent utjecaja je najjednostavniji oblik moralnog agenta koji stvara moralne posljedice bez razvijene svijesti o postojanju tih posljedica. Moralni agent utjecaja izvršavaju naše naredbe kao surogatni agenti i utječu na etičke odluke poput privatnosti, vlasništva i moći. Moor navodi primjer robotskih jahača deva u Kataru koji su smanjili korištenje mladih dječaka kao robova za jahanje deva (Moor, 2006, 19). Implicitni moralni agent održava određene moralne vrijednosti i djeluje moralno jer je programiran ili ima unutarnje funkcije koje potiču moralno djelovanje ili izbjegavaju nemoralno ponašanje. Moor navodi primjer bankomata koji su programirani za isporuku odgovarajuće količine novca.

Eksplicitni moralni agent može prepoznati i obraditi moralno relevantne informacije, donijeti moralne odluke i opravdati ih (Moor, 2006, 20). Eksplicitni moralni agent se može usporediti sa sustavima UI-e koji se koriste za igranje šaha: eksplicitni moralni agent može prepoznati informacije relevantne za šah, obraditi ih i donijeti odluku s ciljem pobjede u igri poštujući pravila igre. Kroz analizu trenutnog položaja figura u igri, eksplicitni moralni agent izračunava koji potez donosi najveću vjerojatnost za pobjedu u danim okolnostima. Za Moora eksplicitni moralni agenti djeluju ne samo u skladu s moralnim smjernicama već i na temelju moralnog razmatranja. Moor bi se složio s Kantom kako eksplicitni moralni agent ipak ne ispunjava sve standarde potpune moralne aktivnosti. Potpuni moralni agent prema Mooru moraju imati dodatne sposobnosti poput svijesti, intencionalnosti, slobodne volje koje posjeduju samo ljudska bića i otvoreno je pitanje mogu li sustavi UI-e ikada postići ta svojstva. Moor smatra kako je prosječan odrasli čovjek koji posjeduje svijest, intencionalnost i slobodnu volju potpuni moralni agent. Moor osporava tvrdnju da sustavi UI-e nikada neće biti potpuni moralni agenti jer smatra kako ne možemo sa sigurnošću reći kako sustavi UI-e u budućnosti neće imati sposobnosti svijesti, intencionalnosti i slobodne volje (Moor, 2006, 20). Umjesto daljnje rasprave o epistemološkom statusu UI-e, Moor tvrdi da je važno ispitati druge kategorije, a posebno istražiti mogućnost razvoja eksplicitnih etičkih agenata koji mogu pomoći u sprječavanju nemoralnih ishoda (Moor, 2006, 21). Moor predlaže razvoj eksplicitnih moralnih agenata iz tri razloga: prvo, zbog toga što ljudi žele da se strojevi odnose prema nama na dobar način, drugo, jer strojevi i sustavi UI-e postaju sve snažniji pa im trebaju jasnija etička pravila, treće, jer smatra kako bi programiranje ili podučavanje etike stroju unaprijedilo naše razumijevanje etike. Moor smatra kako će glavna prepreka stvaranju eksplicitnih moralnih agenata biti nedostatak zdravog razuma i znanja o svijetu. Na primjer, robot se može suzdržati od nanošenja štete ljudima samo kada precizno poznaje koje sve vrste štete su moguće. U tom smislu kodifikacija morala postaje iznimno važna jer je nemoguće kodificirati moral. Čak i ako je moguće kodificirati moral i sastaviti popis preciznih moralnih pravila uz pomoć izuzetno mudrog i pametnog etičara, to i dalje ne znači kako bi se umjetni moralni agent trebao ponašati u određenoj moralnoj dilemi nego bi agent ipak morao sam donositi određene moralne odluke. Umjetni moralni agent se može naći u novoj situaciji koja predstavlja drugačiju moralnu dilemu i zahtijeva da moralna odluka bude korelirana u odnosu na raniju situaciju na temelju koje je moralni agent naučio primijeniti određeno moralno pravilo. Takva interpretacija moralnih pravila u novoj moralnoj dilemi zahtjeva moralno rasuđivanje (Nyholm, 2020, 158).

Zanimljivo je napomenuti kako Moor ističe da bi sustavi UI-e trebali donositi racionalne odluke umjesto nas, ali ne bi nikada trebali odlučivati umjesto nas o osnovnim ciljevima i vrijednostima (Moor, 1979, 227). Ukoliko zamislimo sustave UI-e koji donose objašnjiva predviđanja koja su nepristrana s velikom preciznošću, Moor smatra kako takvi sustavi UI-e trebaju pomoći u ostvarivanju ciljeva ili zadovoljavanju određenih vrijednosti koje je zadao čovjek, ali sustavi UI-e ne bi smjeli mijenjati te ciljeve ili vrijednosti jer takva odgovornost leži samo na čovjeku (Rapaport, 2023, 404).

Zamislimo situaciju u kojoj umjetni moralni agent mora odlučiti treba li obavijestiti obitelj o pogoršanju zdravstvenog stanja pacijenta. Kada umjetnog moralnog agenta suočimo s takvim pitanjem, odgovor bi trebao biti: „Ovisi.“ Značajke o kojima ovisi odgovor umjetnog moralnog agenta mogu se odnositi na pacijentovu mentalnu kompetenciju (koliko je pacijent sposoban racionalno samostalno donositi odluke), ali i na pacijentov odnos prema obitelji (do koje mjere su pacijent i obitelj povezani). S obzirom kako pacijentov odnos prema obitelji može biti obuhvaćen neograničenim brojem različitih utjecaja, on se ne može u potpunosti obuhvatiti značajkama koje bi bile predmet rasuđivanja umjetnog moralnog agenta. Mogli bismo smatrati kako se opća pravila i dalje mogu primjenjivati ako se odnose na tipove značajki koje su dovoljno snažne što dovodi do jasne moralne odluke, npr. pacijentova obitelj je izuzetno zabrinuta za zdravlje pacijenta pa to opravdava razlog na temelju kojeg umjetni moralni agent obavještava obitelj o narušavanju zdravlja pacijenta. Ali dok god nemamo jasna pravila koja kodificiraju oblike i intenzitet ponašanja važnim za donošenje odluke to samo odgađa odluku dok ne kodificiramo potrebna ponašanja. Primjerena razina zabrinutosti se može izraziti na brojne načine kao što su govor, ton glasa, učestalost posjeta, neverbalno ponašanje na razne načine, tako da su sama pravila nekodifikabilna (Graff, 2024, 5).

Kroz proteklih nekoliko desetljeća svjedočimo razvoju različitih sustava UI-e koji su zamišljeni kao moralni ili dobri. Jedan od prvih pokušaja stvaranja pravila moralnog djelovanja sustava UI-e dolazi od Isaaca Asimova koji je predložio 3 zakona robotike:

1) Robot ne smije ozlijediti čovjeka niti dopustiti da se čovjek ozlijedi;
2) Robot mora slušati čovjeka, osim ako to krši prvi zakon;
3) Robot mora čuvati sebe, dok to ne krši prvi ili drugi zakon (Asimov, 1942, 100).

Međutim ne postoji jednostavan način kodifikacije programskog jezika ovih pravila. Kodifikacija pravila za sprječavanje štete čovjeku zahtijeva precizno predviđanje posljedica djelovanja robota i razumijevanje štete što predstavlja iznimno složene procese. Zbog toga bi

pravila koja je predložio Asimov više mogla biti usmjerena na kratkoročnu fizičku sigurnost dok očito postoje drugi načini na koje bi ljudi mogli biti ozlijeđeni. Djelovanje robota ili sustava UI-e može neizravno uzrokovati dugoročnu fizičku štetu ili prouzročiti štete poput pshiološke traume i emocionalno uzrujavanje. Ova pravila također podrazumijevaju robote koji mogu razumjeti ljudske naredbe i potencijalni sukob s prvim zakonom, ali takvo robusno razumijevanje prirodnog jezika još nije postignuto (Murphy, Woods, 2009, 16).

Winfield i suradnici su pokušali razviti „dobre“ robotske sustave s ciljem prevencije ozlijede drugih robota na temelju procjene mogućih putanja u određenom okruženju. Glavni robot ima za cilj pokušati presresti putanju robota za kojeg predvidi da bi mogao upasti u rupu i potencijalno se ozlijediti (Winfield i sur., 2014, 6-7). Međutim, kada glavni robot treba spasiti istovremeno dva robota od potencijalnog nanošenja ozljeda, ne postoji način određivanja prioriteta zbog čega dolazi do kolebanja i nesigurnog odabira putanje. Winfield i suradnici ističu kako smatraju da se njihov robot ponaša u skladu s Asimovljevim prvim zakonom robotike prema kojem robot ne smije ozlijediti ljudsko biće ili svojom neaktivnošću dopustiti nanošenje ozljeda čovjeku (Winfield i sur., 2014, 5). Ovi rezultati su odjeknuli u javnoj raspravi u kojoj se pokušalo doći do zaključka radi li se o moralnim robotima opisujući glavnog robota kao „minimalnog etičnog“. Međutim, dotični roboti su programirani da se ponašaju s ciljem prevencije ozljeda, a iako se čini kao da oklijevaju prilikom dileme istovremenog spašavanja dvaju robota, to oklijevanje je posljedica njihovog programiranja (Sharkey, 2020, 284).

Ron Arkin istraživao je vojne robote i smatrao kako bi roboti mogli donositi moralnije procjene od ljudi koji procjene često donose opterećeni emocijama. Arkin smatra kako bespilotni vojni sustavi na bojnom polju neće biti moralno savršeni, ali mogu donositi moralnije procjene jer se ne moraju štititi, mogu se samožrtvovati, mogu donositi procjene bez emocionalnih utjecaja, mogu integrirati informacije iz više izvora brže od ljudi i samostalno pratiti razvoj događaja na bojnom polju (Arkin, 2008, 6-7). Arkin je kasnije razradio regulatorna pravila za vojno djelovanje robota prema kojima bi roboti minimizirali kolateralnu štetu u skladu s vojnom nuždom, ali bi mogli donositi i odluke o upotrebi smrtonosne sile koje su etički prihvatljive prema propisanim pravilima (Sharkey, 2020, 285). Iako su ova pravila propisana na način da robot djeluje moralno, paradoksalno je kako roboti nemaju mogućnost izbora djelovanja jer njihove radnje uključujući upotrebu smrtonosne sile određuje sustav pravila postavljen od strane programera, slično kao u prethodno opisanom slučaju Winfielda i suradnika. Arkinova pravila naišla su na kritiku od strane Matthiasa koji ističe kako su navedena vojna pravila nedovoljno jasna i kontradiktorna te zahtijevaju tumačenje ljudskih namjera što je složen

postupak. Matthias dodaje kako vojska može zaobići navedena pravila prema vojnoj potrebi pa smatra kako ta pravila imaju više savjetodavnu nego regulativnu ulogu. Uz to, predviđeno je da roboti djeluju prema pravilima bez mogućnosti propitivanja naredbi što izostavlja autonomiju koja je prema Matthiasu ključna za moralno djelovanje (Matthias, 2011, 281).

Odrediti epistemološki status UI-e važno je u kontekstu atribucije moralne odgovornosti sustavima UI-e jer su do sada brojni znanstvenici upravo kroz usporednu analizu ljudskog moralnog djelovanja analizirali potencijalno moralno djelovanje sustava UI-e i u skladu s tim pripisivali odgovarajuću kategoriju moralnosti sustavima UI-e (Firt, 2025, 176; Behdadi, Munthe, 2020, 197; Allen i sur., 2000, 251). Ne postoji univerzalno prihvaćeno stajalište o preciznim kriterijima koje bi sustavi UI-e trebali zadovoljiti niti kako bismo na temelju tih kriterija mogli precizno odrediti epistemološki status UI-e oko kojeg bi se svi znanstvenici složili iako je rasprava o kriterijima široka i uključuje autonomiju, svijest, odgovornost, interaktivnost, prilagodljivost (Sison, Redín, 2023, 48; Himma, 2009, 21; Mosakas, 2021, 430-431; Floridi, Sanders, 2004, 357-358). To posljedično stvara i velik problem u području pripisivanja odgovornosti sustavima UI-e. Na primjer, ukoliko nismo u mogućnosti objasniti kako sustav UI-e dolazi do predviđanja patimo od određene vrste neznanja što otežava prihvaćanje odgovornosti za postupke, ali otežava odgovorno djelovanje prema drugima.

U kontekstu proučavanja epistemološkog statusa UI-e i atribucije odgovornosti u literaturi je često proučavan specifičan izazov rascjepa odgovornosti (Veluwenkamp, Hindriks, 2024, 7; Nyholm, 2023B, 194; Santoni de Sio, Mecacci, 2021, 1060) koji je uveden od strane Andreasa Matthiasa i definiran pri situacijama u kojima stroj djeluje bez ljudske intervencije, ali tradicionalni načini atribucije odgovornosti nisu kompatilibni s ljudskim osjećajem za pravdu niti moralnim okvirom društva. Iako bismo inače proizvođača stroja smatrali odgovornim za posljedice djelovanja, u situacijama u kojima stroj djeluje bez ljudske intervencije nitko nema dovoljno kontrole nad djelovanjem stroja kako bi mogao preuzeti odgovornost za posljedice djelovanja (Matthias, 2004, 177). Rascjep odgovornosti karakterističan je za situacije u kojima se čini prikladnim kako bismo nekoga trebali smatrati odgovornim. Na primjer, trebali bismo nekoga pohvaliti ili ga okriviti za određeni ishod, ili bi nam se moglo činiti kako bismo trebali identificirati nekoga tko se može smatrati odgovornim za određeni ishod, ali ne postoji nitko koga je očito ispravno ili opravdano smatrati odgovornim. Uz to, nitko nije voljan ili sposoban preuzeti odgovornost pa stoga nastaje rascjep odgovornosti. O različitim uzrocima rascjepa odgovornosti kao i o potencijalnim rješenjima detaljno raspravlja Sven Nyholm u poglavlju pod

nazivom *Responsibility Gaps, Value Alignment, and Meaningful Human Control over Artificial Intelligence* (Nyholm, 2023B, 194).

Iako je do sada u velikoj mjeri pisano o epistemološkom statusu UI-e i vrsti moralnog djelovanja kojeg sustavi UI-e potencijalno mogu posjedovati (Llorca Albareda i sur., 2023, 59; Constantinescu i sur., 2021, 804; Himma, 2009, 27; Sison, Redín, 2023, 55) do sada nije dovoljno naglašena veza između atribucije odgovornosti i vrsta djelovanja sustava UI-e. Rascijep odgovornosti često je u literaturi vezan uz različite vrste umjetnih agenata i mogućnosti moralnog djelovanja sustava UI-e, stoga ćemo pažnju posvetiti analizi različitih vrsta umjetnih agenata te mogućnostima moralnog djelovanja sustava UI-e kako bismo detaljnije definirali epistemološki status kojeg sustavi UI-e zaslužuju (Llorca Albareda, 2025, 3861).

Određivanje epistemološkog statusa UI-e podrazumijeva kako bismo trebali moći dati odgovor na neka normativna pitanja strojne etike u užem smislu koja se odnose na moralne standarde koji bi se trebali primijeniti kod umjetnih moralnih agenata. Na primjer: Mogu li agenti učiti moralne vrijednosti? Trebaju li moralni agenti slijediti utilitarne, areatičke ili deontološke normativne principe? Možemo li se pri donošenju odluka osloniti na moralne teorije koje su osmišljene za ljudski društveni život ili su nam potrebni novi pristupi za umjetne moralne agente? Trebaju li se umjetni moralni agenti uopće oslanjati na moralne principe ili bi trebali razmišljati na temelju slučaja? U širem smislu epistemološki status UI-e je važan zbog promišljanja o moralnim implikacijama umjetnih moralnih agenata na individualnoj i društvenoj razini. Neka od pitanja strojne etike koja su u širem smislu relevantna su na primjer: Trebamo li težiti ostvarenju umjetnih moralnih agenata? Mogu li umjetni agenti postići razinu svijesti koju prepoznajemo kod prosječnog čovjeka? Postoje li područja primjene u kojima se umjetni moralni agenti ne bi trebali koristiti, ako se uopće trebaju koristiti? Postoje li moralne odluke koje se ne bi trebale delegirati strojevima? Kakav je moralni i pravni status umjetnih moralnih agenata? Mogu li umjetni moralni agenti promijeniti društveni život i moral ljudi ako postanu sveprisutniji (Misselhorn, 2022, 31)?

Pojedini filozofski izazovi poput definiranja slobodne volje, svijesti ili autonomije stvaraju poprilične izazove s obzirom kako ih je potrebno precizno definirati prije nego se mogu tehnički detaljno razraditi. Na primjer, ukoliko ne znamo što je točno slobodna volja, ne možemo ju tehnološki modelirati. Određivanje svijesti umjetnog agenta uključje još veće poteškoće jer se svi filozofi ne slažu oko toga da li je uopće teorijski moguće da umjetni agent bude svjestan i koju točno vrstu svijesti umjetni agent može doseći. Neki filozofi smatraju kako samo bića koja

su biološki živa mogu biti svjesna, dok drugi vjeruju kako će bilo koji entitet s mozgom koji je složen kao čovjekov proizvesti svijest bez obzira na materijale od kojih je mozak sastavljen. Ukoliko se ipak ispostavi kako je moguće da umjetni agenti budu svjesni, postoje potencijalno nepremostive epistemološke poteškoće u određivanju da li je neki agent svjestan ili ne (Himma, 2009, 28). Taj problem se u literaturi pojavljuje pod nazivom „problem drugih umova“ i ima dugu filozofsku tradiciju blisko vezanu uz skeptičnu poziciju naspram mogućnosti znanja drugog uma. Zbog toga neki autori smatraju kako bi se početak problema drugih umova trebao pripisati Renéu Descartesu koji je u svojim poznatim djelima *Rasprava o metodi* i *Meditacije o prvoj filozofiji* prvi puta opisao dijelove ovog problema (Avramides, 1996, 28). Drugi autori smatraju kako problem drugih umova ne leži u dualizmu uma i tijela kako je to Descartes opisivao, nego u empirističkom tumačenju mišljenja kao niza mentalnih procesa zbog čega ovu tradiciju vežu uz empirističkog filozofa Johna Stuarta Milla (Lagerspetz, 2025, 255). Problem drugih umova ukratko se sastoji u tome kako vlastiti um poznajemo promatranjem vlastitog uma, odnosno introspekcijom, a druge možemo promatrati samo izvana kroz ponašanje (Lagerspetz, 2025, 250). Stoga naše znanje o razmišljanju drugih umova ostaje nesigurno nagađanje. Sven Nyholm ističe kako trenutno problem drugih umova prilikom razmatranja epistemološkog statusa UI-e zaslužuje našu pažnju, posebice u budućnosti kada će se razvijati složeniji oblici UI-e. Važno je istaknuti kako Nyholm i tehnološki stručnjaci upozoravaju kako su ljudi danas općenito previše naivni u vezi s umom robota te u vezi s potencijalom tehnoloških tvrtki u razvoju proizvoda za koje tvrde da imaju um. Zapravo se radi o varljivim uređajima koji nemaju ništa slično ljudskim umovima (Nyholm, 2019, 593-594).

Do zaključka kako sustavi UI-e i umjetni agenti nemaju razvijenu svijest došla je i Deborah Johnson putem usporedne analize ljudskog moralnog djelovanja i minimalnih koncepata moralne agencije. Johnson tvrdi kako sustavi UI-e ne ispunjavaju minimalne preduvjete koji se u klasičnom pogledu dodjeljuju moralnim agentima i koji su nužni za moralno djelovanje (Johnson, 2006, 199). Analiza moralnog djelovanja upućuje na to kako je pojedinac prvenstveno odgovoran za namjerno dobrovljno djelovanje, a nije odgovoran za djelovanje bez namjere ili za posljedice djelovanja s namjerom koje pojedinac nije mogao predvidjeti. Djelovanje s namjerom je povezano uzrokom koji se razlikuje od uzroka nenamjernog ili nevoljnog djelovanja. Dobrovoljno djelovanje s namjerom shvaća se kao vanjsko djelovanje koje je uzrokovano vrstom unutarnjih mentalnih stanja. Unutarnja mentalna stanja uzrokuju vanjsko djelovanje i zbog toga je djelovanje podložno objašnjenju. Suvremene teorije djelovanja smatraju kako bi ljudsko djelovanje moralo ispuniti pet sljedećih uvjeta kako bi

djelovanje moglo biti podložno moralnoj procjeni. Prvo, mora postojati agent s unutarnjim mentalnim stanjem koje se sastoji od želja, uvjerenja i drugih intencionalnih stanja. Jedno od mentalnih stanja nužno mora imati namjeru djelovanja koja zajedno s ostalim intencionalnim stanjima čine razlog za djelovanje. Drugo, mora postojati vanjski događaj u kojem agent nešto čini i pomiče vlastito tijelo na određen način. Treće, unutarnje stanje je uzrok vanjskog događaja, na primjer, kretanje tijela je racionalno usmjereno na neko vanjsko stanje svijeta. Četvrto, vanjsko djelovanje ima vanjski učinak. Peto, učinak mora biti usmjeren na primatelja djelovanja kojemu se može naštetiti ili pomoći. Ovih 5 uvjeta mogu se koristiti kao okvir za razmatranje moralnog djelovanja sustava UI-e (Johnson, 2006, 198). Analiza moralnog djelovanja nema smisla kada je ono nehotično, kao što su na primjer refleksi, kihanje ili slične tjelesne reakcije. Djelovanje je iskaz slobode i sloboda moral čini mogućim. Takvo razumijevanje ljudskog djelovanja ukorijenjeno je u kartezijanskoj doktrini koja smatra kako su životinje, strojevi i prirodni događaji određeni prirodnim silama te je njihovo djelovanje rezultat nužnosti. Životinje i strojevi se stoga ponašaju mehanicistički jer nemaju slobodu ili vrstu intencionalnosti koja bi ih učinila moralno odgovornima ili prikladnim subjektima moralne procjene (Johnson, 2006, 200).

Klasičan pogled na ljudsko moralno djelovanje, koji nam pomaže u određivanju epistemološkog statusa UI-e, veže se uz četiri preduvjeta koje je potrebno zadovoljiti kako bismo nekoga ili nešto smatrali moralnim agentima: racionalnost, slobodna volja, autonomija, svijest, o čemu smo već pisali ranije u tekstu (Zafar, 2025, 2607). Podsjetimo, jedna od najširih definicija agenta glasi kako je agent sve što se može smatrati da opaža svoju okolinu putem senzora i djeluje na tu okolinu putem aktuatora (Russell, Norvig, 2021, 54). Umjetni moralni agenti mogu biti fizički utjelovljeni roboti ili samo softverski agenti,[16] a postoji nekoliko različitih konceptualnih klasifikacija moralnih agenata (Misselhorn, 2022, 31).

Mogli bismo reći kako se epistemološki status sustava UI-e u bitnome određuje u odnosu na dvije glavne odrednice: Možemo li stvoriti umjetne moralne agente? Trebamo li težiti umjetnim moralnim agentima? U tom kontekstu možemo razlikovati 4 temeljna stava znanstvenika u odnosu na mogućnosti i preferencije razvoja umjetnih moralnih agenata:

1) Nije moguće stvoriti umjetnog moralnog agenta niti je poželjno (van Wynsberghe, Robbins, 2019, 731-732).

[16] Iako ontološki označavaju isto, terminološki se u literaturi se često naizmjenično koriste pojmovi poput „entitet“, „stroj“, „računalo“ i „robot“. Vidi: Behdadi, Munthe, 2020, 196.

2) Nije moguće stvoriti umjetnog moralnog agenta, ali je poželjno (Brundage, 2014, 369).;
3) Moguće je stvoriti umjetnog moralnog agenta, ali nije poželjno (Serafimova, 2020, 8);
4) Moguće je i poželjno stvoriti umjetnog moralnog agenta (Formosa, Ryan, 2021, 850; Gibert, 2023, 143; Wallach, Allen, 2009, 16).

Znanstvenici koji smatraju kako nije moguće niti poželjno stvoriti umjetnog moralnog agenta smatraju kako bismo umjesto umjetnih moralnih agenata težnju trebali usmjeriti na razvoj sigurnih robota jer razlozi za stvaranje umjetnih moralnih agenata, kao što su: sprječavanje nemoralne upotrebe, jačanje javnog povjerenja, bolje razumijevanje ljudskog morala, ne opravdavaju niti jamče da će umjetni moralni agenti moći funkcionirati u praksi (van Wynsberghe, Robbins, 2019, 731). Smatraju kako umjetni moralni agenti neće moći funkcionirati u praksi zbog inherentne pristranosti u procesu učenja etičnosti sustava UI-e i nemogućnosti razumijevanja načina kako sustavi UI-e dolaze do predviđanja (van Wynsberghe, Robbins, 2019, 732).

Znanstvenici, koji smatraju kako je poželjno, ali nije moguće stvoriti umjetnog moralnog agenta, ističu ograničenja vezana uz konceptualizacije morala i izgradnje svijesti, osjećajnosti ili slobodne volje. Izazovi konceptualizacije morala, koji su se pojavili još u vrijeme Asimova, danas možda mogu biti nadvladani na isti način kako smo naučili sustave UI-e igrati igru Go. Konceptualizacija morala može biti nadvladana i pomoću razvoja opće UI-e no potrebno je uložiti daljnje napore u jasnije razumijevanje i kontrole opće UI-e (Brundage, 2014, 369).

Silviya Serafimova smatra kako je moguće stvoriti umjetnog moralnog agenta, ali poteškoću predstavlja odabir između deontološke, utilitarističke i areatičke etike te redukcionističke rizike svođenja inteligencije na brojčanu vrijednost kroz računarski postupak u moralnom rasuđivanju. Kroz analizu metodološke neusklađenosti do koje dolazi između računalnog procesa predviđanja i moralnog rasuđivanja Serafimova zaključuje kako niti jedan oblik umjetnih agenata ne dolazi do zadovoljavajućih razina procjena etičkih situacija (Serafimova, 2020, 8).

Formosa i Ryan ističu kako je moguće i poželjno izrađivati umjetne moralne agente jer će nam pomoći smanjiti štete i poboljšati društvene ishode, posebice u situacijama kada se autonomni sustavi već upotrebljavaju kao što je kod autonomnih vozila ili pružanje zdravstvene skrbi putem robota. Iako smatraju da je pri razvoju umjetnih moralnih agenata oprez potreban, navode kako imamo snažne razloge za nastavak odgovornog razvoja umjetnih moralnih agenata u određenim područjima, kao što je to zdravstvo (Formosa, Ryan, 2021, 850). Gibert ističe kako

su mogući i poželjni umjetni moralni agenti čija se arhitektura temelji na aretaičnom pristupu koji pruža početak tehničkog rješenja za izazov moralne percepcije. Izrada umjetnog moralnog agenta temeljenog na aeratičkoj etici prema Gibertu znači odabir određene koherentnosti jer nam omogućuje rješavanje i problema vezanih uz ispravno djelovanje i problema vezanih uz percepciju onoga što je moralno relevantno (Gibert, 2023, 143). Wallach i Allen također ističu kako će posljedice sve većeg broja delegiranih zadataka sustavima UI-e i strojevima voditi do toga kako je neizbježno da sustavi UI-e i strojevi moraju donositi moralne odluke. Stoga će u budućnosti biti nužno stvarati umjetne moralne agente koji mogu donositi ispravne moralne odluke (Wallach, Allen, 2009, 16). Postoje i drugi autori, poput Andersona i Andersona koji smatraju kako postoje minimalno 3 snažna razloga za razvoj umjetnih moralnih agenata: prvo, djelovanje umjetnih moralnih agenata ima etičke posljedice pa bi bilo poželjno da slijede moralna načela, drugo, agenti bi mogli smanjiti strah ljudi prema autonomnim strojevima, treće, pokušaj izgradnje umjetnih agenata pomaže nam razumjeti etiku (Anderson, Anderson, 2007, 16).

Potpuni moralni agent trebao bi prema Purvesu i suradnicima biti sposoban moralno djelovati na temelju razloga. Djelovati na temelju razloga podrazumijeva 2 različita teorijska pristupa: ili djelovanje na temelju uvjerenja i želja, ili djelovanje na temelju prepoznatljivog mentalnog stanja (kao što su autonomija, slobodna volja, intencionalnost). Oba pristupa djelovanja na temelju razloga zahtijevaju ljudski um i svijest, što sustavi UI-e ne posjeduju (Purves i sur., 2015, 860–61). Zbog toga sustavi UI-e ne mogu biti klasificirani kao potpuni moralni agenti.

Postoje nastojanja u kojima se znanstvenici zalažu za koncept meta-odgovornosti koji predviđa dodjeljivanje sve veće odgovornosti i autonomije u djelovanju umjetnih agenata, zbog čega bismo morali razvijati umjetne moralne agente. Koncept meta-odgovornosti postaje sve relevantniji u sustavima UI-e koji se razvijaju gdje se autonomija može mijenjati tijekom vremena, a delegiranje zadataka budućim umjetnim agentima nosi etički rizik, kao što je to slučaj u zdravstvu (Terziyan i sur., 2025, 5171).

Iako se u literaturi često pojavljuje argument emocija koje su potrebne za pretpostavku o postojanju mentalnih stanja, Coeckelbergh smatra kako je taj argument prividan te je posjedovanje moralno relevantnih emocija manje relevantno nego što nam se to čini. Coeckelbergh tvrdi kako moral može ovisiti o emocijama, ali nije nužno imati stvarna mentalna stanja kako bi netko bio percipiran kao moralno biće, dovoljno je da izgleda kao da ih ima. Ukoliko bi se sustavi UI-e i roboti mogli dizajnirati tako da se uvjerljivo čini kako se ponašaju na način koji sugeriraju da imaju moralno relevantne emocije, to bi trebalo biti dovoljno za

kvalifikaciju moralnog agenta prema Coeckelberghu (Coeckelbergh, 2010, 239). Zbog toga argument o postojanju emocija kao pretpostavke za opravdanje epistemološkog statusa UI-e nije dovoljno uvjerljiv kako bismo ga uzeli u obzir za razmatranje.

Valja istaknuti kako tvrditi da su umjetni moralni agent mogući ne znači da ih je tehnički jednostavno postići. Slažemo se s Purvesom i suradnicima kako strojna etika trenutno ne pruža zadovoljavajuća rješenja za kodifikaciju moralnih pravila. Etika je kontekstualno osjetljiva i zahtjeva promišljeno moralno rasuđivanje pa se moralno djelovanje u stvarnom životu ne može jednostavno svesti na popis pravila koja bi se mogla mehanički slijediti (Purves i sur., 2015, 856-857). No, prije svega se slažemo s Nyholmom koji upozorava kako sustavi UI-e i roboti nemaju mentalna svojstva slična onima koje imaju ljudska bića te sustavi UI-e i roboti ne mogu biti u potpunosti odgovorni za predviđanja ili djelovanje do kojeg dolaze. Iako neki sustavi UI-e danas donose odluke, kao što je to slučaj kod autonomnih vozila, čak i ako bismo u budućnosti uspjeli stvoriti potpunog moralnog agenta, ne bismo mu trebali pružiti donošenje odluka koje se tiču odluka o životu i smrti ili situacijama koje se tiču privatnog života (Nyholm, 2020, 157). Slažemo se sa Zafar kako donošenje odluka o životu i smrti predstavljaju visokorizične situacije i odgovornost za posljedice takvih odluka ne možemo prenijeti na sustave UI-e bez nejasnoća i proglašavanja implicitnih epistemoloških izazova koje smo ranije naveli (Zafar, 2025, 2606). Slažemo se i s drugim znanstvenicima koji ističu kako odgovornost za djelovanje strojeva u ovom trenutku treba biti na ljudskom biću (Johnson, 2006, 203; Véliz, 2021, 495; Sparrow, Flenady, 2025, 2). Pa ipak, čini se kako će u budućnosti zbog sve većeg uključivanja sustava UI-e u procese donošenja odluka sustavima UI-e biti pruženo sve više autonomije. Važno je pri takvom razvoju sustava UI-e voditi računa o tome da ne dolazi do blokiranja pripisivanja odgovornosti ljudskim bićima ili ne dovodi do toga da se pojedinci nepravedno smatraju odgovornima za djelovanje nad kojim nemaju kontrolu, što je bit rascjepa odgovornosti o čemu smo ranije pisali. Takav razvoj imao bi za cilj stvoriti sustave UI-e koji blisko aproksimiraju autonomno ljudsko ponašanje i vrijednosti. Autonomija u računalnoj znanosti se odnosi na sposobnosti računala da slijedi složeni sustav UI-e kao odgovor na utjecaje iz okoliša, neovisno o utjecaju ljudi u stvarnom vremenu. (Etzioni, Etzioni, 2016, 149).

Osim toga, s obzirom kako je potpuno zasebno djelovanje čovjeka ili potpuno zasebno djelovanje stroja poprilično nesvakodnevna situacija, realističnije je očekivati kako bismo neke od izazova epistemološkog statusa UI-e mogli riješiti kroz suradnju timova ljudi i strojeva. Uobičajeno je da ljudi rješavaju određene zadatke uz pomoć strojeva, a strojevi dobivaju podatke od ljudi kako bi izvršili određene zadatke na ljudski zahtjev. Timovi ljudi i strojeva bi

mogli zadovoljiti određene epistemološke kriterije mentalnih stanja poput svijesti, autonomije, intencionalnosti. Osim toga, preciznije bismo mogli ustanoviti odgovornost ljudi i potencijalnih moralnih agenata. Uspostavljanjem timova ljudi i strojeva moguće je postići optimalnu suradnju: strojevi mogu preuzeti zadatke u kojima su posebno učinkoviti, kao što su kognitivne sposobnosti analize i prepoznavanje uzoraka, dok se ljudi mogu usredočiti na područja koja zahtijevaju svijest, autonomiju i intencionalnost.

6. Etičko-epistemološki izazovi pri transplantaciji bubrega

Prema Andreasu Holzingeru i Williamu Rapaportu, suradnja timova ljudi i strojeva predstavlja jedan od potencijalnih načina za optimalno korištenje njihovih sposobnosti i unapređenje zdravstvene dijagnostike. Liječnici, prema Holzingeru, pokazuju dobru intuiciju za niskodimenzionalne probleme (s manjim brojem varijabli ili faktora) gdje mogu dobro generalizirati na temelju male količine podataka i prepoznavati simptome pacijenata zahvaljujući svakodnevnoj praksi. Sustavi UI-e u zdravstvu s druge strane se pokazuju efikasni za visokodimenzionalne probleme (s velikim brojem varijabli ili faktora) koje niti jedan liječnik ne bi imao sposobnosti ili vremena obraditi (Holzinger, 2021, 437). Pri tome, Holzinger ističe kako je važno da sustavi UI-e koji pružaju podršku budu objašnjivi zbog regulatornih zahtjeva u visokorizičnim situacijama kao što je zdravstvo (Holzinger, 2021, 427). William Rapaport ističe kako postoje i druge brojne situacije u kojima čovjek surađuje sa strojevima kao što je na primjer prilikom slijetanja zrakopolova ili upravljanja morskim plovilima pri čemu se događaju nepredvidive situacije koje zahtijevaju i ljudsku interpretaciju. Takve situacije koje zahtijevaju suradnju strojeva i ljudi možda je teže postići nego potpunu automatizaciju strojeva, ali Rapaport smatra kako bi takva suradnja trebala ostati stvarnost (Rapaport, 2023, 399). Takav zaključak o suradnji čovjeka i stroja nam je potreban za naše posljednje poglavlje gdje ćemo analizirati upotrebu OUI-e i etičko-epistemološke izazove primjene UI-e u kontekstu transplantacije bubrega.

Već smo upozorili kako sustavi UI-e koji rade prema principu „crne kutije“ generalno predstavljaju problem za razvoj povjerenja u zdravstvu i uzrokom su nerazumijevanja načina rada sustava UI-e. Osim toga, detaljno smo analizirali kako princip „crne kutije“ djeluje na ostvarivanje informiranog pristanka u zdravstvu. Upozorili smo i na druge izazove do kojih dolazi prilikom primjene sustava UI-e u zdravstvu kao što su pristranosti, nepravda, neusklađivanje vrijednosti između sustava UI-e i ljudi. U ovom poglavlju ćemo ispitati etičke i

epistemološke izazove u kontekstu transplantacije bubrega te analizirati na koji način objašnjivost sustava UI-e utječe na etičnost odluka o dodjeli organa. Osim toga, provjerit ćemo do koje mjere su liste koje generira sustav UI-e transparentne i objašnjive. Razmotrit ćemo i situacije u kojima se lista koju generira sustav UI-e razlikuje od one koju formira ljudski stručnjak te analizirati kriterije koji se koriste za sastavljanje liste čekanja.

Javnozdravstveni statistički podaci navode kako je kronična bubrežna bolest glavni globalni javnozdravstveni problem koji pogađa otprilike 800 milijuna ljudi diljem svijeta, odnosno gotovo 10% ukupne populacije (Kovesdy, 2022, 7). Upravo zbog takve raširenosti bolesti, izazov odlučivanja o dodjeli ograničenih resursa pri transplantaciji bubrega postaje osobitno važan i ilustrira etičke dileme u zdravstvu.

Povijesno gledano, u 1960-ima, prije nego što je postala raširenija i dostupna u praksi, hemodijaliza se rutinski uskraćivala. S ciljem utvrđivanja prioriteta, u Seattleu je 1962. godine osnovan odbor pod službenim nazivom „Odbor za prijem i pravila“[17] koji je odlučivao o tome tko će primati hemodijalizu, a tko ne. Ovaj odbor je javnost brzo prozvala „Božji odbor“[18] zbog dramatične uloge koju je odbor imao pri odabiru liječenja koje je spašavalo živote pacijenata te odlučivalo o tome tko će živjeti, a tko umrijeti (Roberts, 2014, 154). Ovaj povijesni primjer bio je izuzetno važan u povijesnom razvoju bioetike te istaknuo potrebu za pravednijim i transparentnijim načinom odlučivanja u zdravstvu.

Procjenjuje se kako više od 4 milijuna ljudi diljem svijeta prima hemodijalizu sa svrhom liječenja kronične bubrežne bolesti (Buturovic Ponikvar i sur., 2025, 1135). Iako dijaliza produljuje život pacijenta, petogodišnja stopa preživljavanja nakon početka dijalize iznosi oko 40% (Flythe, Watnick, 2024, 1559), a kvaliteta života pacijenata često je značajno pogođena (Abecassis i sur., 2008, 472). Transplantacija bubrega jest zlatni standard za liječenje terminalne faze kronične bubrežne bolesti i pruža superiorne ishode preživljavanja i kvalitete života u usporedbi s dijalizom (Khan i sur., 2025, 1). Zbog toga se svaki pacijent s terminalnom fazom kronične bubrežne bolesti smatra kandidatom za transplantaciju bubrega ukoliko za to ne postoje kontraindikacije (Knoll i sur., 2005, S2), usprkos raznim izazovima s kojima se zdravstveni djelatnici i pacijenti suočavaju pri transplantaciji bubrega, kao što su kirurški zahvat i imunosupresivna terapija (Kuttymuratov i sur., 2025, 1). Darivatelj organa može biti preminula osoba s dokazanom moždanom smrti prema hardvarskim kriterijima (Wijdicks,

[17] Engl. *Admissions and Policy Committee*
[18] Engl. *God Committee*

2018, 137) te živa srodna ili nesrodna osoba (Reis Pereira i sur., 2022, 1). Veća uspješnost transplantacije postiže se kod živih darivatelja zbog kraćeg vremena hladne ishemije[19] (Lauronen i sur., 2020, 42) jer su operativni zahvati eksplantacije i transplantacije organa vremenski usklađeni što rezultira boljom funkcijom presađenog organa (Olawade i sur., 2025, 19). Usprkos prednostima uspješnosti transplantacije sa živih darivatelja u dugoročnom pogledu, većina transplantiranih bubrega dolazi od preminulih darivatelja, a osim toga stariji primatelji imaju manju vjerojatnost da će primiti transplantaciju od živih donora nego mlađi (Toapanta i sur., 2024, 1). Detalji kirurškog postupka opisani su u priručnicima kao što su *Kidney transplantation: Principles and practice* (Knechtle, Marson, 2025) ili *Contemporary Kidney Transplantation* (Ramirez, McCauley, 2018). Prije same transplantacije bubrega, darivatelj i primatelj prolaze kroz opsežnu obradu kako bi se ustanovili podudarnost kriterija, kvaliteta organa te otkrila eventualna patologija ili kontraindikacija koji bi mogli ugroziti transplantaciju. Razdoblje pomaže u optimizaciji procesa transplantacije, ključno je za uspješnost transplantacije te pomaže u otklanjanju komplikacija do kojih može doći nakon transplantacije (Tantisattamo, Maggiore, 2024, 1425). Neki znanstvenici ističu kako priprema transplantacije bubrega može započeti prekasno, a vrijeme potrebno za njezin dovršetak iznenađujuće dugo. Zbog toga predlažu internu reviziju procedura svakog transplantacijskog centra (Dirix i sur., 2022, 1100).

Transplantacija organa smatra se generalno jednim od najvažnijih zdravstvenih dostignuća 20. stoljeća, a otkriće učinkovitih imunosupresivnih lijekova krajem 1970-ih pomoglo je u povećanju uspješnosti transplantacije organa (Ghaly, 2012, 175). Republika Hrvatska je članica Eurotransplanta od 2007. godine, međunarodne neprofitne organizacije koja se bavi alokacijom i prekograničnom razmjenom kadaveričnih organa za transplantaciju (Mikulić, Kocman, 2020, 498). U Hrvatskoj 2024. godine je uspješno izvršena transplantacija 142 bubrega (Ministarstvo zdravstva Republike Hrvatske, 2025). Alokacija bubrega određuje se u skladu sa sustavom bodovanja Eurotransplanta koji se neprestano evaluira, a temeljen je na raznim medicinskim kriterijima i kriterijima koji su usmjereni na procjenu pacijenta, kao što su: vrijeme čekanja, duljina dijalize, geografska udaljenost do transplantacijskog centra (Persijn, 2006, 419).

[19] Ishemija je vrijeme tijekom kojeg je bubreg izvađen iz tijela donora i čuvan na hladnom prije nego se transplantira primatelju. Dugo trajanje hladne ishemije znači kako je bubreg dugo bio bez protoka krvi pa nakon transplantacije dolazi do odgođene funkcije presatka i primatelju je privremeno potrebna dijaliza. Vidi: Lauronen i sur., 2020, 42.

Za što uspješniju transplantaciju bubrega, kao najčešćeg postupka presađivanja organa u svijetu prema statistikama iz SAD-a i Europe (Lewis i sur., 2021, 4), važna je postoperativna skrb koju uvelike olakšavaju minimalno invazivni kirurški zahvati (Uwechue i sur., 2017, 2), ali i kasnije liječenje i praćenje pacijenta. Za poboljšanje uspješnosti transplantacije bubrega razvijeni su razni sustavi UI-e za predviđanje ishoda transplantacije bubrega, poput preživljavanja presađenog bubrega i imunosupresije (Basuli, Roy, 2023, 391-392), te za poboljšanje podudaranja donora i primatelja (Connor i sur., 2021, 726), a razvijeni su sustavi OUI-e za dijagnozu i predviđanje odbacivanja presatka bubrega, odgođene funkcije presatka, zatajenja presatka i drugih povezanih ishoda nakon transplantacije bubrega (Assis de Souza i sur., 2025, 123).

Važno je uočiti kako globalno tržište dijalize, koje je vrijedno nekoliko milijardi dolara, cvjeta, dok se područje transplantacije organa bori s financiranjem, dostupnošću i političkom podrškom (Muller, 2025, 902). Uz procijenjena četiri miliona pacijenata koji se liječe dijalizom, stručnjaci zagovaraju stav kako je osnovno pravo pacijenta liječenje bubrežnog zatajenja, a ne samo prava na transplantaciju bubrega (Buturovic Ponikvar i sur., 2025, 1136). Liste čekanja za sve dobne skupine smatraju se predugima zbog nedostatka donacija organa, i od preminulih i od živih donora, a nedostaju i sredstva za stvaranje i održavanje jedinica intenzivne njege, transplantacijskih ustanova i posttransplantacijske skrbi. Stoga, čak i povećanjem doniranja organa, s obzirom na trenutnu demografiju zatajenja bubrega, neće doći do skorog pražnjenja odjela za dijalizu (Caldwell i sur., 2025, 9). Transplantacija bubrega, posebno tamo gdje se primjenjuje zdrava politika proširenih kriterija, uključujući starije donore, donore s neoptimalnom funkcijom bubrega ili donaciju nakon kardiocirkulacijske smrti, vjerojatno neće biti doživotni zdravstveni postupak, posebno za mlade pacijente (Aubert i sur., 2015, 2).

Ipak, za brojne pacijente koji boluju od kroničnog zatajenja bubrega ne postoji dovoljno odgovarajućih donora i nedostatak organa što je najveći izazov s kojim se danas suočava transplantacijska medicina i zbog toga neki znanstvenici predlažu propitivanje kriterija sa svrhom proširenja baze donora organa (Saidi, Hejazii Kenari, 2014, 87). Osim toga, zbog neravnoteže između potražnje za organima i dostupnom ponudom bubrega se javlja interes za proučavanjem pravedne alokacije organa (Bunnik, 2023, 192; Yeung i sur., 2022, 1).

Do sada su pokušani različiti pristupi s ciljem proširenja baze donora organa, uključujući darivanje organa živih donora, proširenje darivanja od preminulih donora, podijeljeno darivanje organa (Howard i sur., 2005, 565), nacionalni modeli dijeljenja i veće korištenje donora s proširenim kriterijima (Klein i sur., 2010, 978). Propisani su različiti programi, politike i

kriteriji za transplantaciju bubrega koji se u literaturi često nalaze pod sintagmom sustav raspodjele bubrega[20] i koji su izrađeni od konzorcija stručnjaka iz područja nefrologije i drugih zdravstvenih područja koji vrijede u različitim regijama svijeta kao što su SAD-e (Lentine i sur., 2025, 1), Europa (Assfalg i sur., 2024, 1200) ili globalno (Chadban i sur., 2020, S11). Evaluacija tih programa, politika i kriterija se često pojavljuju u sve savršenijem obliku s ciljem unapređenja zdravstvenih aspekata postupka transplantacije bubrega (Freitas, 2011, 1; Boerstra i sur., 2025, 1; Mudiayi, i sur., 2022, 1113; Zambeli-Ljepović i sur., 2025, e1691). Postupak procjene i utvrđivanja prikladnosti pacijenta za transplantaciju bubrega je složen i dugotrajan, a čak i nakon što se potvrdi podobnost vrijeme provedeno na listi čekanja može se produljiti (Watters i sur., 2024, 1444). Usprkos postojanju stručnih programa, politika i kriterija, možemo reći kako postoje zamjetne nejednakosti (rasne, etničke, geografske) koje mogu utjecati na pogoršanje socioekonomskih razlika kod pacijenata zbog čega su programi, politike i kriteriji za transplantaciju bubrega u kontinuiranom razvoju od strane brojnih organizacija s ciljem pružanja jednakog pristupa transplantaciji bubrega (Boerstra i sur., 2025, 1).

Postoje različiti etički i epistemološki izazovi transplantacije bubrega koji se ponekad obrađuju u znanstvenoj literaturi i koji ističu potrebu za transparentnim i pravednim pristupom transplantaciji bubrega, bez diskriminacije na temelju rase, spola ili geografskog položaja (Rebollo Mateos, López Castro, 2025, 11). Najčešći etički i epistemološki izazovi koji se spominju u kontekstu transplantacije bubrega tiču se analize principa pravednosti, neškodljivosti, autonomije, dobročinstva, motivacije donora te analize informiranog pristanka (Martinelli i sur., 2024, 1), ali proučavali su se i izazovi pristranosti i uloga financijskih poticaja u kontekstu transplantacije bubrega (Parker, Chin, 2020, 410-411), te neetične prakse, uključujući iskorištavanje ranjivih skupina stanovništva, ilegalnu trgovinu organima i korištenje organa pogubljenih zatvorenika što je zabranjeno Istanbulskom deklaracijom (Buturovic Ponikvar i sur., 2025, 1136). Iz etičke perspektive su analizirane rasne i etničke nejednakosti do kojih dolazi u pristupu transplantaciji bubrega i dolasku na listu čekanja zbog rasne ili etničke pripadnosti pacijenta (Clark-Cutaia i sur., 2024, 2). Osim toga, istraživane su nejednakosti koje su uzrokovane geografskim položajem pacijenata u SAD-u te se geografske nejednakosti smatraju neprihvatljivo visokima (Zhou i sur., 2018, 7). Međutim, rijetki su radovi posvećeni etičkim i epistemološkim izazovima primjene sustava UI-e u području transplantacije bubrega (Robinson, 2022, 2). Zbog toga će se naša analiza usmjeriti upravo na

[20] Engl. *Kidney Allocation System*

etičke i epistemološke izazove primjene UI-e u području transplantacije bubrega na primjeru američkog Sustava raspodjele bubrega (engl. *Kidney Allocation System*), ali ćemo u obzir uzeti i kriterije drugih sustava raspodjele bubrega.

6.1. Kriteriji za transplantaciju bubrega

Do sada su raspisani brojni kriteriji putem smjernica ili politika različitih država i svjetskih regija koje moraju zadovoljiti primatelj i donor bubrega. Kriteriji se razmatraju kroz nekoliko različitih elemenata, kao što su pristup transplantaciji, sociodemografski i zdravstveni čimbenici pacijenta te imunološka i psihosocijalna procjena. U obzir se uzimaju različiti čimbenici rizika i komorbiditeta koji utječu na prikladnost pojedinca za transplantaciju, kao što su pridržavanje terapije, pušenje, dijabetes, pretilost, uzroci zatajenja bubrega i slično (Chadban i sur., 2020, S1). Najčešći kandidati za transplantaciju bubrega su pacijenti s terminalnom fazom kronične bubrežne bolesti brzinom glomerularne filtracije <15 mL/min/1,73 m² (brzina glomerularne filtracije je mjera kojom se izražava koliko kvalitetno bubrezi filtriraju krv, što je broj manji, funkcija filtracije je lošija), ili pacijenti s kroničnom bubrežnom bolešću stadija 4 koji pokazuju znakove progresije bolesti i brzina glomerularne filtracije iznosi 15–29 mL/min/1,73 m² (Novacescu i sur., 2024, 2).

Detaljne preporuke za kriterije za transplantaciju bubrega nalaze se u KDIGO smjernicama i obuhvaćaju 19 sljedećih elemenata: pristup transplantaciji, dob, pedijatrijska procjena, psihosocijalna procjena, pridržavanje preporuka, konzumacija duhana, kirurška pitanja koja uključuju i pretilost, dijabetes, uzrok završnog stadija bubrežne bolesti, infekcije, procjena maligniteta, plućne bolesti, srčane bolesti, periferna arterijska bolest, neurološke bolesti, bolesti probavnog sustava i jetrene bolesti, hematološki poremećaji, metabolizam kostiju i minerala, imunološka procjena (Chadban i sur., 2020, S28). U SAD-u, Sustav za raspodjelu bubrega prioritet dodjeljuje kandidatima za transplantaciju na temelju kliničkih i imunoloških čimbenika, uključujući vrijeme čekanja, osjetljivost imunološkog sustava i podudarnost donora bubrega s očekivanim životnim vijekom primatelja (Yacoub i sur., 2018, 484). Pedijatrijski i hitni slučajevi dobivaju dodatnu važnost pri odabiru prioriteta, a geografska udaljenost također utječe na alokaciju bubrega. HLA podudarnost ima ograničen učinak na određivanje prioriteta jer je potvrđeno kako HLA podudarnost može utjecati na način da će bijeli pacijenti biti pristrano favorizirani u odnosu na pacijente druge boje kože (Roberts i sur., 2004, 550).

Postoje i kriteriji na temelju kojih se kandidat isključuje iz doniranja bubrega na temelju kontraindikacija i znanstvenih podataka o medicinskom riziku, kirurškom riziku, psihološkoj procjeni ili konsenzusu. Nacionalne smjernice za obradu i procjenu primatelja i darivatelja bubrega Republike Hrvatske tako definiraju čimbenike rizika za darivanje bubrega te savjetuju odustajanje od darivanja bubrega u slučajevima kada je prisutno više od jednog rizičnog čimbenika: hipertenzija, pretilost, proteinurija, oštećena tolerancija glukoze, hematurija. (Ministarstvo zdravstva Republike Hrvatske, 2016, 33). Apsolutne kontraindikacije za transplantaciju bubrega su nemogućnost podnošenja operativnog zahvata zbog teške srčane ili plućne bolesti, neliječenih malignih bolesti, aktivnih infekcija, neliječenog HIV-a, aktivne zlouporabe droga, nekontrolirane psihijatrijske bolesti, medicinska stanja s očekivanim životnim vijekom manjim od 2 godine i maligni melanom unutar 5 prethodnih godina (Novacescu i sur., 2024, 2). Relativne kontraindikacije su morbidna pretilost (preporučeni indeks tjelesne mase <40 kg/m2), povijest nepoštivanja rasporeda dijalize ili režima lijekova, slabost, komorbiditeti, infekcija virusom hepatitisa B ili C, prethodne maligne bolesti (Abramowicz i sur., 2015, 2-8).

Utvrđivanje kriterija za transplantaciju bubrega zaista je važan proces koji se dodatno regulira uravnotežujući etička načela korisnosti (maksimiziranja koristi od transplantacije), jednakosti (jednakog pristupa transplantaciji) i poštovanja osoba. Iako se može reći kako su biomedicinski kriteriji propisani poprilično detaljno, posebno je zanimljivo kako neki autori ističu da različiti moralni i sociodemografski kriteriji nisu dovoljno jasni, mogu ovisiti o interpretaciji te mogu uzrokovati nejednakosti prilikom izrade liste čekanja (Oedingen i sur., 2020, 677-678). Takvi nedovoljno jasni kriteriji tiču se, na primjer, informacija o pacijentovom načinu života te autodestruktivnim mehanizmima ponašanja, kao što su konzumacija alkohola ili duhana. Iako su ove informacije relevantne za donošenje odluka o alokaciji organa, zbog njih se može razviti kontroverzna rasprava. Tako su sudionici jednog od istraživanja o alokaciji organa istaknuli kako zlouporabu alkohola ili duhana smatraju ovisnošću što bi trebalo biti klasificirano kao bolest. Međutim, s obzirom na nedostatak organa generalno, možda ne bi bilo etično prednost dati ovisniku jer bi to impliciralo kako drugi pacijent koji nije razvio ovisničko ponašanje neće moći dobiti organ. Zatajenje organa uzrokovano konzumacijom alkohola ili cigareta povezuje se s nižim stopama preživljavanja nakon transplantacije. Neki sudionici istraživanja željeli su izbjeći dilemu uključivanja pacijenata s ovisničkim ponašanjem na listu čekanja, pa su koristili zamjenske pokazatelje učinkovitosti i primjetili kako su stope preživljavanja slične kod pacijenata s ovisničkim ponašanjem kao i kod pacijenata bez takvog ponašanja (Oedingen i

sur., 2020, 677). Osim navedenog, znanstvenici su zaključili kako bi se moralni i sociodemografski kriteriji poput vremena čekanja, životnog stila, reciprociteta i sociodemografskog statusa trebali koristiti kao dodatni ili zamjenski kriteriji. Predlažu i da se spol, financijski status ili obiteljske odgovornosti obuhvate načelom jednakosti (Oedingen i sur., 2020, 676).

U drugom istraživanju, skupina autora koji su analizirali smjernice Kanadskog društva za transplantaciju ocjenjuju kako uspostavljeni kriteriji nisu empirijski testirani te obrazloženja za definicije kontraindikacija smatraju varijabilnim. Navode primjer kako bi imunosupresija bolesnika s aktivnom infekcijom ili rakom uzrokovala veću štetu nego korist. Isti autori napominju kako su neke navedene kontraindikacije potencijalno reverzibilne (nepridržavanje terapije, aktivna bolest bubrega), a odgođeno vremensko razdoblje moglo bi omogućiti bolji relativni ishod pacijentima (Kiberd i sur., 2015, 1).

S obzirom kako je više od 40% pacijenata u SAD-u u terminalnoj fazi kronične bubrežne bolesti posebno je važno kako klinički stručnjaci smatraju da je etički neprihvatljivo alokaciju organa vršiti isključivo na temelju očekivanog razdoblja života pacijenata. Iako se radi o više od 40% pacijenata u terminalnoj fazi bolesti u dobi od 65 ili više godina, samo je 20% svih primatelja bubrega što predstavlja zabrinutost za primjenu kriterija transplantacije bubrega (Shrestha i sur., 2022, 2). Krhkost i kognitivno oštećenje mjere su fiziološkog i kognitivnog starenja koje su prikladne za procjenu podobnosti starijih osoba koje se podvrgavaju procjeni za transplantaciju bubrega. Iako se krhkost i kognitivno oštećenje ponekad procjenjuju, nedostaju etičke smjernice o tome kako pravilno koristiti ove mjere pri donošenju odluka o stavljanju na listu čekanja. Iako u SAD-u postoje smjernice za dodjelu organa koje proizlaze iz Mreže za nabavu i transplantaciju organa, te smjernice ne adresiraju kako točno uravnotežiti etička načela korisnosti, jednakosti i poštovanja osoba. Također, smjernice ne propisuju hoće li se pacijent koji je krhak ili ima kognitivno oštećenje uvrstiti na listu čekanja ili ne zbog čega posljedično postoje nejednakosti u listama čekanja (Shrestha i sur., 2022, 2).

Neki autori pozivaju na implementaciju sustava UI-e s ciljem poboljšanja podudaranja donora i primatelja, posttransplantacijsku skrb te smanjenje odbacivanja organa identifikacijom podudaranja i predviđanjem posttransplantacijskih komplikacija (Tantisattamo, Maggiore, 2024, 1427). Implementaciju sustava UI-e smatramo opravdanom jer posljednih godina sustavi UI-e u području transplantacije organa pružaju poboljšane prediktivne performanse za preživljavanje presađenog organa i pacijenta u usporedbi s konvencionalnim metodama (Cheah i sur., 2025, 2). Poboljšane prediktivne performanse mogu biti dovoljan razlog kako bi netko

pomislio da su predmet donošenja predviđanja sustava UI-e i dizajn samih sustava UI-e isključivo tehničke prirode, te kako o tome isključivo trebaju voditi računa programeri i stručnjaci iz tehničkog područja. Ali to nije sasvim točno jer su sustavi UI-e više od izvođenja matematičkih formulacija. Sustavi UI-e nisu izgrađeni samo na podacima nego i na ljudskoj prosudbi o tome koje su vrijednosti najvažnije i na koji način izraziti te vrijednosti (Robinson, 2022, 1). Jedno od vodećih izazova etike i epistemologije primjene UI-e u području transplantacije bubrega tiče se usklađivanja vrijednosti, to jest, kako uskladiti sustav UI-e koji odlučuje o transplantaciji bubrega s našim vrijednostima, o čemu smo ranije već pisali. Osim toga, analiza i usklađivanje vrijednosti u sustave UI-e postaje sve važnija znanstvena tema jer upućuje na odnos čovjeka i stroja (Johnson, Verdicchio, 2025, 67; van de Poel, 2020, 385; Coeckelbergh, 2020A, 167; Conwill i sur., 2024, 25).

6.2. Dvije mogućnosti rangiranja

Istraživanje iz Australije upozorava kako se sustavi UI-e za raspodjelu organa od preminulih donora razvijaju s nedovoljnim unosom vrijednosti pacijenata i temelje se na kombinaciji sljedećih kriterija: vrijeme čekanja, medicinska hitnost, podudaranje humanog leukocitnog antigena (HLA), senzibilizacije i pedijatrijskog statusa. Kao rezultat, istraživači smatraju kako je nedostatak vrijednosti pacijenata dodao nepotrebne kontroverze u formiranje politike za transplantaciju bubrega bez dovoljno zalaganja za pravednu alokaciju organa (Tong i sur., 2012, 1).

Spomenuli smo kako pojedini kriteriji za transplantaciju bubrega mogu uzrokovati nejednakosti pri izradi liste čekanja. Pretpostavimo sada kako postoje dva izvora za izradu liste čekanja: lista A i lista B. Autor liste A je ljudski stručnjak (liječnik) koji listu stvara na temelju biomedicinskih i sociodemografskog kriterija vezanog uz starenje. Autor liste B je sustav UI-e koji listu stvara na temelju biomedicinskih kriterija i predikciji vjerojatnosti odbacivanja presatka. Sociodemografski kriteriji poput čimbenika povezanih sa starenjem klasičan su primjer kriterija koji mogu uzrokovati različit pristup prilikom definiranja liste čekanja i transplantacije bubrega (Boerstra i sur., 2025, 10). Kao rezultat, dobit ćemo dvije liste s različitim prioritetom pacijenata. Lista A uključivat će pacijente starosti 65 godina ili starije koji neće zauzimati posljednja mjesta nego će biti ravnopravni kao i pacijenti mlađe dobne populacije te će pristup transplantaciji biti pružen prema onima koji bi od transplantacije mogli imati koristi, u smislu očekivanog životnog vijeka i poboljšanja kvalitete života. Lista B

uključivat će samo pacijente starosti do 65 godina bez iznimke u dobnoj granici temeljem propisanih biomedicinskih kriterija i predikciji vjerojatnosti odbacivanja presatka. Lista A ne bi automatski isključivala pacijente u dobi od 65 godina ili starije nego bi u obzir uzela procjenu poboljšanja kvalitete života, smanjenu ovisnost o dijalizi i ostale biomedicinske kriterije pacijenta. Lista B isključila bi pacijente starije od 65 godina iz liste čekanja i sustav UI-e bi na taj način maksimizirao predvidljivi medicinski ishod transplantacije, ali bi žrtvovao pravednost i jednak pristup transplantaciji. Dok se Lista A oslanja na postizanje pravednosti i procjenjuje individualnu korist medicinskog postupka, lista B je isključivo usmjerena na optimalne rezultate. Lista A time pokazuje osjetljivost za društvenu pravednost, a ne samo medicinsku učinkovitost te uključuje vrijednosti poput suosjećanja, prepoznavanja konteksta, osobnog pristupa i pravednosti. Lista B naglasak stavlja na maksimiziranje numeričkih ishoda do čega često dolazi na štetu pravednosti i pokazuje kako ne razumije širi kontekst.

Lista A time pokazuje uključivost prema pacijentima 65 godina starosti ili starijima i etički je prihvatljiviji pristup raspodjeli organa. Stručnjaci spomenutog istraživanja u velikoj mjeri se slažu kako je prikladno uvrstiti pacijenta koji ima 65 ili više godina na popis za transplantaciju te predlažu sljedeće čimbenike koje treba uzeti u obzir za procjenu pacijenata koji imaju 65 ili više godina starosti: socijalna podrška, pridržavanje lijekova, dobar funkcionalni status, kvaliteta života povezana s zdravljem, niska ili nikakva krhkost, bez medicinskih kontraindikacija i nisko opterećenje komorbiditetom (Shrestha i sur., 2022, 2).

Lista A time bi bila bliže deontološkim i aretaičkim normativnim teorijama jer liječnik želi spriječiti štetu konkretnoj osobi i štiti integritet svakog pojedinog pacijenta. Deontološka normativna teorija uzima u obzir i širi kontekst te ne diskriminira pacijente starije od 65 godina starosti. Deontološka normativna teorija u skladu je s teorijom pravednosti jer promiče pravednu raspodjelu i štiti ranjive skupine. Lista A se preklapa i sa značajkama aretaičke normativne teorije jer vrline poput pravednosti, brige i suosjećanja naglašavaju uključivost. Etika vrline primjenjuje moralni karakter liječnika, a ne stroga pravila i ishode. Zbog toga, ali i zbog ranije spomenutih značajki kojima se vodi liječnik koji primjenjuje etiku vrline, bliži smo zaključku kako sustavi UI-e u budućnosti trebaju biti razvijani uz podršku etike vrline. Već smo ranije spomenuli kako liječnik koji se vodi etikom vrline može svoju odluku usmjeriti u odnosu na 4 osnovne vrline kao što su: pravednost, poštenje, odgovornost i briga. Ukoliko pri donošenju odluke liječnik nastoji minimizirati diskriminaciju i osigurati pravedan tretman svih pacijenata, kao temelj kriterija za donošenje odluke, može uzeti pravednost kao vrlinu (Hagendorff, 2022, 13).

Lista B najbliže je utilitarističkoj normativnoj teoriji jer maksimizira ukupni zdravstveni rezultat i minimalizira odbacivanje presatka, neovisno o individualnim karakteristikama ili sociodemografskom kriteriju starosti. Takva lista ne uzima dovoljno u obzir pravednost i naglasak stavlja na učinkovitost, a diskriminaciju starijih ne tretira kao relevantan faktor jer analizira ukupnu korist mlađih pacijenata. Takva lista ne pruža objašnjenje za predviđanje do kojeg je sustav UI-e došao što predstavlja poseban izazov u kontekstu zdravstva.

6.3. Objašnjiva umjetna inteligencija i transplantacija bubrega

Sve više znanstvenika upozorava kako tehnološka rješenja u zdravstvu moraju voditi računa o etičkim izazovima koji se pojavljuju zbog upotrebe sustava UI-e u zdravstvu. Sustavi UI-e mogu pogoršati postojeće nejednakosti među pacijentima na temelju ekonomskih čimbenika ili pripadnosti rasi i time dodatno ugroziti pravednost (Solanki i sur., 2023, 227). Isto vrijedi i u kontekstu kriterija koji se primjenjuju za transplantaciju bubrega.

OUI predmet je proučavanja brojnih znanstvenika koji tvrde kako je objašnjivost potrebna za povjerenje, kako UI mora biti objašnjiva da bi bila pouzdana, kako je objašnjivost preduvjet za povjerenje te kako je objašnjivost nužan uvjet za povjerenje. Ovdje ćemo navesti neke od najvažnijih izjava koje potkrepljuju takve i slične stavove: „Kako bi ljudi vjerovali metodama crne kutije, potrebna nam je objašnjivost“ (Gilpin i sur., 2018, 80). „Potreba za objašnjivom umjetnom inteligencijom motivirana je uglavnom s tri razloga: potrebom za povjerenjem“ ... (Fox i sur., 2007, 24). „Umjetni agenti moraju objasniti svoju odluku korisniku kako bi stekli povjerenje...“ (Pieters, 2011, 53). „Osim toga, objašnjivost modela preduvjet je za izgradnju povjerenja i prihvaćanje sustava UI-e“ (Gade i sur., 2019, 3203). „…postoji potreba za objašnjenjem ... kako bi korisnici i donositelji odluka mogli razviti odgovarajuće povjerenje.“ (Mathews, 2019, 1271).

Ranije u tekstu smo naglasili kako pomoću OUI-e zapravo želimo razumjeti način na koji sustav UI-e dolazi do određenog predviđanja, posebno u primjenama koje imaju etičke posljedice kao što je to kod transplantacije bubrega (Bartneck i sur., 2021, 39). Programeri i filozofi i brojne međunarodne zajednice više su puta pozivali na razvoj i implementaciju OUI-e koja promiče jednakost i pravednost (Wulf, Seizov, 2024, 236). Opća uredba o zaštiti podataka (GDPR) Europske unije među prvim je zakonodavnim dokumentima koji regulira etiku UI-e i objašnjivost kroz čl. 13., čl. 14., čl. 15., čl. 21. i čl. 22. (Europski parlament i Vijeće Europske unije, 2016). Osnovna ideja je da kada se odluka donosi isključivo automatiziranom obradom

(bez ljudske intervencije), korisnik ima pravo dobiti objašnjenje o tome kako je odluka donesena. Spominjali smo kako neki znanstvenici smatraju da upravo GDPR potvrđuje pravo pacijenta na smisleno objašnjenje o logici uključenoj u stvaranje predviđanja sustava UI-e (Astromskė i sur., 2021, 515). No, ne slažu se svi s takvim razumijevanjem posebice po pitanju upravljanja rizika koji se razlikuje u tri kategorije: neprihvatljiv rizik, visok i nizak rizik (Finocchiaro, 2024, 1964-1965). Uz to, postoje i drugi pristupi regulacije etike UI-e koja je prikazana unutar usporednog okvira europskog, američkog, kineskog zakonodavnog okvira (Poslon, Čartolovni, 2025, 49-50). Posljedice neslaganja mogu proizlaziti iz takozvane Collingridgeove dileme. Reguliranje takve, još uvijek nove tehnologije predstavlja izazov našeg vremena, poznat kao Collingridgeova dilema, koja opisuje situacije preranog reguliranja tehničkih posljedica u trenutku kada su rizici nejasni. Ova dilema je posebno izražena za UI-e gdje se tehnologija i neizvjesnost brzo mijenjaju, ali Collingridgeova dilema opisuje i situacije rizika neusklađenosti između propisa i rizika koje su oni trebali ublažiti (Emery-Xu i sur., 2025, 3021).

Poznata su tehnička ograničenja OUI u vidu kompromisa koji se uspostavlja između preciznosti i objašnjenja. O tom ograničenju smo pisali kada smo opisivali situacije u kojima sustavi UI-e koji pokazuju jednake ili bolje sposobnosti od liječnika (Shmatko i sur., 2022, 1028) mogu raditi na složen način kojeg ljudi ne mogu razumjeti (Hatherley i sur., 2022, 7). Zbog toga se ponekad događa kako liječnici prednost daju sustavima UI-e koji mogu objasniti predviđanja i način rada sustava UI-e. No poznati su primjeri znanstvenika koji iz moralnih razloga brane objašnjivost, ne samo za situacije primjene OUI-e u zdravstvu, nego i zbog toga što UI koja djeluje na principu „crne kutije“ ne pruža uvijek precizna predviđanja, a dijelom zato što krši proceduralne dužnosti zbog kojih je vjerojatno da će ignorirati relevantne podatke i uključiti podatke koji ne bi trebali biti uključeni (Grant i sur., 2025, 56).

Ponekad se može dogoditi da stručnjaci iz područja zdravstva ili tehnološke domene pokušavaju nametnuti zaključke koji se tiču moralnih izbora. To je bilo uobičajeno u zdravstvu koje se prakticiralo prije 50-ak godina. Filozof s posebnim zanimanjem u području transplantacijske etike, Robert Veatch definirao je 1973. godine sindrom koji se može prepoznati kod pojedinih zdravstvenih stručnjaka i nazvao ga „As-a“[21] sindrom. Do toga dolazi kada liječnik, kirurg ili

[21] Veatch koristi primjer iz engleskog govornog područja koje glasi „As-a-psychiatrist, I cannot condone involuntary psychiatric commitment…“ što bi se na hrvatskom jeziku moglo prevesti na sljedeći način: „Kao psihijatar, ja ne mogu odobriti prisilno psihijatrijsko liječenje…“ Sindrom „As-a“ ovdje upućuje na situaciju u kojoj osoba, iako je stručnjak u jednom području zdravstvene skrbi, bez stvarnog autoriteta pokušava donositi ili nametati stručne zaključke izvan okvira svoje stvarne kompetencije. Vidi: Veatch, 1973, 36.

drugi zdravstveni stručnjak zauzme poziciju moralnog autoriteta pri odlučivanju koji je od medicinskih postupaka dostupnih pacijentu najmudrije slijediti, čak i kada dolazi do protivljenja pacijenta (Veatch, 1973, 36). U takvim situacijama liječnik uzima moralnu ulogu koja, kako Veatch kaže, može pripadati pacijentu, obitelji ili javnosti. Veatch smatra kako u situacijama u kojima liječnik izražava moralno mišljenje može djelovati na pacijenta na specifičan način. Iako pacijent formalno ima pravo odlučiti sam, način na koji liječnik iznosi moralna uvjerenja može neizravno utjecati na pacijenta da se s tim mišljenjem pacijent složi. Naravno da mogu postojati dobri razlozi zbog kojih bi se pacijent mogao složiti s određenim moralnim uvjerenjem i ti razlozi mogu ići pacijentu u korist jer su liječnici promatrali postupke liječenja kroz koje mnogi pacijenti prolaze i mogu biti u dobroj poziciji predvidjeti s kojim moralnim izborima će pacijent ili obitelj kasnije biti sretni što su napravili, a zbog kojih izbora će požaliti. Ali čak i kada moralno uvjerenje liječnika zaslužuje posebnu težinu, posljednja odluka bi trebala ostati na pacijentu (Robinson, 2022, 29).

S obzirom kako danas više nije dovoljno da sustavi UI-e budu samo precizni, učinkoviti ili profitabilni nego je potrebno da budu i pravedni, sustavi OUI-e mogu pomoći u regulaciji pravednosti. Smisao je koristiti OUI-u kako bismo došli do ostvarenja pravednosti što će rezultirati povećanjem povjerenja u korištenje alata UI-e i širu društvenu prihvaćenost digitalnih tehnologija. Potrebno je objašnjenje prilagoditi svakom pojedinom korisniku (Hoffman i sur., 2023, 8). Stoga se općenito pretpostavlja da objašnjenja korištena za algoritamski pristup moraju biti informativna u kontekstu određenog pojedinca ili slučaja (Upadhyay i sur., 2025, 446).

Ne postoji široko prihvaćeno jedinstveno etičko načelo za pravednu raspodjelu organa za transplantaciju. Iako većina sustava raspodjele koristi kombinacije načela raspodjele, postoji posebna napetost između davanja prioriteta onima koji su najpogođeniji bolešću te maksimiziranja ukupne koristi za sve pacijente. Često se sugerira kako bi empirijska istraživanja o javnim preferencijama trebala pomoći u rješavanju dileme između pravednosti i učinkovitosti u donošenju politika raspodjele (Bunnik, 2023, 192).

Razni izazovi konceptualizacije pravednosti poput terminoloških nejasnoća, pristranosti, nepravednosti, nedostatka objašnjivosti i povjerenja su već prikazani u literaturi i objašnjeno je zašto čine pukotine prilikom primjene UI-e u zdravstvu (Poslon, 2025, 281-282). Iako ne postoji puno različitih primjena, OUI može pomoći pri suočavanju s izazovom pravednosti u kontekstu transplantacije bubrega. To je vidljivo u primjeru sustava OUI-e koji koristi objašnjive metode s ciljem identifikacije najvažnijih značajki predviđanja i pravovremenoj identifikaciji sustavnih

pristranosti što podržava pravednu alokaciju organa za sve primatelje presatka (Fan i sur., 2025, 7). SHAP metoda je u ovom istraživanju korištena za povećanje objašnjivosti sustava, a analiza podskupina doprinijela je razvoju pravednosti sustava UI-e modela. Takvi primjeri su dokaz kako sustavi UI-e mogu biti objašnjivi i doprinositi pravednoj alokaciji organa. Naime, analiza doprinosa pojedinih značajki do koje možemo doći putem SHAP-a omogućuje identifikaciju potencijalnih pristranosti kojima sustav UI-e može težiti. Takve pristranosti mogu se, na primjer, odražavati u konačnim predviđanjima ukoliko sustav UI-e prekomjerno favorizira pacijente određene dobne skupine, spolne kategorije ili zdravstvenih karakteristika. S obzirom na to, pomoću analiza doprinosa do koje dolazimo putem SHAP-a, možemo provesti korekcije ili prilagodbe kako bi sustav UI-e bio pravedniji prema svim kandidatima. Dodatno, analiza skupina pacijenata pomaže u razumijevanju načina na koji sustav UI-e tretira različite populacije, čime se osigurava da odluke o raspodjeli organa prednost ne daju jednoj društvenoj skupini na štetu druge. Na taj način SHAP doprinosi ne samo objašnjivosti sustava, već i konkretnoj implementaciji pravednijih i transparentnijih odluka u alokaciji organa.

Iako se transparentnost i objašnjivost široko zagovaraju u primjeni UI-e u zdravstvu, još uvijek nisu dovoljno zastupljeni u području alokacije organa niti transplantacije bubrega. Neki radovi koriste *post-hoc* interpretaciju kako bi pomogli liječnicima u razumijevanju pokretača predviđanja (Salaün i sur., 2024, 5; Dili i sur., 2024, 2), no te se tehnike općenito ne proširuju na sustave UI-e za alokaciju organa. Većina rasprava o pravednosti i objašnjivosti tako ostaje izolirana i vezana uz validaciju sustava UI-e umjesto da bude povezana s revizijom na razini alokacije organa što ističe značajan jaz u implementaciji sustava OUI-e (Firuzpour i sur., 2025, 9).

Navedeno potpoglavlje prikazuje kako OUI u zdravstvu nadilazi tehničku dimenziju te predstavlja ključni alat za izgradnju povjerenja među pacijentima i zdravstvenim osobljem. Također, OUI pomaže u prepoznavanju pristranosti te uspostavljanju pravednosti pri čemu doprinosi odgovornom i etički opravdanom donošenju odluka. OUI može pomoći pri prilagođavanju objašnjenja koja su relevantna za specifične korisnike i kontekst što posebno dolazi do izražaja u primjeni sustava UI-e za transplantaciju bubrega.

7. Zaključak

U ovom doktorskom radu proveli smo etičku i epistemološku analizu upotrebe UI-e koja nam nalaže kako nije važno samo koje ciljeve zadajemo sustavu UI-e da ih ispuni nego je važan i način na koji sustav UI-e dolazi do ostvarenja tih ciljeva. Budućnost razvoja sustava UI-e će stoga biti važno promatrati pod vidom postizanja ravnoteže između objašnjivosti i preciznosti. Razumijevanje odluka i postupaka složenih sustava umjetne inteligencije ključan je izazov, što je dovelo do razvoja OUI-e radi povećanja povjerenja, prihvaćanja i usklađenosti s regulativom.

U prvom poglavlju smo približili pojam UI-e koja poboljšava ljudske kognitivne sposobnosti, ali istovremeno donosi značajne etičke i epistemološke izazove. Ključni izazovi uključuju odgovornost, pristranost, objašnjivost i povjerenje u sustave UI-e. Posebno u zdravstvu, OUI omogućuje bolje razumijevanje odluka sustava i smanjenje rizika za pacijente, dok društvena percepcija i antropomorfizacija utječu na prihvaćanje tehnologije.

U drugom poglavlju smo detaljno analizirali pojam UI-e i definirali UI-u kao sustav sposoban učiti, percipirati i djelovati kako bi postigao zadane ciljeve. UI se razvijala od oponašanja ljudske inteligencije do racionalnih agenata i dijeli se na slabu (usku), jaku (opću) i superinteligenciju, s različitim sposobnostima i autonomijom. Strojno učenje omogućuje otkrivanje pravila iz podataka, dok duboko učenje koristi neuronske mreže za prepoznavanje složenih uzoraka. Međutim, sustavi UI-e često funkcioniraju kao „crna kutija“, gdje je način donošenja odluka netransparentan, što izaziva etičke i pravne izazove, posebno u zdravstvu, ali i u pravosuđu i u upravljanju autonomnim vozilima. OUI nastoji pružiti objašnjenja i vizualizacije koje povećavaju razumljivost, povjerenje i odgovornost, iako još uvijek ne postoji univerzalni standard za procjenu objašnjivosti.

U trećem poglavlju smo predstavili primjenu UI-e u zdravstvu gdje može ubrzati dijagnostiku, omogućuti personalizaciju postupaka i nadmašiti ljude u prepoznavanju uzoraka, ali nosi rizike poput pogrešaka i neobjašnjivih odluka. Pouzdana primjena UI-e zahtijeva regulaciju, etičke smjernice, verifikaciju, validaciju i ljudski nadzor kako bi se osigurala preciznost i sigurnost. Multimodalni i objašnjivi sustavi UI-e podržavaju, ali ne zamjenjuju, kliničku prosudbu liječnika.

U četvrtom poglavlju smo istaknuli kako OUI omogućuje ljudima razumijevanje odluka sustava UI-e, što je ključno za povjerenje, etiku i sigurnost. OUI nadilazi ograničenost sustava UI-e koji rade na principu „crne kutije“ pružajući razumljiva objašnjenja, a primjenjuje se u dijagnostici,

terapiji i procjeni rizika. Unatoč prednostima, izazovi primjene OUI-e uključuju ljudske kognitivne ograničenosti, rizik *post-fact* racionalizacije i ograničeno usvajanje u praksi.

U petom poglavlju detaljno smo analizirali koncept povjerenja u UI-u koje je ključno za uspješnu integraciju tehnologije u zdravstvu jer omogućava suradnju, smanjuje rizik i potiče učinkovitost. Transparentnost i objašnjivost pomažu izgradnji povjerenja, a brojni autori tvrde kako je objašnjivost nužna za izgradnju povjerenja u sustave UI-e. U petom poglavlju smo detaljno analizirali ulogu informiranog pristanka za pacijente u kontekstu sustava UI-e koji rade na principu „crne kutije" i pristranosti algoritama, što zahtijeva redefiniranje pristanka temeljenog na povjerenju u liječnika ili sustav. Povjerenje se može razviti ako sustavi kontinuirano pokazuju kompetentno i pouzdano ponašanje, uz usklađivanje s ljudskim vrijednostima i nadzorom. U petom poglavlju smo analizirali ulogu i mogućnost moralnih agenata. Suvremena društva ovise o epistemološkim autoritetima zbog hiperspecijalizacije znanja. Pitanje može li stroj misliti otvara raspravu o epistemološkom statusu UI-e. Sustavi UI-e ne posjeduju svijest, slobodnu volju ni moralnu odgovornost, već djeluju kroz algoritamska predviđanja i statističke modele. Koncept umjetnih moralnih agenata istražili smo kroz hijerarhiju Moorovih agenata, ali potpuni moralni agenti zahtijevaju ljudske osobine koje UI ne posjeduje. Rasprava uključuje probleme atribucije odgovornosti i rascjepa odgovornosti kada strojevi djeluju autonomno. Iako UI može pomagati u moralnim odlukama, odgovornost za konačne odluke i dalje pripada ljudima. Suradnja ljudi i strojeva nudi najbolji način za balansiranje epistemoloških i moralnih izazova.

Također, u petom poglavlju smo analizirali ulogu usklađivanja vrijednosti jer etika UI-e u zdravstvu zahtijeva usklađivanje sustava s ljudskim vrijednostima i moralnim načelima pacijenata i liječnika. Normativni i tehnički izazovi uključuju izbor etičkih pristupa (utilitaristička normativna teorija, deonotološka normativna teorija, areatička normativna teorija) i strategija učenja (nadzirano učenje, nenadzirano učenje, učenje s potkrepljenjem, inverzno učenje). Ključno je stvoriti pouzdane sustave koji poštuju moralne principe i podržavaju odluke u skladu s ljudskim vrijednostima, balansirajući objašnjivost i preciznost, te kritički primjenjuju objašnjenja kako bi se spriječile pogreške, manipulacije ili smanjeno prihvaćanje UI-e.

U posljednjem šestom poglavlju smo analizirali kako kroz primjenu UI-e u transplantaciji bubrega dolazi do razvoja etičkih i epistemoloških izazova, uključujući pravednost, transparentnost i objašnjivost odluka. Sustavi OUI mogu smanjiti pristranost, povećati pravednost, poboljšati podudaranje donora i primatelja te povećati povjerenje pacijenata, ali

njihova implementacija mora biti vođena etikom vrline, poštovanjem autonomije pacijenta i sociodemografskim vrijednostima. Posebna pažnja treba se posvetiti uključivim kriterijima kako bi se izbjegle nepravedne liste čekanja, a u budućnosti sustavi UI-e imat će sve veću ulogu u podršci donošenju odluka, ugrađujući vrijednosti koje osiguravaju pravedan pristup transplantaciji.

Zaključno, cilj je povećati povjerenje i odgovornost u visokorizičnim situacijama, kao što je zdravstvo, te omogućiti ljudima da razumiju predviđanja sustava UI-e. Iako postoji velik napredak, još uvijek ne postoji jedinstveni standard za objašnjivost i evaluaciju rezultata. Sustavi UI-e koji rade na principu „crne kutije“ izazivaju probleme u informiranom pristanku jer pacijentima i liječnicima nije jasno kako sustav donosi dijagnoze i preporuke. Pristranosti sustava (vođene podacima, algoritamske, ljudske) dodatno ugrožavaju pravednost i pouzdanost odluka. Nedostatak objašnjivosti otežava da pacijent stvarno razumije rizike i koristi liječenja, što može oslabiti pravnu i etičku težinu pristanka. Neki autori predlažu redefiniciju informiranog pristanka, temeljenog na povjerenju u liječnika ili sustav, umjesto potpune informiranosti. U praksi se preporučuje „kopilot upotreba“ UI-a, gdje sustav pomaže liječniku, ali konačna odluka ostaje ljudska.

Povjerenje je ključno za prihvaćanje UI-e, ali se može narušiti zbog grešaka ili krhkih algoritamskih predviđanja, što ilustriraju obrađeni primjeri lažnih nuklearnih dojava i nesreća autonomnih vozila. U skladu s tim, sustavi UI-e trebaju biti usklađeni s ljudskim vrijednostima i korišteni uz odgovarajući nadzor.

Utvrdili smo kako sociodemografski kriteriji mogu dovesti do stvaranja nepravednih lista čekanja zbog čega je potrebno posebnu pažnju posvetiti uključivim kriterijima za transplantaciju bubrega. U budućnosti će sustavi UI-e imati sve veću ulogu u podršci prilikom donošenja odluka i ugrađivanje vrijednosti u sustave UI-e bit će još važnije.

Iako primjena UI-e u transplantaciji bubrega donosi značajne etičke i epistemološke izazove, posebno u pogledu pravednosti, transparentnosti i objašnjivosti odluka, oni mogu biti nadvladani sustavima OUI-e koji pomažu u smanjenju pristranosti, poboljšanju pravednosti, poboljšanju podudaranja donora i primatelja te povećanju povjerenja pacijenata, ali njihova implementacija mora biti vođena etikom vrline, poštovanjem autonomije pacijenta i sociodemografskim vrijednostima kako bi se osigurao pravedan pristup transplantaciji.

Literatura

- Abecassis, M., Bartlett, S. T., Collins, A. J., Davis, C. L., Delmonico, F. L., Friedewald, J. J., Hays, R., Howard, A., Jones, E., Leichtman, A. B., Merion, R. M., Metzger, R. A., Pradel, F., Schweitzer, E. J., Velez, R. L., & Gaston, R. S. (2008). Kidney transplantation as primary therapy for end-stage renal disease: a National Kidney Foundation/Kidney Disease Outcomes Quality Initiative (NKF/KDOQITM) conference. Clinical journal of the American Society of Nephrology : CJASN, 3(2), 471–480. https://doi.org/10.2215/CJN.05021107
- Abid, M. H., Kumah, A., Newera, A., & Hafez, P. (2024). Patient-Centered Healthcare: From Patient Experience to Human Experience. Global journal on quality and safety in healthcare, 7(4), 144–148. https://doi.org/10.36401/JQSH-24-X2
- Abraham, T. H. (2002). (Physio)logical circuits: The intellectual origins of the McCulloch–Pitts neural networks. Journal of the History of the Behavioral Sciences, 38(1), 3–25. https://doi.org/10.1002/jhbs.1094
- Abramowicz, D., Cochat, P., Claas, F. H., Heemann, U., Pascual, J., Dudley, C., Harden, P., Hourmant, M., Maggiore, U., Salvadori, M., Spasovski, G., Squifflet, J. P., Steiger, J., Torres, A., Viklicky, O., Zeier, M., Vanholder, R., Van Biesen, W., & Nagler, E. (2015). European Renal Best Practice Guideline on kidney donor and recipient evaluation and perioperative care. Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association, 30(11), 1790–1797. https://doi.org/10.1093/ndt/gfu216
- Acemoglu, D., & Johnson, S. (2024). Learning from Ricardo and Thompson: Machinery and labor in the early industrial revolution, and in the age of AI (NBER Working Paper No. 32416). National Bureau of Economic Research. https://ssrn.com/abstract=4826001

- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052
- Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). Trust in AI: Progress, challenges, and future directions. Humanities & Social Sciences Communications, 11, Article 1568. https://doi.org/10.1057/s41599-024-04044-8
- Agozzino, E., Borrelli, S., Cancellieri, M., Carfora, F. M., Di Lorenzo, T., & Attena, F. (2019). Does written informed consent adequately inform surgical patients? A cross sectional study. BMC Medical Ethics, 20, Article 1. https://doi.org/10.1186/s12910-018-0340-z
- Ahmed, I., Jeon, G., & Piccialli, F. (2022). From artificial intelligence to explainable artificial intelligence in Industry 4.0: A survey on what, how, and where. IEEE Transactions on Industrial Informatics, 18(8), 5031–5042. https://doi.org/10.1109/TII.2022.3146552
- Ai, L., Muggleton, S. H., Hocquette, C., & others. (2021). Beneficial and harmful explanatory machine learning. Machine Learning, 110(4), 695–721. https://doi.org/10.1007/s10994-020-05941-0
- Al Kuwaiti, A., Nazer, K., Al-Reedy, A., Al-Shehri, S., Al-Muhanna, A., Subbarayalu, A. V., Al Muhanna, D., & Al-Muhanna, F. A. (2023). A Review of the Role of Artificial Intelligence in Healthcare. Journal of personalized medicine, 13(6), 951. https://doi.org/10.3390/jpm13060951
- Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion, 99, 101805. https://doi.org/10.1016/j.inffus.2023.101805
- Al-Jardani, R., Al-Hadrami, W., Al-Balushi, M. H., Al-Adawi, M., & Al-Nasseri, S. (2024). Factors Influencing the Quality and Extent of the Informed Consent Process in Patients Undergoing Surgical Procedures: A Cross-Sectional Study. Cureus, 16(8), e68149. https://doi.org/10.7759/cureus.68149
- Alkhanbouli, R., Matar Abdulla Almadhaani, H., Alhosani, F., Simsekler, M. C. E. (2025). The role of explainable artificial intelligence in disease prediction: A

systematic literature review and future research directions. BMC Medical Informatics and Decision Making, 25, 110. https://doi.org/10.1186/s12911-025-02944-6

- Allen, C., Smit, I., and Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. Ethics and Information Technology, 7(3), 149–155. https://doi.org/10.1007/s10676-006-0004-4
- Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. Journal of Experimental & Theoretical Artificial Intelligence, 12(3), 251–261. https://doi.org/10.1080/09528130050111428
- Alpsancar, S., Buhl, H. M., Matzner, T., & Scharlau, I. (2025). Explanation needs and ethical demands: Unpacking the instrumental value of XAI. AI and Ethics, 5, 3015–3033. https://doi.org/10.1007/s43681-024-00622-3
- Alum, E. U., & Ugwu, O. P. (2025). Artificial intelligence in personalized medicine: transforming diagnosis and treatment. Deleted Journal, 7(3). https://doi.org/10.1007/s42452-025-06625-x
- Alvarado, R. (2023). AI as an epistemic technology. Science and Engineering Ethics, 29, 32. https://doi.org/10.1007/s11948-023-00451-3
- Aljunaid, S. K., Almheiri, S. J., Dawood, H., & Khan, M. A. (2025). Secure and Transparent Banking: Explainable AI-Driven Federated Learning Model for Financial Fraud Detection. Journal of Risk and Financial Management, 18(4), 179. https://doi.org/10.3390/jrfm18040179
- American Medical Association. (2016A). Patient rights. AMA Code of Medical Ethics. https://code-medical-ethics.ama-assn.org/ethics-opinions/patient-rights, pristup: 30. rujna 2025.
- American Medical Association. (2016B). Ethics Opinions. AMA Code of Medical Ethics. https://code-medical-ethics.ama-assn.org/ethics-opinions/patient-rights, pristup: 2. listopada 2025.
- Anderson, M., & Anderson, S. L. (2007). Machine ethics: Creating an ethical intelligent agent. AI Magazine, 28(4), 15–26. https://doi.org/10.1609/aimag.v28i4.2065
- Andrews, R., Diederich, J., & Tickle, A. (1995). Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowledge-Based Systems, 8(6), 373–389. https://doi.org/10.1016/0950-7051(96)81920-4
- Ankuda, C. K., Block, S. D., Cooper, Z., Correll, D. J., Hepner, D. L., Lasic, M., Gawande, A. A., & Bader, A. M. (2014). Measuring critical deficits in shared decision

making before elective surgery. Patient education and counseling, 94(3), 328–333. https://doi.org/10.1016/j.pec.2013.11.013

- Aramini, M. (2009). Uvod u bioetiku. Kršćanska sadašnjost.
- Araujo, T., Helberger, N., Kruikemeier, S., Vreese, C. H. de, & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI & Society, 35(3), 1–13.
- Arendt, H. (1951). The origins of totalitarianism. Harcourt Brace & Company.
- Aristotle. (1999). Nicomachean ethics. Batoche Books.
- Arkin, R. C. (2008). Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. In Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI '08) (121–128). Association for Computing Machinery. https://doi.org/10.1145/1349822.1349839
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
- Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial intelligence and human trust in healthcare: Focus on clinicians. Journal of Medical Internet Research, 22(6), e15154. https://doi.org/10.2196/15154
- Ashby, W. R. (1956). An introduction to cybernetics. New York, J. Wiley, https://doi.org/10.5962/bhl.title.5851
- Asimov, I. (1942). Runaround. Astounding Science Fiction, 29(1), 94–103.
- Assfalg, V., Miller, G., Stocker, F., Hüser, N., Hartmann, D., Heemann, U., Tieken, I., Zanen, W., Vogelaar, S., Rosenkranz, A. R., Schneeberger, S., Függer, R., Berlakovich, G., Ysebaert, D. R., Jacobs-Tulleneers-Thevissen, D., Mikhalski, D., van Laecke, S., Kuypers, D., Mühlfeld, A. S., Viebahn, R., … Wahba, R. (2024). Rescue Allocation Modes in Eurotransplant Kidney Transplantation: Recipient Oriented Extended Allocation Versus Competitive Rescue Allocation-A Retrospective Multicenter Outcome Analysis. Transplantation, 108(5), 1200–1211. https://doi.org/10.1097/TP.0000000000004878
- Assis de Souza, A., Stubbs, A. P., Hesselink, D. A., Baan, C. C., & Boer, K. (2025). Cherry on Top or Real Need? A Review of Explainable Machine Learning in Kidney

Transplantation. Transplantation, 109(1), 123–132. https://doi.org/10.1097/TP.0000000000005063

- Astromskė, K., Peičius, E., & Astromskis, P. (2021). Ethical and legal challenges of informed consent applying artificial intelligence in medical diagnostic consultations. AI & Society, 36(932). https://doi.org/10.1007/s00146-020-01008-9
- Atmar, J. W. (1976). Speculation on the evolution of intelligence and its possible realization in machine form (Sc.D. dissertation). New Mexico State University, Las Cruces.
- Atzmueller, M., Sylvester, S., & Kanawati, R. (2023). Exploratory and explanation-aware network intrusion profiling using subgroup discovery and complex network analysis. In Proceedings of the 2023 European Interdisciplinary Cybersecurity Conference (EICC '23) (153–158). Association for Computing Machinery. https://doi.org/10.1145/3590777.3590803
- Aubert, O., Kamar, N., Vernerey, D., Viglietti, D., Martinez, F., Duong-Van-Huyen, J. P., Eladari, D., Empana, J. P., Rabant, M., Verine, J., Rostaing, L., Congy, N., Guilbeau-Frugier, C., Mourad, G., Garrigue, V., Morelon, E., Giral, M., Kessler, M., Ladrière, M., Delahousse, M., … Loupy, A. (2015). Long term outcomes of transplantation using kidneys from expanded criteria donors: prospective, population based cohort study. BMJ (Clinical research), 351, h3557. https://doi.org/10.1136/bmj.h3557
- Avanzo, M., Trianni, A., Botta, F., Talamonti, C., Stasi, M., & Iori, M. (2021). Artificial Intelligence and the Medical Physicist: Welcome to the Machine. Applied Sciences, 11(4), 1691. https://doi.org/10.3390/app11041691
- Averbach, S., Ha, D., Meadows, A., Brubaker, L., & Gyamfi-Bannerman, C. (2023). Failure to progress: structural racism in women's healthcare. EClinicalMedicine, 57, 101861. https://doi.org/10.1016/j.eclinm.2023.101861
- Avramides, A. (1996). Descartes and Other Minds. Teorema: Revista Internacional de Filosofía, 16(1), 27–46. http://www.jstor.org/stable/43046178
- Babic, B., Gerke, S., Evgeniou, T., & Cohen, I. G. (2021). Beware explanations from AI in health care. Science, 373(6552), 284–286. https://doi.org/10.1126/science.abg1834

- Babushkina, D. (2023). Are we justified attributing a mistake in diagnosis to an AI diagnostic system? AI And Ethics, 3(2), 567–584. https://doi.org/10.1007/s43681-022-00189-x
- Bahalul Haque, A. K. M., Najmul Islam, A. K. M., & Mikalef, P. (2023). Explainable artificial intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. Technological Forecasting and Social Change, 186(Part A), 122120. https://doi.org/10.1016/j.techfore.2023.122120
- Baier, A. (1986). Trust and antitrust. Ethics, 96(2), 231–260. https://www.jstor.org/stable/2381376
- Baker, S., Xiang, W., (2023). Explainable AI is Responsible AI: How Explainability Creates Trustworthy and Socially Responsible Artificial Intelligence. 10.48550/arXiv.2312.01555.
- Bal, B. S., & Choma, T. J. (2012). What to disclose? Revisiting informed consent. Clinical orthopaedics and related research, 470(5), 1346–1356. https://doi.org/10.1007/s11999-011-2232-0
- Balagurunathan, Y., Mitchell, R., & El Naqa, I. (2021). Requirements and reliability of AI in the medical context. Physica medica : PM : an international journal devoted to the applications of physics to medicine and biology : official journal of the Italian Association of Biomedical Physics (AIFB), 83, 72–78. https://doi.org/10.1016/j.ejmp.2021.02.024
- Band, S. S., Yarahmadi, A., Hsu, C.-C., Biyari, M., Sookhak, M., Ameri, R., Dehzangi, I., Chronopoulos, A. T., Liang, H.-W. (2023). Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. Informatics in Medicine Unlocked, 40, 101286. https://doi.org/10.1016/j.imu.2023.101286
- Barocas, S., Selbst, A. D., & Raghavan, M. (2020). The hidden assumptions behind counterfactual explanations and principal reasons. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT '20) (80–89). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372830
- Baron, S. (2025). Trust, explainability and AI. Philosophy & Technology, 38(4). https://doi.org/10.1007/s13347-024-00837-6
- Bartneck, C., Lütge, C., Wagner, A. R., & Welsh, S. (2021). An Introduction to Ethics in Robotics and AI (1. izd.). Springer. https://doi.org/10.1007/978-3-030-51110-4

- Basuli, D., & Roy, S. (2023). Beyond Human Limits: Harnessing Artificial Intelligence to Optimize Immunosuppression in Kidney Transplantation. Journal of clinical medicine research, 15(8-9), 391–398. https://doi.org/10.14740/jocmr5012
- Baum, K., Mantel, S., Schmidt, E., & Speith, T. (2022). From responsibility to reason-giving explainable artificial intelligence. Philosophy & Technology, 35(12). https://doi.org/10.1007/s13347-022-00510-w
- Bayor, A. A., Li, J., Yang, I. A., & Varnfield, M. (2025). Designing Clinical Decision Support Systems (CDSS)-A User-Centered Lens of the Design Characteristics, Challenges, and Implications: Systematic Review. Journal of medical Internet research, 27, e63733. https://doi.org/10.2196/63733
- Beauchamp, T. L., & Childress, J. F. (2001). Principles of bioethics (5. izd.). Oxford University Press.
- Behdadi, D., & Munthe, C. (2020). A normative approach to artificial moral agency. Minds and Machines, 30(2), 195–218. https://doi.org/10.1007/s11023-020-09525-8
- Beheshti, A., & Kerridge, I. (2025). Understanding the artificial intelligence revolution and its ethical implications. Bioethical Inquiry. https://doi.org/10.1007/s11673-025-10427-6
- Beisbart, C. (2025). Digging deeper with deep learning? Explanatory understanding and deep neural networks. European Journal for Philosophy of Science, 15(1), 40. https://doi.org/10.1007/s13194-025-00668-y
- Belle, V. (2017). Logic meets probability: Towards explainable AI systems for uncertain worlds. In Proceedings of the 26th International Joint Conference on Artificial Intelligence (5116–5120).
- Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. Frontiers in Big Data, 4. https://doi.org/10.3389/fdata.2021.688969
- Bellucci, G., Molter, F., & Park, S. Q. (2019). Neural representations of honesty predict future trust behavior. Nature Communications, 10, 5184. https://doi.org/10.1038/s41467-019-13261-8
- Bellucci, M., Delestre, N., Malandain, N., Zanni-Merk, C. (2021). Towards a terminology for a fully contextualized XAI. Procedia Computer Science, 192, 241–250. https://doi.org/10.1016/j.procs.2021.08.025

- Beltramin, D., Lamas, E., & Bousquet, C. (2022). Ethical Issues in the Utilization of Black Boxes for Artificial Intelligence in Medicine. Studies in health technology and informatics, 295, 249–252. https://doi.org/10.3233/SHTI220709
- Benani, A., & Messas, E. (2025). Prix Nobel de physique 2024 : John J. Hopfield et Geoffrey E. Hinton - De Hopfield et Hinton à AlphaFold : le Prix Nobel 2024 récompense les pionniers de l'apprentissage profond [Nobel Prize in physics 2024 : John J. Hopfield and Geoffrey E. Hinton. From Hopfield and Hinton to AlphaFold: The 2024 Nobel Prize honors the pioneers of deep learning]. Medecine sciences : M/S, 41(3), 277–280. https://doi.org/10.1051/medsci/2025036
- Bench-Capon, T., & Atkinson, K. (2009). Abstract argumentation and values. In G. Simari & I. Rahwan (Ur.), Argumentation in artificial intelligence (3–22). Springer. https://doi.org/10.1007/978-0-387-98197-0_3
- Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., Dragan, A., Torr, P., Russell, S., Kahneman, D., Brauner, J., & Mindermann, S. (2024). Managing extreme AI risks amid rapid progress. Science, 384(6698), 842–845. https://doi.org/10.1126/science.adn0117
- Benitez, J. M., Castro, J. L., & Requena, I. (1997). Are artificial neural networks black boxes?. IEEE transactions on neural networks, 8(5), 1156–1164. https://doi.org/10.1109/72.623216
- Bergomi, L., Nicora, G., Orlowska, M. A., Podrecca, C., Bellazzi, R., Fregosi, C., Salinaro, F., Bonzano, M., Crescenzi, G., Speciale, F., Di Pietro, S., Zuccaro, V., Asperges, E., Sacchi, P., Valsecchi, P., Pagani, E., Catalano, M., Bortolotto, C., Preda, L., & Parimbelli, E. (2025). Which explanations do clinicians prefer? A comparative evaluation of XAI understandability and actionability in predicting the need for hospitalization. BMC medical informatics and decision making, 25(1), 269. https://doi.org/10.1186/s12911-025-03045-0
- Bertrand, A., Belloum, R., Eagan, J. R., Maxwell, W. (2022). How cognitive biases affect XAI-assisted decision-making: A systematic review. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (78–91). https://doi.org/10.1145/3514094.3534164

- Bhatnagar, S., Alexandrova, A., Avin, S., Cave, S., Cheke, L., Crosby, M., Feyereisl, J., Halina, M., Loe, B. S., Ó hÉigeartaigh, S., Martínez-Plumed, F., Price, H., Shevlin, H., Weller, A., Winfield, A., & Hernández-Orallo, J. (2018). Mapping intelligence: Requirements and possibilities. Müller, V. (Ur.) Philosophy and Theory of Artificial Intelligence 2017. PT-AI 2017. Studies in Applied Philosophy, Epistemology and Rational Ethics, Vol 44. Springer, Cham. https://doi.org/10.1007/978-3-319-96448-5_13
- Bickmore, T. W., Pfeifer, L. M., & Paasche-Orlow, M. K. (2009). Using computer agents to explain medical documents to patients with low health literacy. Patient education and counseling, 75(3), 315–320. https://doi.org/10.1016/j.pec.2009.02.007
- Bjerring, J. C., & Busch, J. (2021). Artificial intelligence and patient-centered decision-making. Philosophy & Technology, 34(2), 349–371. https://doi.org/10.1007/s13347-019-00391-6
- Boddington, P. (2023). AI ethics: A textbook. Springer Nature Singapore. https://doi.org/10.1007/978-981-19-9382-4
- Boddington, P., (2017). Towards a Code of Ethics for Artificial Intelligence. New York: Springer.
- Boden, M. A. (1977). Artificial Intelligence and Natural Man. Basic Books, New York.
- Boerstra, B. A., Pippias, M., Kramer, A., Dirix, M., Daams, J., Jager, K. J., Hellemans, R., & Stel, V. S. (2025). The evaluation of kidney transplant candidates prior to waitlisting: A scoping review. Clinical Kidney Journal, 18(1), sfae377. https://doi.org/10.1093/ckj/sfae377
- Bok, S. (1999). Lying: Moral choice in public and private life. Vintage.
- Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. Science, 352(6293), 1573–1576. https://doi.org/10.1126/science.aaf2654
- Borole, P., & Rajan, A. (2024). Building trust in deep learning-based immune response predictors with interpretable explanations. Communications Biology, 7, 279. https://doi.org/10.1038/s42003-024-05968-2
- Borrego-Díaz, J., Galán-Páez, J. (2022). Explainable artificial intelligence in data science: from foundational issues towards socio-technical considerations. Minds Mach. 32, 485–531. doi: 10.1007/s11023-022-09603-z

- Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.
- Bourguignon, L., & Mossé, M. (2025). How to count sore throats. Analysis. Advance online publication. https://doi.org/10.1093/analys/anae068
- Bracic, A., Callier, S. L., & Price, W. N., 2nd (2022). Exclusion cycles: Reinforcing disparities in medicine. Science (New York, N.Y.), 377(6611), 1158–1160. https://doi.org/10.1126/science.abo2788
- Bradley, D., (2025), Vegetative electron microscopy https://www.science-base.com/science-blog/vegetative-electron-microscopy.html, pristup: 1. travnja 2025.
- Brewer, D., & Carlson, A. (2024). Addressing sample selection bias for machine learning methods. Journal of Applied Econometrics, 39(3), 383–400. https://doi.org/10.1002/jae.3029
- Brezis, M., Israel, S., Weinstein-Birenshtock, A., Pogoda, P., Sharon, A., & Tauber, R. (2008). Quality of informed consent for invasive procedures. International journal for quality in health care : journal of the International Society for Quality in Health Care, 20(5), 352–357. https://doi.org/10.1093/intqhc/mzn025
- Brown, M. J. (2024). For values in science: Assessing recent arguments for the ideal of value-free science. Synthese, 204, 112. https://doi.org/10.1007/s11229-024-04762-1
- Brożek, B., Furman, M., Jakubiec, M., & Kucharzyk, B. (2024). The black box problem revisited: Real and imaginary challenges for automated legal decision making. Artificial Intelligence and Law, 32(3), 427–440. https://doi.org/10.1007/s10506-023-09356-9
- Bruckert, S., Finzel, B., Schmid, U. (2020). The next generation of medical decision support: A roadmap toward transparent expert companions. Frontiers in Artificial Intelligence, 3, 75. https://doi.org/10.3389/frai.2020.00075
- Brundage, M. (2014). Limitations and risks of machine ethics. Journal of Experimental & Theoretical Artificial Intelligence, 26(3), 355–372. https://doi.org/10.1080/0952813X.2014.895108
- Brunton, S. L., & Kutz, J. N. (2022). Data-driven science and engineering: Machine learning, dynamical systems, and control (2. izd.). Cambridge University Press.

- Bryson, J. J. (2020). The artificial intelligence of the ethics of artificial intelligence: An introductory overview for law and regulation. In M. D. Dubber, F. Pasquale, & S. Das (Ur.), The Oxford handbook of ethics of AI. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.1
- Buchanan, B. G., Shortliffe, E. H., (1985). Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. Reading, MA: Addison-Wesley.
- Buchweitz, O., Wülfing, P., & Malik, E. (2005). Interobserver variability in the diagnosis of minimal and mild endometriosis. European journal of obstetrics, gynecology, and reproductive biology, 122(2), 213–217. https://doi.org/10.1016/j.ejogrb.2005.02.002
- Bunge, M. (1963). A General Black Box Theory. Philosophy of Science, 30(4), 346–358. doi:10.1086/287954
- Bunnik E. M. (2023). Ethics of allocation of donor organs. Current opinion in organ transplantation, 28(3), 192–196. https://doi.org/10.1097/MOT.0000000000001058
- Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research, 70, 245–317. https://doi.org/10.1613/jair.1.12228
- Burns, H., Luckhardt, C. A., Parlett, J. W., & Redfield, C. L. (2014). Intelligent tutoring systems: Evolutions in design. Routledge. https://doi.org/10.4324/9781315783718
- Bushnell, S. T. (1922). The Truth about Henry Ford, Chicago: Reilly & Lee Co
- Buturovic Ponikvar, J., Orozco Guillen, A., & Piccoli, G. B. (2025). Epochal changes and ethical issues, reflections on kidney transplantation. Journal of Nephrology, 38, 1135–1137. https://doi.org/10.1007/s40620-025-02350-1
- Caldwell, J. S., Cheng, X. S., Chertow, G. M., & Goldhaber-Fiebert, J. D. (2025). Kidney Transplant Wait Times Under Waiting List Expansion Scenarios. JAMA network open, 8(3), e251665. https://doi.org/10.1001/jamanetworkopen.2025.1665
- Callaway, E. (2024). Chemistry Nobel goes to developers of AlphaFold AI that predicts protein structures. Nature, 634(8034), 525–526. https://doi.org/10.1038/d41586-024-03214-7

- Cao, N., Sun, B., Li, W., Jin, X., Zhang, J., & Yue, G. (2025). The implicit theories of trust: The more individuals believe trust to be unchangeable, the more they tend to trust others. BMC Psychology, 13, 950. https://doi.org/10.1186/s40359-025-03267-x
- Cappelli, O., Aliberti, M., & Praino, R. (2024). The 'Implicit Intelligence' of artificial intelligence. Investigating the potential of large language models in social science research. Political Research Exchange, 6(1). https://doi.org/10.1080/2474736X.2024.2351794
- Casey, S., Avalos, G., & Dowling, M. (2018). Critical care nurses' knowledge of a-larm fatigue and practices towards alarms: A multicentre study. Intensive & critical care nursing, 48, 36–41. https://doi.org/10.1016/j.iccn.2018.05.004
- Castelvecchi D. (2016). Can we open the black box of AI?. Nature, 538(7623), 20–23. https://doi.org/10.1038/538020a
- Celi, L. A. (2025). Teaching machines to doubt. Nature Medicine. https://doi.org/10.1038/s41591-025-04013-x
- Celi, L. A., Cellini, J., Charpignon, M. L., Dee, E. C., Dernoncourt, F., Eber, R., Mitchell, W. G., Moukheiber, L., Schirmer, J., Situ, J., Paguio, J., Park, J., Wawira, J. G., Yao, S., & for MIT Critical Data (2022). Sources of bias in artificial intelligence that perpetuate healthcare disparities-A global review. PLOS digital health, 1(3), e0000022. https://doi.org/10.1371/journal.pdig.0000022
- Chadban, S. J., Ahn, C., Axelrod, D. A., Foster, B. J., Kasiske, B. L., Kher, V., Kumar, D., Oberbauer, R., Pascual, J., Pilmore, H. L., Rodrigue, J. R., Segev, D. L., Sheerin, N. S., Tinckam, K. J., Wong, G., & Knoll, G. A. (2020). KDIGO Clinical Practice Guideline on the Evaluation and Management of Candidates for Kidney Transplanta-tion. Transplantation, 104(4S1 Suppl 1), S11–S103. https://doi.org/10.1097/TP.0000000000003136
- Chalmers, D. J., (2022). Reality+: Virtual worlds and the problems of philosophy. Penguin Books.
- Chan, B., (2023). Black-box assisted medical decisions: AI power vs. ethical physi-cian care. Medicine, health care, and philosophy, 26(3), 285–292. https://doi.org/10.1007/s11019-023-10153-z
- Chandrasekaran, B., Tanner, M. C., & Josephson, J. R. (1989). Explaining control strategies in problem solving. IEEE Expert, 4(1), 9–15. https://doi.org/10.1109/64.21896

- Chang, V., Sivakulasingam, S., Wang, H., Wong, S. T., Ganatra, M. A., & Luo, J. (2024). Credit Risk Prediction Using Machine Learning and Deep Learning: A Study on Credit Card Customers. Risks, 12(11), 174. https://doi.org/10.3390/risks12110174
- Chang, V., Xu, Q. A., Akinloye, S. H., Benson, V., & Hall, K. (2024). Prediction of bank credit worthiness through credit risk analysis: An explainable machine learning study. Annals of Operations Research, 1–25. https://doi.org/10.1007/s10479-024-06134-x
- Chapman, C. S. (1997). Reflections on a contingent view of accounting. Accounting, Organizations and Society, 22(2), 189–205. https://doi.org/10.1016/S0361-3682(97)00001-9
- Charniak, E., McDermott, D. V., (1985). Introduction to Artificial Intelligence. Reading, MA: Addison-Wesley.
- Charon R. (2001). The patient-physician relationship. Narrative medicine: a model for empathy, reflection, profession, and trust. JAMA, 286(15), 1897–1902. https://doi.org/10.1001/jama.286.15.1897
- Chauhan, V. K., Clifton, L., Salaün, A., Lu, H. Y., Branson, K., Schwab, P., Nigam, G., & Clifton, D. A. (2025). Sample selection bias in machine learning for healthcare. ACM Transactions on Computing for Healthcare, Advance online publication. https://doi.org/10.1145/3761822
- Cheah, B. C. J., Vicente, C. R., & Chan, K. R. (2025). Machine Learning and Artificial Intelligence for Infectious Disease Surveillance, Diagnosis, and Prognosis. Viruses, 17(7), 882. https://doi.org/10.3390/v17070882
- Chelouche, T., Brahmer, G., & Benedict, S. (2013). Casebook on bioethics and the Holocaust. Haifa: UNESCO Chair in Bioethics.
- Chen, R. J., Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., & Mahmood, F. (2022). Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer cell, 40(8), 865–878.e6. https://doi.org/10.1016/j.ccell.2022.07.004
- Chollet, F. (2021). Deep learning with Python (2. izd.). Manning Publications.
- Chollet, F., (2017), The implausibility of intelligence explosion, https://medium.com/@francois.chollet/the-impossibility-of-intelligence-explosion-5be4a9eda6ec, pristup: 2. lipnja 2025.

- Chomanski, B., & Lauwaert, L. (2024). Online consent: How much do we need to know? AI & Society, 39, 2879–2889. https://doi.org/10.1007/s00146-023-01790-2
- Chomsky, N. (1995). Language and nature. Mind, 104(413), 1–61. https://doi.org/10.1093/mind/104.413.1
- Christian, B. (2020). *The alignment problem: How can machines learn human values?* Atlantic Books.
- Chustecki M. (2024). Benefits and Risks of AI in Health Care: Narrative Review. Interactive journal of medical research, 13, e53616. https://doi.org/10.2196/53616
- Cieciuch, J., Schwartz, S. H., & Davidov, E. (2015). Values, social psychology of. U J. D. Wright (Ur.), International encyclopedia of the social & behavioral sciences, (2. izd.), (41–46). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.25098-8
- Clark-Cutaia, M. N., Menon, G., Li, Y., Metoyer, G. T., Bowring, M. G., Kim, B., Orandi, B. J., Wall, S. P., Hladek, M. D., Purnell, T. S., Segev, D. L., & McAdams-DeMarco, M. A. (2024). Identifying when racial and ethnic disparities arise along the continuum of transplant care: a national registry study. Lancet regional health. Americas, 38, 100895. https://doi.org/10.1016/j.lana.2024.100895
- Coeckelbergh, M. (2010). Moral appearances: Emotions, robots, and human morality. Ethics and Information Technology, 12(3), 235–241. https://doi.org/10.1007/s10676-010-9221-y
- Coeckelbergh, M. (2012). Growing moral relations: Critique of moral status ascription. New York, NY: Palgrave Macmillan.
- Coeckelbergh, M. (2020A). AI Ethics. In The MIT Press eBooks. https://doi.org/10.7551/mitpress/12549.001.0001
- Coeckelbergh, M. (2020B). Artificial intelligence, responsibility attribution, and a relational justification of explainability. Science and Engineering Ethics, 26(5), 2051–2068. https://doi.org/10.1007/s11948-019-00146-8
- Coeckelbergh, M. (2025). AI and Epistemic Agency: How AI Influences Belief Revision and Its Normative Implications. Social Epistemology, 1–13. https://doi.org/10.1080/02691728.2025.2466164
- Cohen, I. G. (2020). Informed consent and medical artificial intelligence: What to tell the patient? Georgetown Law Journal, 108, 1425–1469. Harvard Public Law Working Paper No. 20-03. https://doi.org/10.2139/ssrn.3529576

- Cohen, I. G., & Slottje, A. (2024). Artificial intelligence and the law of informed consent. In B. Solaiman & I. G. Cohen (Ur.), Research handbook on health, AI and the law. Edward Elgar Publishing. https://doi.org/10.4337/9781802205657.ch10
- Cohen, J. (2010). The arc of the moral universe and other essays. Harvard University Press.
- Cohen, S. (2015). Forced supererogation. European Journal of Philosophy, 23(4), 1006–1024. https://doi.org/10.1111/ejop.12023
- Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A. L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., Hoffman, M. M., … Logullo, P. (2024). TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ (Clinical research), 385, e078378. https://doi.org/10.1136/bmj-2023-078378
- Connor, K. L., O'Sullivan, E. D., Marson, L. P., Wigmore, S. J., & Harrison, E. M. (2021). The Future Role of Machine Learning in Clinical Transplantation. Transplantation, 105(4), 723–735. https://doi.org/10.1097/TP.0000000000003424
- Conradie, N. H., & Nagel, S. K. (2024). No agent in the machine: Being trustworthy and responsible about AI. Philosophy & Technology, 37(72). https://doi.org/10.1007/s13347-024-00760-w
- Constantinescu, M., Voinea, C., Uszkai, R., & Vică, C. (2021). Understanding responsibility in Responsible AI: Dianoetic virtues and the hard problem of context. Ethics and Information Technology, 23, 803–814. https://doi.org/10.1007/s10676-021-09616-9
- Conwill, L., Levis Scheirer, M., & Scheirer, W. J. (2024). Virtue in virtual spaces: Catholic social teaching and technology. Liturgical Press.
- Council of Europe (2010). Guide for Research Ethics, Committee Members Steering Committee on Bioethics, (CDBI-CO-GT2)
- Coupland, H., Scheidwasser, N., Katsiferis, A., Davies, M., Flaxman, S., Rod, N. H., Mishra, S., Bhatt, S., Unwin, H. J. T. (2025). Exploring the potential and limitations of deep learning and explainable AI for longitudinal life course analysis. BMC Public Health, 25, 1520. https://doi.org/10.1186/s12889-025-22705-4

- Cramer, H., Evers, V., Ramlal, S., van Someren, M., Rutledge, L., Stash, N., Aroyo, L., & Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. User Modeling and User-Adapted Interaction, 18(5), 455–496. https://doi.org/10.1007/s11257-008-9051-3
- Craven, M. W., & Shavlik, J. W. (1992). Visualizing learning and computation in artificial neural networks. International Journal of Artificial Intelligence Tools, 1(3), 399–425.
- Creel, K. A. (2020). Transparency in complex computational systems. Philosophy of Science, 87(4), 568–589. https://doi.org/10.1086/709729
- Cross, J. L., Choma, M. A., & Onofrey, J. A. (2024). Bias in medical AI: Implications for clinical decision-making. PLOS digital health, 3(11), e0000651. https://doi.org/10.1371/journal.pdig.0000651
- Cultrera, L., Becattini, F., Seidenari, L., Pala, P., & Del Bimbo, A. (2023). Explaining autonomous driving with visual attention and end-to-end trainable region proposals. Journal of Ambient Intelligence and Humanized Computing, 14(2), 1–15. https://doi.org/10.1007/s12652-023-04550-8
- Cummings, M. (2021). Rethinking the maturity of artificial intelligence in safety-critical settings. AI Magazine, 42(1), 6–15. https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/7394
- Čartolovni, A., Tomičić, A., & Lazić Mosler, E. (2022). Ethical, legal, and social considerations of AI-based medical decision-support tools: A scoping review. International journal of medical informatics, 161, 104738. https://doi.org/10.1016/j.ijmedinf.2022.104738
- Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: A systematic literature review. Artificial Intelligence Review, 57(216). https://doi.org/10.1007/s10462-024-10854-8
- Dahlin, E. (2025). Trust in AI. AI & Society. https://doi.org/10.1007/s00146-025-02429-0
- Dalton-Brown S. (2020). The Ethics of Medical AI and the Physician-Patient Relationship. Cambridge quarterly of healthcare ethics : CQ : the international journal of healthcare ethics committees, 29(1), 115–121. https://doi.org/10.1017/S0963180119000847

- Dang, Q., & Li, G. (2025). Unveiling trust in AI: The interplay of antecedents, consequences, and cultural dynamics. AI & Society. https://doi.org/10.1007/s00146-025-02477-6
- Darrach, B., (1970). Meet Shakey, the first electronic person, *Life*
- Daugherty, P. R., Wilson, H. J. (2018). Human + machine: Reimagining work in the age of AI. Harvard Business Review Press.
- David, M. (2012). Lehrer on trustworthiness and acceptance. Philosophical Studies, 161(1), 7–15. https://doi.org/10.1007/s11098-012-9945-0
- Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. Science (New York, N.Y.), 243(4899), 1668–1674. https://doi.org/10.1126/science.2648573
- de Fine Licht, K., & Brülde, B. (2021). On defining "reliance" and "trust": Purposes, conditions of adequacy, and new definitions. Philosophia, 49(5), 1981–2001. https://doi.org/10.1007/s11406-021-00339-1
- De Giorgis, S., Gangemi, A., & Damiano, R. (2022). Basic human values and moral foundations theory in ValueNet ontology. In O. Corcho, L. Hollink, O. Kutz, N. Troquard, & F. J. Ekaputra (Ur.), Knowledge engineering and knowledge management, Vol. 13514. Springer. https://doi.org/10.1007/978-3-031-17105-5_1
- de Quervain, D. J., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishment. Science (New York, N.Y.), 305(5688), 1254–1258. https://doi.org/10.1126/science.1100735
- Degnan, A. J., Ghobadi, E. H., Hardy, P., Krupinski, E., Scali, E. P., Stratchko, L., Ulano, A., Walker, E., Wasnik, A. P., & Auffermann, W. F. (2019). Perceptual and Interpretive Error in Diagnostic Radiology-Causes and Potential Solutions. Academic radiology, 26(6), 833–845. https://doi.org/10.1016/j.acra.2018.11.006
- DeGrave, A. J., Cai, Z. R., Janizek, J. D., Daneshjou, R., & Lee, S.-I. (2025). Auditing the inference processes of medical-image classifiers by leveraging generative AI and the expertise of physicians. Nature Biomedical Engineering, 9, 294–306. https://doi.org/10.1038/s41551-023-01160-9
- DeGrave, A. J., Janizek, J. D., & Lee, S. I. (2021). AI for radiographic COVID-19 detection selects shortcuts over signal. Nature Machine Intelligence, 3(7), 610–619. https://doi.org/10.1038/s42256-021-00338-7

- Dennett, D. C. (2004). Can machines think? U C. Teuscher (Ur.), Alan Turing: Life and legacy of a great thinker (295–316). Springer. https://doi.org/10.1007/978-3-662-05642-4_12
- Densen P. (2011). Challenges and opportunities facing medical education. Transactions of the American Clinical and Climatological Association, 122, 48–58.
- Dessimoz, C., & Thomas, P. D. (2024). AI and the democratization of knowledge. Scientific Data, 11(1), Article 268. https://doi.org/10.1038/s41597-024-03099-1
- Díaz-Pernas, F. J., Martínez-Zarzuela, M., Antón-Rodríguez, M., & González-Ortega, D. (2021). A Deep Learning Approach for Brain Tumor Classification and Segmentation Using a Multiscale Convolutional Neural Network. Healthcare, 9(2), 153. https://doi.org/10.3390/healthcare9020153
- Dietz, T., Ostrom, E., & Stern, P. C. (2003). The struggle to govern the commons. Science, 302, 1907–1912. https://doi.org/10.1126/science.1091015
- Dignum, V. (2017). Responsible artificial intelligence: Designing AI for human values. ITU Journal: ICT Discoveries, 1(1), 1–8.
- Dignum, V. (2019). Responsible artificial intelligence: How to develop and use AI in a responsible way. In Artificial intelligence: Foundations, theory, and algorithms (1-21). Springer International Publishing. https://doi.org/10.1007/978-3-030-30371-6
- Dignum, V. (2020). Responsibility and artificial intelligence. In M. D. Dubber, F. Pasquale, & S. Das (Ur.), The Oxford handbook of ethics of AI. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.12
- Dignum, V. (2022). Responsible Artificial Intelligence – from principles to practice. arXiv. https://arxiv.org/abs/2205.10785
- Dili, G., Balan, A., Davis, T., Balan, A. C. M., & Yazeen, M. (2024). Explainable AI revolutionizing kidney transplants: Seeing clearly to save lives. In 2024 11th International Conference on Advances in Computing and Communications (ICACC) (1–6). IEEE. https://doi.org/10.1109/ICACC63692.2024.10845738
- Dimoka, A. (2010). What Does the Brain Tell Us About Trust and Distrust? Evidence from a Functional Neuroimaging Study. MIS Q., 34, 373-396.
- Ding, Y., Sohn, J. H., Kawczynski, M. G., Trivedi, H., Harnish, R., Jenkins, N. W., Lituiev, D., Copeland, T. P., Aboian, M. S., Mari Aparici, C., Behr, S. C., Flavell, R. R., Huang, S. Y., Zalocusky, K. A., Nardo, L., Seo, Y., Hawkins, R. A., Hernandez Pampaloni, M., Hadley, D., & Franc, B. L. (2019). A Deep Learning Model to Predict

a Diagnosis of Alzheimer Disease by Using 18F-FDG PET of the Brain. Radiology, 290(2), 456–464. https://doi.org/10.1148/radiol.2018180958

- Director, S. (2025). Does black box AI in medicine compromise informed consent? Philosophy & Technology, 38(62). https://doi.org/10.1007/s13347-025-00860-1
- Dirix, M., Philipse, E., Vleut, R., Hartman, V., Bracke, B., Chapelle, T., Roeyen, G., Ysebaert, D., Van Beeumen, G., Snelders, E., Massart, A., Leyssens, K., Couttenye, M. M., Abramowicz, D., & Hellemans, R. (2022). Timing of the pre-transplant work-up for renal transplantation: is there room for improvement?. Clinical kidney journal, 15(6), 1100–1108. https://doi.org/10.1093/ckj/sfac006
- Dixon, D., Sattar, H., Moros, N., Kesireddy, S. R., Ahsan, H., Lakkimsetti, M., Fatima, M., Doshi, D., Sadhu, K., & Junaid Hassan, M. (2024). Unveiling the Influence of AI Predictive Analytics on Patient Outcomes: A Comprehensive Narrative Review. Cureus, 16(5), e59954. https://doi.org/10.7759/cureus.59954
- Donnelly, J. (2007). The relative universality of human rights. Human Rights Quarterly, 29(2), 281–306. https://doi.org/10.1353/hrq.2007.0016
- Doran, D., Schulz, S., Besold, T. R. (2018). What does explainable AI really mean? A new conceptualization of perspectives. CEUR Workshop Proceedings, 2071.
- Dorsch, J., & Deroy, O. (2024). Quasi-metacognitive machines: Why we don't need morally trustworthy AI and communicating reliability is enough. Philosophy & Technology, 37(62). https://doi.org/10.1007/s13347-024-00752-w
- Douglas, D. M. (2025). Researchers' perceptions of automating scientific research. AI & Society, 40, 4131–4144. https://doi.org/10.1007/s00146-025-02190-4
- Doyle, D., Tsymbal, A., & Cunningham, P. (2003). A review of explanation and explanation in case-based reasoning (Technical Report). Trinity College Dublin, Department of Computer Science.
- Dubhashi, D., & Lappin, S. (2017). AI dangers. Communications of the ACM, 60(2), 43–45. https://doi.org/10.1145/2953876
- Dunbar, R. (2021). Friends: Understanding the power of our most important relationships. Little, Brown.
- Dung, L. (2025A). Evaluating approaches for reducing catastrophic risks from AI. AI Ethics, 5, 1177–1188. https://doi.org/10.1007/s43681-024-00475-w
- Dung, L. (2025B). The argument for near-term human disempowerment through AI. AI & Society, 40, 1195–1208. https://doi.org/10.1007/s00146-024-01930-2

- Dung, L. (2023). Current cases of AI misalignment and their implications for future risks. Synthese, 202(1), 138. https://doi.org/10.1007/s11229-023-04367-0
- Durán, J. M., & Formanek, N. (2018). Grounds for trust: Essential epistemic opacity and computational reliabilism. Minds & Machines, 28, 645–666. https://doi.org/10.1007/s11023-018-9481-6
- Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. Journal of Medical Ethics, 47(5), 329–335. https://doi.org/10.1136/medethics-2020-106598
- Durand, M.-A., Moulton, B., Cockle, E., Mann, M., & Elwyn, G. (2015). Can shared decision-making reduce medical malpractice litigation? A systematic review. BMC Health Services Research, 15, 167. https://doi.org/10.1186/s12913-015-0823-2
- Durand-Zaleski, I. S., Alberti, C., Durieux, P., Duval, X., Gottot, S., Ravaud, P., Gainotti, S., Vincent-Genod, C., Moreau, D., & Amiel, P. (2008). Informed consent in clinical research in France: assessment and factors associated with therapeutic misconception. Journal of medical ethics, 34(9), e16. https://doi.org/10.1136/jme.2007.023473
- Duyx, B., Urlings, M. J. E., Swaen, G. M. H., Bouter, L. M., & Zeegers, M. P. (2017). Scientific citations favor positive results: a systematic review and meta-analysis. Journal of clinical epidemiology, 88, 92–101. https://doi.org/10.1016/j.jclinepi.2017.06.002
- Dyer C. (2023). Patients launch legal action against AstraZeneca over its covid-19 vaccine. BMJ (Clinical research), 380, 725. https://doi.org/10.1136/bmj.p725
- Egger, J., Gsaxner, C., Pepe, A., Pomykala, K. L., Jonske, F., Kurz, M., Li, J., Kleesiek, J. (2022). Medical deep learning-A systematic meta-review. Computer methods and programs in biomedicine, 221, 106874. https://doi.org/10.1016/j.cmpb.2022.106874
- Eiband, M., Völkel, S. T., Buschek, D., Cook, S., & Hussmann, H. (2019). When people and algorithms meet. Proceedings of the 24th International Conference on Intelligent User Interfaces, 96–106.
- Eke, C. I., & Shuib, L. (2025). The role of explainability and transparency in fostering trust in AI healthcare systems: A systematic literature review, open issues and potential solutions. Neural Computing and Applications, 37(3), 1999–2034. https://doi.org/10.1007/s00521-024-10868-x

- Emanuel, E. J., & Emanuel, L. L. (1992). Four models of the physician-patient relationship. JAMA, 267(16), 2221–2226.
- Emery-Xu, N., Jordan, R., & Trager, R. (2025). International governance of advancing artificial intelligence. AI & Society, 40, 3019–3044. https://doi.org/10.1007/s00146-024-02050-7
- Epstein, R. M., Fiscella, K., Lesser, C. S., & Stange, K. C. (2010). Why the nation needs a policy push on patient-centered health care. Health affairs, 29(8), 1489–1495.
- Erasmus, A., Brunet, T. D. P., Fisher, E. What is Interpretability?. Philos. Technol. 34, 833–862 (2021). https://doi.org/10.1007/s13347-020-00435-2
- Erion, G., Janizek, J. D., Hudelson, C., Utarnachitt, R. B., McCoy, A. M., Sayre, M. R., White, N. J., & Lee, S. (2022). A cost-aware framework for the development of AI models for healthcare applications. Nature Biomedical Engineering, 6(12), 1384–1398. https://doi.org/10.1038/s41551-022-00872-8
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118. https://doi.org/10.1038/nature21056
- Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. Nature medicine, 25(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z
- Etzioni, A., & Etzioni, O. (2016). AI assisted ethics. Ethics and Information Technology, 18(2), 149–156. https://doi.org/10.1007/s10676-016-9400-6
- Etzioni, O. (2014). AI won't exterminate us; it will empower us. https://medium.com/backchannel/ai-wont-exterminate-us-it-will-empower-us-5b7224735bf3, pristup: 26. travnja 2025.
- European Data Protection Supervisor. (2023). TechDispatch #2/2023 – Explainable Artificial Intelligence. https://www.edps.europa.eu/system/files/2023-11/23-11-16_techdispatch_xai_en.pdf
- Europski parlament i Vijeće Europske unije. (2016). Uredba (EU) 2016/679 Europskog parlamenta i Vijeća od 27. travnja 2016. o zaštiti pojedinaca u vezi s obradom osobnih podataka i o slobodnom kretanju takvih podataka te o stavljanju izvan snage Direktive 95/46/EZ (Opća uredba o zaštiti podataka). Službeni list Europske unije, L 119/1. https://eur-lex.europa.eu/legal-content/HR/TXT/PDF/?uri=CELEX%3A32016R0679

- Europski parlament i Vijeće Europske unije. (2024). Uredba (EU) 2024/1689 Europskog parlamenta i Vijeća o utvrđivanju usklađenih pravila o umjetnoj inteligenciji i izmjeni uredbi (EZ) br. 300/2008, (EU) br. 167/2013, (EU) br. 168/2013, (EU) 2018/858, (EU) 2018/1139 i (EU) 2019/2144 te direktiva 2014/90/EU, (EU) 2016/797 i (EU) 2020/1828 (Akt o umjetnoj inteligenciji). Službeni list Europske unije, L 1689, 12. srpnja 2024. https://eur-lex.europa.eu/legal-content/HR/TXT/?uri=CELEX:32024R1689
- Europski parlament, Vijeće Europske unije i Europska komisija. (2016). Povelja Europske unije o temeljnim pravima (SL C 202/389). Službeni list Europske unije. https://eur-lex.europa.eu/legal-content/HR/TXT/PDF/?uri=CELEX:12016P/TXT&from=RO
- Eurostat, (2024), https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Healthcare_expenditure_statistics_by_function,_provider_and_financing_scheme, pristup: 1. travnja 2025.
- Evans, H., & Snead, D. (2024). Why do errors arise in artificial intelligence diagnostic tools in histopathology and how can we minimize them?. Histopathology, 84(2), 279–287. https://doi.org/10.1111/his.15071
- Eyal N. (2014). Using informed consent to save trust. Journal of medical ethics, 40(7), 437–444. https://doi.org/10.1136/medethics-2012-100490
- Faden, R. R., Beauchamp, T. L., & King, N. M. P. (1986). A history and theory of informed consent. Oxford University Press.
- Faden, R. R., Kass, N. E., Goodman, S. N., Pronovost, P., Tunis, S., & Beauchamp, T. L. (2013). An ethics framework for a learning health care system: A departure from traditional research ethics and clinical ethics. Hastings Center Report, 43(s1), S16–S27. https://doi.org/10.1002/hast.134
- Falagas, M. E., Korbila, I. P., Giannopoulou, K. P., Kondilis, B. K., & Peppas, G. (2009). Informed consent: how much and what do patients understand?. American journal of surgery, 198(3), 420–435. https://doi.org/10.1016/j.amjsurg.2009.02.010
- Fan, B., Schürch, M., Tian, Y., Mallone, A., Frischknecht, L., Koller, M., Van Delden, C., Leichtle, A., Golshayan, D., Villard, J., Schachtner, T., Sidler, D., Schaub, S., Nilsson, J., Krauthammer, M., Swiss Transplant Cohort Study. (2025). Enhancing post kidney transplant prognostication: An interpretable machine learning approach

for longitudinal outcome prediction. npj Digital Medicine, 8, 684. https://doi.org/10.1038/s41746-025-02049-4

- Fang, X., Tao, L., & Li, Z. (2025). Closer to language than steam: AI as the cognitive engine of a new productivity revolution. arXiv. https://arxiv.org/abs/2506.10281
- Fawzy, A., Wu, T. D., Wang, K., Robinson, M. L., Farha, J., Bradke, A., Golden, S. H., Xu, Y., & Garibaldi, B. T. (2022). Racial and Ethnic Discrepancy in Pulse Oximetry and Delayed Identification of Treatment Eligibility Among Patients With COVID-19. JAMA internal medicine, 182(7), 730–738. https://doi.org/10.1001/jamainternmed.2022.1906
- Fehr, E., & Fischbacher, U. (2003). The nature of human altruism. Nature, 425(6960), 785–791. https://doi.org/10.1038/nature02043
- Ferber, D., El Nahhas, O. S. M., Wölflein, G., Wiest, I. C., Clusmann, J., Leßmann, M.-E., Foersch, S., Lammert, J., Tschochohei, M., Jäger, D., Salto-Tellez, M., Schultz, N., Truhn, D., & Kather, J. N. (2025). Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. Nature Cancer, 6(8), 1337–1349. https://doi.org/10.1038/s43018-025-00991-6
- Fernández-Loría, C., Provost, F., & Han, X. (2022). Explaining data-driven decisions made by AI systems: The counterfactual approach. MIS Quarterly, 46(3), 1635-1660.
- Ferrario, A., & Loi, M. (2022). How explainability contributes to trust in AI. 2022 ACM Conference on Fairness, Accountability, and Transparency, 1457–1466. https://doi.org/10.1145/3531146.3533202
- Finocchiaro, G. (2024). The regulation of artificial intelligence. AI & Society, 39, 1961–1968. https://doi.org/10.1007/s00146-023-01650-z
- Firt, E. (2025). What makes full artificial agents morally different. AI & Society, 40, 175–184. https://doi.org/10.1007/s00146-024-01867-6
- Firuzpour, F., Pasha, A. A., Oliaei, F., Nasirimehr, K., Khosravi, M., Rostami, G., & Saeidnia, H. R. (2025). Artificial intelligence-driven kidney organ allocation: systematic review of clinical outcome prediction, ethical frameworks, and decision-making algorithms. BMC nephrology, 26(1), 639. https://doi.org/10.1186/s12882-025-04576-4
- Fletcher, K. H., (1951). Matter with a mind; a neurological research robot. Research; a journal of science and its applications, 4(7), 305–307.

- Floridi, L. (2025A). AI as agency without intelligence: On artificial intelligence as a new form of artificial agency and the multiple realisability of agency thesis. Philosophy & Technology, 38(30). https://doi.org/10.1007/s13347-025-00858-9
- Floridi, L., (2025B). Floridi, L. (2025). Distant writing: Literary production in the age of artificial intelligence. Centre for Digital Ethics (CEDE) Research Paper. https://doi.org/10.2139/ssrn.5232088
- Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. Minds and Machines, 14(3), 349 – 379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d
- Flythe, J. E., & Watnick, S. (2024). Dialysis for Chronic Kidney Failure: A Review. JAMA, 332(18), 1559–1573. https://doi.org/10.1001/jama.2024.16338
- Fok, R., & Weld, D. S. (2024). In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. AI Magazine, 45, 317–332. https://doi.org/10.1002/aaai.12182
- Formosa, P., & Ryan, M. (2021). Making moral machines: Why we need artificial moral agents. AI & Society, 36, 839–851. https://doi.org/10.1007/s00146-020-01089-6
- Fossa, F. (2019). I don't trust you, you faker! On trust, reliance, and artificial agency. Teoria. Rivista di Filosofia, 39(1), 63–80.
- Fox, J., Glasspool, D., Grecu, D., Modgil, S., South, M., & Patkar, V. (2007). Argumentation-based inference and decision making: A medical perspective. IEEE Intelligent Systems, 22(6), 34–41. https://doi.org/10.1109/MIS.2007.102
- Franke, U. (2021). Rawls's original position and algorithmic fairness. Philosophy & Technology, 34(5), 1803–1817. https://doi.org/10.1007/s13347-021-00488-x
- Freitas, M., C., (2011). Kidney transplantation in the US: an analysis of the OPTN/UNOS registry. Clinical transplants, 1–16.
- Fresz, B., Dubovitskaya, E., Brajovic, D., Huber, M. F., & Horz, C. (2025). How should AI decisions be explained? Requirements for explanations from the perspective of European law. In Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (438–450). AAAI Press.
- Friedman, B., Kahn, P. H., Jr., & Borning, A. (2003). Value sensitive design: Theory and methods. University of Washington.

- Fukuyama, F. (1995). Trust: The social virtues and the creation of prosperity. Free Press.
- Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds & Machines, 30, 411–437. https://doi.org/10.1007/s11023-020-09539-2
- Gade, K., Geyik, S. C., Kenthapadi, K., Mithal, V., & Taly, A. (2019). Explainable AI in industry. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19) (3203–3204). Association for Computing Machinery. https://doi.org/10.1145/3292500.3332281
- Gert, J. (2013). Overridingness, moral. U H. LaFollette (Ur.), The international encyclopedia of ethics. John Wiley & Sons. https://doi.org/10.1002/9781444367072.wbiee010
- Ghaly, M. (2012). The ethics of organ transplantation: How comprehensive the ethical framework should be? Medicine, Health Care and Philosophy, 15(2), 175–179. https://doi.org/10.1007/s11019-012-9379-7
- Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. JAMA internal medicine, 178(11), 1544–1547. https://doi.org/10.1001/jamainternmed.2018.3763
- Giannini, A., & Kwik, J. (2023). Negligence failures and negligence fixes: A comparative analysis of criminal regulation of AI and autonomous vehicles. Criminal Law Forum, 34(1), 43–85. https://doi.org/10.1007/s10609-023-09451-1
- Gibert, M. (2023). The case for virtuous robots. AI Ethics, 3, 135–144. https://doi.org/10.1007/s43681-022-00185-1
- Gibney, E., & Castelvecchi, D. (2024). Physics Nobel scooped by machine-learning pioneers. Nature, 634(8034), 523–524. https://doi.org/10.1038/d41586-024-03213-8
- Gigerenzer, G., Mata, J., & Frank, R. (2009). Public knowledge of benefits of breast and prostate cancer screening in Europe. Journal of the National Cancer Institute, 101(17), 1216–1220. https://doi.org/10.1093/jnci/djp237
- Giubilini, A., Gur-Arie, R., & Jamrozik, E. (2025). Expertise, disagreement, and trust in vaccine science and policy. The importance of transparency in a world of experts. Diametros, 22(82), 7–27. https://doi.org/10.33392/diam.1871
- Glaser, J., Nouri, S., Fernandez, A., Sudore, R. L., Schillinger, D., Klein-Fedyshin, M., & Schenker, Y. (2020). Interventions to Improve Patient Comprehension in

Informed Consent for Medical and Surgical Procedures: An Updated Systematic Review. Medical decision making : an international journal of the Society for Medical Decision Making, 40(2), 119–143. https://doi.org/10.1177/0272989X19896348

- Goldman, A., & Beddor, B. (2016). Reliabilist epistemology. In E. N. Zalta (Ur.), The Stanford encyclopedia of philosophy. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/reliabilism
- Gonçalves, B. (2023A). Irony with a point: Alan Turing and his intelligent machine utopia. Philosophy & Technology, 36, 50. https://doi.org/10.1007/s13347-023-00650-7
- Gonçalves, B. (2023B). Can machines think? The controversy that led to the Turing test. AI & Society, 38(4), 2499–2509. https://doi.org/10.1007/s00146-021-01318-6
- Goodman, B., Flaxman, S. (2017). European Union regulations on algorithmic decision making and a "right to explanation". AI Magazine, 38(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741
- Gorban, A. N., Mirkes, E. M., & Tyukin, I. Y. (2020). How deep should be the depth of convolutional neural networks: A backyard dog case study. Cognitive Computation, 12(2), 388–397. https://doi.org/10.1007/s12559-019-09667-7
- Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. Intelligence, 24, 13–23.
- Goyal, K., Garg, M., & Malik, S. (2025). Adoption of artificial intelligence-based credit risk assessment and fraud detection in the banking services: A hybrid approach (SEM-ANN). Future Business Journal, 11, 44. https://doi.org/10.1186/s43093-025-00464-3
- Graff, J. (2024). Moral sensitivity and the limits of artificial moral agents. Ethics and Information Technology, 26, 13. https://doi.org/10.1007/s10676-024-09755-9
- Graham, S., & Brookey, J. (2008). Do patients understand?. The Permanente journal, 12(3), 67–69. https://doi.org/10.7812/TPP/07-144
- Grant, D. G., Behrends, J., & Basl, J. (2025). What we owe to decision-subjects: Beyond transparency and explanation in automated decision-making. Philosophical Studies, 182(1), 55–85. https://doi.org/10.1007/s11098-023-02013-6
- Grant, S. C. (2021). Informed consent—We can and should do better. JAMA Network Open, 4(4), e2110848. https://doi.org/10.1001/jamanetworkopen.2021.10848

- Green, C. D. (2005). Was Babbage's analytical engine intended to be a mechanical model of the mind? History of Psychology, 8(1), 35–45. https://doi.org/10.1037/1093-4510.8.1.35
- Green, D. S., & MacKenzie, C. R. (2007). Nuances of informed consent: the paradigm of regional anesthesia. HSS journal : the musculoskeletal journal of Hospital for Special Surgery, 3(1), 115–118. https://doi.org/10.1007/s11420-006-9035-y
- Gross M. L. (2019). Military medical research in Britain and the USA: the challenge of informed consent. Journal of the Royal Army Medical Corps, 165(4), 298–302. https://doi.org/10.1136/jramc-2018-001023
- Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. Journal of medical ethics, 46(3), 205–211. https://doi.org/10.1136/medethics-2019-105586
- Grote, T., & Keeling, G. (2022). On Algorithmic Fairness in Medical Practice. Cambridge Quarterly of Healthcare Ethics, 31(1), 83–94. doi:10.1017/S0963180121000839
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 93:1–93:42. https://doi.org/10.1145/3236009
- Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA, 316(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216
- Gundersen, T., & Bærøe, K. (2022). The future ethics of artificial intelligence in medicine: Making sense of collaborative models. Science and Engineering Ethics, 28(1), 17. https://doi.org/10.1007/s11948-022-00369-2
- Gunkel, D. J. (2018). 3 S1→S2: Robots can have rights; Robots should have rights. In Robot rights (79–106). The MIT Press.
- Gunkel, D. J. (2023). Person, thing, robot: A moral and legal ontology for the 21st century and beyond. The MIT Press.
- Gunning, D., Vorm, E., Wang, J. Y., & Turek, M. (2021). DARPA's explainable AI (XAI) program: A retrospective. Applied AI Letters, 2, e61. https://doi.org/10.1002/ail2.61

- Hacker, P., Krestel, R., Grundmann, S., Samin, A., & de Vries, K. (2020). Explainable AI under contract and tort law: Legal incentives and technical challenges. Artificial Intelligence and Law, 28(3), 415–439. https://doi.org/10.1007/s10506-020-09260-6
- Hagendorff, T. (2022). A virtue-based framework to support putting AI ethics into practice. Philosophical Technology, 35, 55. https://doi.org/10.1007/s13347-022-00553-z
- Halpern J. (2003). What is clinical empathy?. Journal of general internal medicine, 18(8), 670–674. https://doi.org/10.1046/j.1525-1497.2003.21017.x
- Hammann, D., & Wouters, M. (2025). Explainability versus accuracy of machine learning models: The role of task uncertainty and need for interaction with the machine learning model. European Accounting Review, 1–34. https://doi.org/10.1080/09638180.2025.2463961
- Han, P. K. J., Strout, T. D., Gutheil, C., Germann, C., King, B., Ofstad, E., Gulbrandsen, P., & Trowbridge, R. (2021). How Physicians Manage Medical Uncertainty: A Qualitative Study and Conceptual Taxonomy. Medical decision making : an international journal of the Society for Medical Decision Making, 41(3), 275–291. https://doi.org/10.1177/0272989X21992340
- Han, S., Kelly, E., Nikou, S., & Svee, E.-O. (2022). Aligning artificial intelligence with human values: Reflections from a phenomenological perspective. AI & Society, 37, 1383-1395. https://doi.org/10.1007/s00146-021-01247-4
- Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature medicine, 25(1), 65–69. https://doi.org/10.1038/s41591-018-0268-3
- Haque, A. B., Islam, A. N., Mikalef, P. (2023). Explainable artificial intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. Technological Forecasting and Social Change, 186, 122120. https://doi.org/10.1016/j.techfore.2022.122120
- Harari, Y. N. (2011). Sapiens: A brief history of humankind. Penguin Random House UK.
- Harari, Y. N. (2024). Nexus: A brief history of information networks from the Stone Age to AI (1. izd.). Fern Press.

- Harvey, H. B., & Gowda, V. (2022). Clinical applications of AI in MSK imaging: a liability perspective. Skeletal radiology, 51(2), 235–238. https://doi.org/10.1007/s00256-021-03782-z
- Hasanzadeh, F., Josephson, C. B., Waters, G., Adedinsewo, D., Azizi, Z., & White, J. A. (2025). Bias recognition and mitigation strategies in artificial intelligence healthcare applications. npj Digital Medicine, 8, 154. https://doi.org/10.1038/s41746-025-01503-7
- Hassabis, D. (2017). Artificial intelligence: Chess match of the century. Nature, 544(7648), 413–414. https://doi.org/10.1038/544413a
- Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., & Hussain, A. (2023). Interpreting Black-Box Models: A review on Explainable Artificial intelligence. Cognitive Computation, 16(1), 45–74. https://doi.org/10.1007/s12559-023-10179-8
- Hatherley, J. J. (2020). Limits of trust in medical AI. Journal of Medical Ethics, 46(7), 478–481. https://doi.org/10.1136/medethics-2019-105935
- Hatherley, J., Sparrow, R., & Howard, M. (2022). The Virtues of Interpretable Medical Artificial Intelligence. Cambridge Quarterly of Healthcare Ethics, 1–10. doi:10.1017/S0963180122000305
- Hauswald, R. (2025). Artificial Epistemic Authorities. Social Epistemology, 39(6), 716–725. https://doi.org/10.1080/02691728.2025.2449602
- Hayashi D. (2021). Deep Learning for Lumbar Spine MRI Reporting: A Welcome Tool for Radiologists. Radiology, 300(1), 139–140. https://doi.org/10.1148/radiol.2021210730
- Heaven, P. C. L. (1993). Human values and personality dimensions of the Social Values Inventory. Personality and Individual Differences, 15, 307–312.
- Heidegger, M., (1973). The End of Philosophy, New York: Harper & Row
- Hellrigel-Holderbaum, M., & Dung, L. (2025). Misalignment or misuse? The AGI alignment tradeoff. Philosophical Studies. https://doi.org/10.1007/s11098-025-02403-y
- Helou, M. A., DiazGranados, D., Ryan, M. S., & Cyrus, J. W. (2020). Uncertainty in Decision Making in Medicine: A Scoping Review and Thematic Analysis of Conceptual Models. Academic medicine : journal of the Association of American Medical Colleges, 95(1), 157–165. https://doi.org/10.1097/ACM.0000000000002902

- Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). Explaining collaborative filtering recommendations. In Proceedings of the 2000 ACM conference on Computer supported cooperative work (CSCW '00) (241–250). Association for Computing Machinery. https://doi.org/10.1145/358916.358995
- Hildebrandt, M. (2016). Smart technologies and the end(s) of law. Edward Elgar Publishing.
- Hills, A. (2009). Moral testimony and moral epistemology. Ethics, 120(1), 94–127. https://doi.org/10.1086/648610
- Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? Ethics and Information Technology, 11(1), 19–29. https://doi.org/10.1007/s10676-008-9167-5
- Himmelreich, J. (2019). Responsibility for killer robots. Ethical Theory and Moral Practice, 22, 731–747. https://doi.org/10.1007/s10677-019-10007-9
- Hobbes, T. (1651). Leviathan. London: Andrew Crooke.
- Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors, 57(3), 407–434. https://doi.org/10.1177/0018720814547570
- Hoffman, R., Mueller, S., Klein, G., Jalaeian, M., & Tate, C. (2023). Explainable AI: Roles and stakeholders, desirements and challenges. Frontiers in Computer Science, 5, 1117848. https://doi.org/10.3389/fcomp.2023.1117848
- Hoffmann, A. (2010). Can machines think? An old question reformulated. Minds & Machines, 20(2), 203–212. https://doi.org/10.1007/s11023-010-9193-z
- Hofstadter, D. R., (1985). Metamagical Themas: Questing for the Essence of Mind and Pattern. New York: Basic Books
- Holland, C., Perry, G., & Neyedli, H. F. (2024). Calibrating Trust, Reliance and Dependence in Variable-Reliability Automation. Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting. Human Factors and Ergonomics Society. Annual meeting, 68(1), 604–610. https://doi.org/10.1177/10711813241277531
- Hollnaicher, S. (2025). Value judgments in IAMs. In Assessing feasibility with value-laden models. J. B. Metzler. Springer. https://doi.org/10.1007/978-3-662-70714-2_6
- Holm, H., & Banerjee, S. (2025). Intelligence in animals, humans and machines: A heliocentric view of intelligence? AI & Society, 40, 1169–1171. https://doi.org/10.1007/s00146-024-01931-1

- Holzinger, A. (2021). The next frontier: AI we can really trust. In Kamp, M., Koprinska, I., Bibal, A., Bouadi, T., Frénay, B., Galárraga, L., Oramas, J., Adilova, L., Krishnamurthy, Y., Kang, B., Largeron, C., Lijffijt, J., Viard, T., Welke, P., Ruocco, M., Aune, E., Gallicchio, C., Schiele, G., Pernkopf, F., Graça, G. (Ur.), Machine learning and principles and practice of knowledge discovery in databases (Communications in Computer and Information Science, Vol. 1524, (427–440). Springer. https://doi.org/10.1007/978-3-030-93736-2_33
- Hoppe, J. M., Auer, M. K., Strüven, A., Massberg, S., & Stremmel, C. (2024). ChatGPT With GPT-4 Outperforms Emergency Department Physicians in Diagnostic Accuracy: Retrospective Analysis. Journal of medical Internet research, 26, e56110. https://doi.org/10.2196/56110
- Hoque F. (2024). Shared Decision-Making in Patient Care: Advantages, Barriers and Potential Solutions. Journal of Brown hospital medicine, 3(4), 13–15. https://doi.org/10.56305/001c.122787
- Horn, J. M. (2024). Accident, scandal, disaster: The media framing of corporate crime. Crime, Law and Social Change, 81, 203–229. https://doi.org/10.1007/s10611-023-10111-z
- Hornuf, L., Kück, T., & Schwienbacher, A. (2022). Initial coin offerings, information disclosure, and fraud. Small Business Economics, 58, 1741–1759. https://doi.org/10.1007/s11187-021-00471-y
- Hosseini, R., & Askari, N. (2023). A review of neurological side effects of COVID-19 vaccination. European journal of medical research, 28(1), 102. https://doi.org/10.1186/s40001-023-00992-0
- Howard, R. J., Schold, J. D., & Cornell, D. L. (2005). A 10-year analysis of organ donation after cardiac death in the United States. Transplantation, 80(5), 564–568. https://doi.org/10.1097/01.tp.0000168156.79847.46
- Hristova, T., Magee, L., & Soldatic, K. (2025). The problem of alignment. AI & Society, 40(4), 1439–1453. https://doi.org/10.1007/s00146-024-02039-2
- Hruschka, D., Hackman, J., & Macfarlan, S. (2015). Why do humans help their friends? Proximal and ultimate hypotheses from evolutionary theory. In V. Zeigler-Hill, L. Welling, & T. Shackelford (Ur.), Evolutionary perspectives on social psychology (337–356). Springer. https://doi.org/10.1007/978-3-319-12697-5_20

- Hrvatska liječnička komora. (2006). Kodeks medicinske etike i deontologije. https://www.hlk.hr/EasyEdit/UserFiles/1-kodeks-medicinske-etike-i-deontologije-10062006.pdf, pristup: 30. rujna 2025.
- Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., & Camerer, C. F. (2005). Neural systems responding to degrees of uncertainty in human decision-making. Science, 310(5754), 1680–1683. https://doi.org/10.1126/science.1115327
- Huang H. (2024). Eight challenges in developing theory of intelligence. Frontiers in computational neuroscience, 18, 1388166. https://doi.org/10.3389/fncom.2024.1388166
- Huang, J., Neill, L., Wittbrodt, M., Melnick, D., Klug, M., Thompson, M., Bailitz, J., Loftus, T., Malik, S., Phull, A., Weston, V., Heller, J. A., & Etemadi, M. (2023). Generative Artificial Intelligence for Chest Radiograph Interpretation in the Emergency Department. JAMA network open, 6(10), e2336100. https://doi.org/10.1001/jamanetworkopen.2023.36100
- Huettel, S. A., Song, A. W., & McCarthy, G. (2005). Decisions under uncertainty: probabilistic context influences activation of prefrontal and parietal cortices. The Journal of neuroscience : the official journal of the Society for Neuroscience, 25(13), 3304–3311. https://doi.org/10.1523/JNEUROSCI.5070-04.2005
- Hugo, V. (1887). Les Misérables (Isabel F. Hapgood, Trans.). Thomas Y. Crowell & Co. (Original work published 1862)
- Hurley, M. E., Lang, B. H., Kostick-Quenet, K. M., Smith, J. N., & Blumenthal-Barby, J. (2025). Patient Consent and The Right to Notice and Explanation of AI Systems Used in Health Care. The American journal of bioethics : AJOB, 25(3), 102–114. https://doi.org/10.1080/15265161.2024.2399828
- Hyde, B. V. (2025). Lying increases trust in science. Theory and Society. Advance online publication. https://doi.org/10.1007/s11186-025-09635-1
- Ignatieff, M. (2001). The attack on human rights. Foreign Affairs, 80(6), 102–116. https://doi.org/10.2307/20050331
- Iniesta, R. (2025). The human role to guarantee an ethical AI in healthcare: A five-facts approach. AI and Ethics, 5, 385–397. https://doi.org/10.1007/s43681-023-00353-x

- Institute of Medicine (US) Committee on Quality of Health Care in America. (2001). Crossing the Quality Chasm: A New Health System for the 21st Century. National Academies Press (US). https://doi.org/10.17226/10027
- Intemann, K. (2024). Value transparency and promoting warranted trust in science communication. Synthese, 203(1), 42. https://doi.org/10.1007/s11229-023-04471-1
- Islam, M. R., Ahmed, M. U., Barua, S., & Begum, S. (2022). A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks. Applied Sciences, 12(3), 1353. https://doi.org/10.3390/app12031353
- Jackson, P. (1998). Introduction to expert systems (3. izd.). Addison-Wesley Longman Publishing Co. Inc.
- Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21) (624–635). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445923
- Jain, S., & Wallace, B. C. (2019). Attention is not explanation (arXiv:1902.10186v3). arXiv. https://arxiv.org/abs/1902.10186
- Jamieson, K. H., McNutt, M., Kiermer, V., & Sever, R. (2019). Signaling the trustworthiness of science. Proceedings of the National Academy of Sciences of the United States of America, 116(39), 19231–19236. https://doi.org/10.1073/pnas.1913039116
- Janizek, J. D., Dincer, A. B., Celik, S., Chen, H., Chen, W., Naxerova, K., & Lee, S. (2023). Uncovering expression signatures of synergistic drug responses via ensembles of explainable machine-learning models. Nature Biomedical Engineering, 7(6), 811–829. https://doi.org/10.1038/s41551-023-01034-0
- Jargon, J. (2025). He Had Dangerous Delusions. ChatGPT Admitted It Made Them Worse. The Wall Street Journal. https://www.wsj.com/tech/ai/chatgpt-chatbot-psychology-manic-episodes-57452d14, pristup: 2. prosinca 2025.
- Jha, S., & Topol, E. J. (2016). Adapting to Artificial Intelligence: Radiologists and Pathologists as Information Specialists. JAMA, 316(22), 2353–2354. https://doi.org/10.1001/jama.2016.17438

- Jin, W., Zhang, L., Yang, Z., & Zhang, L. (2020). On failure modes in molecule generation and optimization. Nature Machine Intelligence, 2(8), 437–445. https://doi.org/10.1038/s41563-020-0701-4
- Joffe, S., Manocchia, M., Weeks, J. C., & Cleary, P. D. (2003). What do patients value in their hospital care? An empirical perspective on autonomy centred bioethics. Journal of medical ethics, 29(2), 103–108. https://doi.org/10.1136/jme.29.2.103
- Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. Ethics and Information Technology, 8(4), 195–204. https://doi.org/10.1007/s10676-006-9111-5
- Johnson, D. G., & Verdicchio, M. (2025). The sociotechnical entanglement of AI and values. AI & Society, 40(1), 67–76. https://doi.org/10.1007/s00146-023-01852-5
- Jones, K. (2012). Trustworthiness. Ethics, 123(1), 61–85. https://doi.org/10.1086/667838
- Jongsma, K. R., Sand, M., & Milota, M. (2024). Why we should not mistake accuracy of medical AI for efficiency. NPJ digital medicine, 7(1), 57. https://doi.org/10.1038/s41746-024-01047-2
- Joxhe, M. (2023). Trust in institutions. U K. F. Zimmermann (Ur.), Handbook of labor, human resources and population economics (1–28). Springer. https://doi.org/10.1007/978-3-319-57365-6_252-1
- Kahneman, D. (2011). Thinking, fast and slow. Penguin Books.
- Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). Noise: A flaw in human judgment. Little, Brown Spark.
- Kahrass, H., Bossert, S., Schürmann, C., & Strech, D. (2021). Details of risk-benefit communication in informed consent documents for phase I/II trials. Clinical trials (London, England), 18(1), 71–80. https://doi.org/10.1177/1740774520971770
- Kaji, D. A., Zech, J. R., Kim, J. S., Cho, S. K., Dangayach, N. S., Costa, A. B., & Oermann, E. K. (2019). An attention based deep learning model of clinical events in the intensive care unit. PloS one, 14(2), e0211057. https://doi.org/10.1371/journal.pone.0211057
- Kakad, S., Kale, V., Kakare, A., Kalokhe, A., & Yewale, A. (2023). The evolution of computers. In Proceedings of the 5th International Conference on Communication and Information Processing (ICCIP)–2023. SSRN. https://doi.org/10.2139/ssrn.4687940

- Kamm, F. M. (2019). The trolley problem mysteries. Oxford University Press.
- Kamm, F. M. (2020). Cost-effectiveness analysis and fairness. In Almost over: Aging, dying, dead. Oxford University Press. https://doi.org/10.1093/oso/9780190097158.005.0001
- Kaplan, A., Haenlein, M., (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business Horizons, 62(1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004
- Kapsetaki, M. E. (2024). Polymath: To be or not to be? Postgraduate Medical Journal, 100(1184), 440–441. https://doi.org/10.1093/postmj/qgad091
- Karaca, O., Çalışkan, S. A., & Demir, K. (2021). Medical artificial intelligence readiness scale for medical students (MAIRS-MS) – Development, validity and reliability study. BMC Medical Education, 21, 112. https://doi.org/10.1186/s12909-021-02546-6
- Karimi, A.-H., Barthe, G., Schölkopf, B., & Valera, I. (2023). A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. ACM Computing Surveys, 55(5), Article 95. https://doi.org/10.1145/3527848
- Karnes, H. T., & March, C. (1993). Precision, accuracy, and data acceptance criteria in biopharmaceutical analysis. Pharmaceutical Research, 10, 1420–1426. https://doi.org/10.1023/A:1018958805795
- Kasirzadeh, A. (2025). Two types of AI existential risk: Decisive and accumulative. Philosophical Studies, 182, 1975–2003. https://doi.org/10.1007/s11098-025-02301-3
- Kasser, T., & Ryan, R. M. (1996). Further examining the American dream: Differential correlates of intrinsic and extrinsic goals. Personality and Social Psychology Bulletin, 22(3), 280–287. https://doi.org/10.1177/0146167296223006
- Kelley, D. (2025). Persons, agents, patients – Conceptual housekeeping in a world with group agents. Topoi, 44, 1087–1094. https://doi.org/10.1007/s11245-025-10182-8
- Kelley, J. M., Kraft-Todd, G., Schapira, L., Kossowsky, J., & Riess, H. (2014). The influence of the patient-clinician relationship on healthcare outcomes: a systematic review and meta-analysis of randomized controlled trials. PloS one, 9(4), e94207. https://doi.org/10.1371/journal.pone.0094207

- Kelly, B., S., (2025). From Wile E. Coyote to Nobel laureate: reflections on Geoffrey Hinton's impact on radiology and AI. European radiology, 35(5), 2642–2643. https://doi.org/10.1007/s00330-024-11166-9
- Kelly, B., Judge, C., Bollard, S. M., Clifford, S. M., Healy, G. M., Yeom, K. W., Lawlor, A., & Killeen, R. P. (2020). Radiology artificial intelligence, a systematic evaluation of methods (RAISE): a systematic review protocol. Insights into imaging, 11(1), 133. https://doi.org/10.1186/s13244-020-00929-9
- Kelly, D., Barrett, J., Brand, G., Leech, M., & Rees, C. (2024). Factors influencing decision-making processes for intensive care therapy goals: A systematic integrative review. Australian critical care : official journal of the Confederation of Australian Critical Care Nurses, 37(5), 805–817. https://doi.org/10.1016/j.aucc.2024.02.007
- Kelly, T. A. (1990). The role of values in psychotherapy: A critical review of process and outcome effects. Clinical Psychology Review, 10, 171–186.
- Kerasidou, A., (2020). Artificial intelligence and the ongoing need for empathy, compassion and trust in healthcare. Bulletin of the World Health Organization, 98(4), 245–250. https://doi.org/10.2471/BLT.19.237198
- Kerasidou, A., & Kerasidou, C. (2025). Epistemic authority and medical AI: Epistemological differences and challenges in medical practice. Medicine, Health Care and Philosophy. https://doi.org/10.1007/s11019-025-10306-2
- Khader, F., Müller-Franzes, G., Wang, T., Han, T., Tayebi Arasteh, S., Haarburger, C., Stegmaier, J., Bressem, K., Kuhl, C., Nebelung, S., Kather, J. N., & Truhn, D. (2023). Multimodal Deep Learning for Integrating Chest Radiographs and Clinical Parameters: A Case for Transformers. Radiology, 309(1), e230806. https://doi.org/10.1148/radiol.230806
- Khan, M. A., Hanna, A., Sridhara, S., Chaudhari, H., Me, H. M., Attieh, R. M., & Abu Jawdeh, B. G. (2025). Maintenance Immunosuppression in Kidney Transplantation: A Review of the Current Status and Future Directions. Journal of clinical medicine, 14(6), 1821. https://doi.org/10.3390/jcm14061821
- Khan, M., & Ewuoso, C. (2024). Epistemic (in)justice, social identity and the Black Box problem in patient care. Medicine, health care, and philosophy, 27(2), 227–240. https://doi.org/10.1007/s11019-024-10194-y

- Kiberd, B. A., AlBugami, M. M., Panek, R., & Tennankore, K. (2015). Contraindications to kidney transplantation: uneven grounds?. Transplantation research, 4, 2. https://doi.org/10.1186/s13737-015-0024-x
- Kiener, M. (2021). Artificial intelligence in medicine and the disclosure of risks. AI & society, 36(3), 705–713. https://doi.org/10.1007/s00146-020-01085-w
- Kim, E. E., & Toole, B. A. (1999). Ada and the First Computer. Scientific American, 280(5), 76–81. http://www.jstor.org/stable/26058246
- Kim, J., Maathuis, H., & Sent, D. (2024). Human-centered evaluation of explainable AI applications: a systematic review. Frontiers in artificial intelligence, 7, 1456486. https://doi.org/10.3389/frai.2024.1456486
- King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: reputation and trust in a two-person economic exchange. Science (New York, N.Y.), 308(5718), 78–83. https://doi.org/10.1126/science.1108062
- Kiseleva, A., Kotzinos, D., & De Hert, P. (2022). Transparency of AI in Healthcare as a Multilayered System of Accountabilities: Between Legal Requirements and Technical Limitations. Frontiers in artificial intelligence, 5, 879603. https://doi.org/10.3389/frai.2022.879603
- Kizildag Yirgin, I., Koyluoglu, Y. O., Seker, M. E., Ozkan Gurdal, S., Ozaydin, A. N., Ozcinar, B., Cabioğlu, N., Ozmen, V., & Aribal, E. (2022). Diagnostic Performance of AI for Cancers Registered in A Mammography Screening Program: A Retrospective Analysis. Technology in cancer research & treatment, 21, 15330338221075172. https://doi.org/10.1177/15330338221075172
- Klein, A. S., Messersmith, E. E., Ratner, L. E., Kochik, R., Baliga, P. K., & Ojo, A. O. (2010). Organ donation and utilization in the United States, 1999-2008. American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons, 10(4 Pt 2), 973–986. https://doi.org/10.1111/j.1600-6143.2009.03008.x
- Kluckhohn, C. (1951). Values and value-orientations in the theory of action: An exploration in definition and classification. In T. Parsons & E. A. Shils (Ur.), Toward a general theory of action (388–433). Harvard University Press. https://doi.org/10.4159/harvard.9780674863507.c8

- Knechtle, S. J., & Marson, L. P. (Ur.). (2025). Kidney transplantation: Principles and practice (9. izd.). Elsevier.
- Kneer, M., & Viehoff, J. (2025). The hard problem of AI alignment: Value forks in moral judgment. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25) (2671–2681). Association for Computing Machinery. https://doi.org/10.1145/3715275.3732174
- Knoll, G., Cockfield, S., Blydt-Hansen, T., Baran, D., Kiberd, B., Landsberg, D., Rush, D., Cole, E., & Kidney Transplant Working Group of the Canadian Society of Transplantation (2005). Canadian Society of Transplantation: consensus guidelines on eligibility for kidney transplantation. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne, 173(10), S1–S25. https://doi.org/10.1503/cmaj.1041588
- Kochenderfer, M. J., Amato, C., Chowdhary, G., How, J. P., Davison Reynolds, H. J., Thornton, J. R., Torres-Carrasquillo, P. A., Ure, N. K., & Vian, J. (2015). Integrating automation with humans. In Decision making under uncertainty: Theory and application (291–316). MIT Press.
- Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y. C., & Shaw, T. H. (2021). Measurement of Trust in Automation: A Narrative Review and Reference Guide. Frontiers in psychology, 12, 604977. https://doi.org/10.3389/fpsyg.2021.604977
- Kovesdy, C., P., (2022). Epidemiology of chronic kidney disease: an update 2022. Kidney international supplements, 12(1), 7–11. https://doi.org/10.1016/j.kisu.2021.11.003
- Kraemer, F., van Overveld, K., & Peterson, M. (2011). Is there an ethics of algorithms? Ethics and Information Technology, 13(3), 251–260. https://doi.org/10.1007/s10676-010-9233-7
- Kramer, O., Even, A., Matot, I., Steinberg, Y., & Bitan, Y. (2021). The impact of data quality defects on clinical decision-making in the intensive care unit. Computer methods and programs in biomedicine, 209, 106359. https://doi.org/10.1016/j.cmpb.2021.106359
- Krishna Nk, R., R S, A., & K, S. (2025). Artificial intelligence in radiology: Augmentation, not replacement. Cureus, 17(6), e86247. https://doi.org/10.7759/cureus.86247

- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90. https://doi.org/10.1145/3065386
- Krueger, F., McCabe, K., Moll, J., Kriegeskorte, N., Zahn, R., Strenziok, M., Heinecke, A., & Grafman, J. (2007). Neural correlates of trust. Proceedings of the National Academy of Sciences of the United States of America, 104(50), 20084–20089. https://doi.org/10.1073/pnas.0710103104
- Kudina, O., & van de Poel, I. (2024). A sociotechnical system perspective on AI. Minds & Machines, 34(1), 21. https://doi.org/10.1007/s11023-024-09680-2
- Kuhn, T. S. (1962). The structure of scientific revolutions. University of Chicago Press.
- Kulesza, T., Burnett, M., Wong, W.-K., & Stumpf, S. (2015). Principles of explanatory debugging to personalize interactive machine learning. In Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15) (126–137). Association for Computing Machinery. https://doi.org/10.1145/2678025.2701399
- Kulkova, J., Kulkov, I., Rohrbeck, R., Lu, S., Khwaja, A., Karjaluoto, H., & Mero, J. (2023). Medicine of the future: How and who is going to treat us? Futures, 146, 103097. https://doi.org/10.1016/j.futures.2023.103097
- Kundu, S., (2021). AI in medicine must be explainable. Nature medicine, 27(8), 1328. https://doi.org/10.1038/s41591-021-01461-z
- Kunitomo, K., Harada, T., & Watari, T. (2022). Cognitive biases encountered by physicians in the emergency room. BMC emergency medicine, 22(1), 148. https://doi.org/10.1186/s12873-022-00708-3
- Kurzweil, R. (2005). The singularity is near. Viking Penguin.
- Kuttymuratov, G., Saliev, T., Ainakulov, A., Ayaganov, A., Oshakbayev, K., Zharassov, D., Tuleuzhan, A., & Uderbayev, N. (2025). Kidney and Bladder Transplantation: Advances, Barriers, and Emerging Solutions. Medicina, 61(6), 1045. https://doi.org/10.3390/medicina61061045
- Kyrimi, E., McLachlan, S., Wohlgemut, J. M., Perkins, Z. B., Lagnado, D. A., & Marsh, W.; the ExAIDSS Expert Group. (2025). Explainable AI: Definition and attributes of a good explanation for health AI. AI and Ethics, 5, 3883–3896. https://doi.org/10.1007/s43681-025-00668-x

- Lábaj, M., Oleš, T., & Procházka, G. (2025). Impact of robots and artificial intelligence on labor and skill demand: Evidence from the UK. Eurasian Business Review. https://doi.org/10.1007/s40821-025-00314-w
- Lacave, C., & Díez, F. J. (2002). A review of explanation methods for Bayesian networks. The Knowledge Engineering Review, 17(2), 107–127. https://doi.org/10.1017/S026988890200019X
- Lagerspetz, O. (2025). Was Descartes responsible for the problem of other minds? Philosophical Investigations, 48(3), 249–268. https://doi.org/10.1111/phin.12458
- Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. The Behavioral and brain sciences, 40, e253. https://doi.org/10.1017/S0140525X16001837
- Lakkaraju, H., Slack, D., Chen, Y., Tan, C., & Singh, S. (2022). Rethinking Explainability as a Dialogue: A Practitioner's Perspective. ArXiv, abs/2202.01875.
- Landgrebe, J., & Smith, B. (2025). Why Machines Will Never Rule the World: Artificial Intelligence without Fear. (2nd edition). Abingdon, UK: Routledge.
- Lang, B. H., Nyholm, S., & Blumenthal-Barby, J. (2023). Responsibility Gaps and Black Box Healthcare AI: Shared Responsibilization as a Solution. Digital society : ethics, socio-legal and governance of digital technology, 2(3), 52. https://doi.org/10.1007/s44206-023-00073-z
- Langer, M., Baum, K., Hartmann, K., Hessel, S., Speith, T., & Wahl, J. (2021A). Explainability auditing for intelligent systems: A rationale for multi-disciplinary perspectives. In 2021 IEEE 29th International Requirements Engineering Conference Workshops (REW) (164–168). IEEE. https://doi.org/10.1109/REW53955.2021.00030
- Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021B). What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artificial Intelligence, 296, 103473. https://doi.org/10.1016/j.artint.2021.103473
- Lash, M. T. (2024). HEX: Human-in-the-loop explainability via deep reinforcement learning. Decision Support Systems, 187, 114304. https://doi.org/10.1016/j.dss.2024.114304

- Lauronen, J., Peräsaari, J. P., Saarinen, T., Jaatinen, T., Lempinen, M., & Helanterä, I. (2020). Shorter Cold Ischemia Time in Deceased Donor Kidney Transplantation Reduces the Incidence of Delayed Graft Function Especially Among Highly Sensitized Patients and Kidneys From Older Donors. Transplantation proceedings, 52(1), 42–49. https://doi.org/10.1016/j.transproceed.2019.11.025
- Laux, J., Wachter, S., & Mittelstadt, B. (2023). Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. Regulation & Governance. Advance online publication. https://doi.org/10.1111/rego.12512
- Leblang, D., Smith, M. D., & Wesselbaum, D. (2022). The effect of trust on economic performance and financial access. Economics Letters, 220, 110884. https://doi.org/10.1016/j.econlet.2022.110884
- Lee, E. A. (2020). The coevolution: The entwined futures of humans and machines. The MIT Press. https://doi.org/10.7551/mitpress/12307.001.0001
- Lee, J. D., & See, K. A. (2004). Trust in automation: designing for appropriate reliance. Human factors, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
- Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. Minds & Machines, 17(4), 391–444. https://doi.org/10.1007/s11023-007-9079-x
- Lemaire, R. (2006). Informed consent – a contemporary myth?. The Journal of Bone & Joint Surgery British Volume, 88-B(1), 2-7. https://doi.org/10.1302/0301-620X.88B1.16435
- Lentine, K. L., Smith, J. M., Lyden, G. R., Miller, J. M., Booker, S. E., Dolan, T. G., Temple, K. R., Weiss, S., Handarova, D., Israni, A. K., & Snyder, J. J. (2025). OPTN/SRTR 2023 Annual Data Report: Kidney. American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons, 25(2S1), S22–S137. https://doi.org/10.1016/j.ajt.2025.01.020
- Lewis, A., Koukoura, A., Tsianos, G. I., Gargavanis, A. A., Nielsen, A. A., & Vassiliadis, E. (2021). Organ donation in the US and Europe: The supply vs demand imbalance. Transplantation reviews (Orlando, Fla.), 35(2), 100585. https://doi.org/10.1016/j.trre.2020.100585

- Li, X., Wang, P., Zhu, Y., Zhao, W., Pan, H., & Wang, D. (2024A). Interpretable machine learning model for predicting acute kidney injury in critically ill patients. BMC Medical Informatics & Decision Making, 24, Article 148. https://doi.org/10.1186/s12911-024-02537-9
- Li, Y., Wu, B., Huang, Y., & Luan, S. (2024B). Developing trustworthy artificial intelligence: Insights from research on interpersonal, human-automation, and human-AI trust. Frontiers in Psychology, 15, 1382693. https://doi.org/10.3389/fpsyg.2024.1382693
- Liao, Q. V., Varshney, K. R. (2021). Human-centered explainable AI (XAI): From algorithms to user experiences. arXiv. https://doi.org/10.48550/arXiv.2110.10790
- Lighthill, J. (1973). Artificial intelligence: A paper symposium. London: Science Research Council.
- Lin, P., Abney, K., & Jenkins, R. (Ur.). (2017). Robot ethics 2.0: From autonomous cars to artificial intelligence. Oxford University Press. https://doi.org/10.1093/oso/9780190652951.001.0001
- Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy (Basel, Switzerland), 23(1), 18. https://doi.org/10.3390/e23010018
- Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., Ledsam, J. R., Schmid, M. K., Balaskas, K., Topol, E. J., Bachmann, L. M., Keane, P. A., & Denniston, A. K. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet. Digital health, 1(6), e271–e297. https://doi.org/10.1016/S2589-7500(19)30123-2
- Liu, X., Glocker, B., McCradden, M. M., Ghassemi, M., Denniston, A. K., & Oakden-Rayner, L. (2022). The medical algorithmic audit. The Lancet. Digital health, 4(5), e384–e397. https://doi.org/10.1016/S2589-7500(22)00003-6
- Liu, Y., Gadepalli, K. K., Norouzi, M., Dahl, G., Kohlberger, T., Venugopalan, S., Boyko, A. S., Timofeev, A., Nelson, P. Q., Corrado, G., Hipp, J. D., Peng, L., & Stumpe, M. C. (2017). Detecting cancer metastases on gigapixel pathology images (arXiv:1703.02442). arXiv. https://arxiv.org/abs/1703.02442
- Liu, Y., Jain, A., Eng, C., Way, D. H., Lee, K., Bui, P., Kanada, K., de Oliveira Marinho, G., Gallegos, J., Gabriele, S., Gupta, V., Singh, N., Natarajan, V., Hofmann-

Wellenhof, R., Corrado, G. S., Peng, L. H., Webster, D. R., Ai, D., Huang, S. J., Liu, Y., … Coz, D. (2020). A deep learning system for differential diagnosis of skin diseases. Nature medicine, 26(6), 900–908. https://doi.org/10.1038/s41591-020-0842-3

- Liu, Y., Liu, C., Zheng, J., Xu, C., & Wang, D. (2025). Improving explainability and integrability of medical AI to promote health care professional acceptance and use: Mixed systematic review. Journal of Medical Internet Research, 27, e73374. https://doi.org/10.2196/73374
- Llorca Albareda, J., (2025). Uncovering the gap: Challenging the agential nature of AI responsibility problems. AI and Ethics, 5, 3857–3870. https://doi.org/10.1007/s43681-025-00685-w
- Llorca Albareda, J., García, P., & Lara, F. (2023). The moral status of AI entities. In F. Lara & J. Deckers (Ur.), Ethics of artificial intelligence (The International Library of Ethics, Law and Technology, Vol. 41, (59–83). Springer. https://doi.org/10.1007/978-3-031-48135-2_4
- Logg, J. M., Minson, J. A., & Moore, D. A. (2018). Do people trust algorithms more than companies realize? Harvard Business Review. https://hbr.org/2018/10/do-people-trust-algorithms-more-than-companies-realize, pristup: 1. travnja 2025.
- Logg, J. M. (2017). Theory of machine: When do people rely on algorithms? Harvard Business School Working Paper, No. 17-086.
- Lombardo, P. A. (2025). "The craze for legal proceedings": Another look at Schloendorff v. New York Hospital. Health Matrix, 35(1), 89–116. https://scholarlycommons.law.case.edu/healthmatrix/vol35/iss1/4
- London, A. J. (2019). Artificial intelligence and black-box medical decisions: Accuracy versus explainability. The Hastings Center Report, 49(1), 15–21. https://doi.org/10.1002/hast.973
- Lopez-Sanchez, M., Serramia, M., Rodriguez-Aguilar, J. A., Morales, J., & Wooldridge, M. (2017). Automating decision making to help establish norm-based regulations. In Proceedings of the 16th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '17) (1613–1615). International Foundation for Autonomous Agents and Multiagent Systems.
- Loyola-González, O. (2019). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. IEEE Access, 7, 154096–154113. https://doi.org/10.1109/ACCESS.2019.2949286

- Lu, M. Y., Chen, B., Williamson, D. F. K., Chen, R. J., Zhao, M., Chow, A. K., Ikemura, K., Kim, A., Pouli, D., Patel, A., Soliman, A., Chen, C., Ding, T., Wang, J. J., Gerber, G., Liang, I., Le, L. P., Parwani, A. V., Weishaupt, L. L., & Mahmood, F. (2024). A multimodal generative AI copilot for human pathology. Nature, 634(8033), 466–473. https://doi.org/10.1038/s41586-024-07618-3
- Luhmann, N. (1988). Familiarity, confidence, trust. In D. Gambetta (Ur.), Trust: Making and breaking cooperative relations (94–107). Basil Blackwell.
- Luhmann, N. (2017). Trust and power. Polity Press.
- Luo, J., Meyer, J. J., & Knobbout, M. (2017). Reasoning about opportunistic propensity in multi-agent systems. In G. Sukthankar & J. Rodriguez-Aguilar (Ur.), Autonomous agents and multiagent systems: AAMAS 2017, (205–221). Springer. https://doi.org/10.1007/978-3-319-71682-4_13
- Luther, V. P., & Crandall, S. J. (2011). Commentary: ambiguity and uncertainty: neglected elements of medical education curricula?. Academic medicine : journal of the Association of American Medical Colleges, 86(7), 799–800. https://doi.org/10.1097/ACM.0b013e31821da915
- Maalsen, S. (2023). Algorithmic epistemologies and methodologies: Algorithmic harm, algorithmic care and situated algorithmic knowledges. Progress in Human Geography, 47(2), 197–214. https://doi.org/10.1177/03091325221149439
- Maiter, A., Hocking, K., Matthews, S., Taylor, J., Sharkey, M., Metherall, P., Alabed, S., Dwivedi, K., Shahin, Y., Anderson, E., Holt, S., Rowbotham, C., Kamil, M. A., Hoggard, N., Balasubramanian, S. P., Swift, A., & Johns, C. S. (2023). Evaluating the performance of artificial intelligence software for lung nodule detection on chest radiographs in a retrospective real-world UK population. BMJ open, 13(11), e077348. https://doi.org/10.1136/bmjopen-2023-077348
- Mancosu, P., Poggiolesi, F., & Pincock, C. (2023). Mathematical explanation. In E. N. Zalta & U. Nodelman (Ur.), The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/fall2023/entries/mathematics-explanation/
- Mandava, A., Pace, C., Campbell, B., Emanuel, E., & Grady, C. (2012). The quality of informed consent: mapping the landscape. A review of empirical data from developing and developed countries. Journal of medical ethics, 38(6), 356–365. https://doi.org/10.1136/medethics-2011-100178

- Marcus, G., (2024), Superhuman AGI is not nigh https://garymarcus.substack.com/p/superhuman-agi-is-not-nigh, pristup: 2. lipnja 2025.
- Marey, A., Arjmand, P., Alerab, A. D. S., Eslami, M. J., Saad, A. M., Sanchez, N., Umair, M. (2024). Explainability, transparency and black box challenges of AI in radiology: Impact on patient care in cardiovascular radiology. Egyptian Journal of Radiology and Nuclear Medicine, 55, 183. https://doi.org/10.1186/s43055-024-01356-2
- Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. Journal of biomedical informatics, 113, 103655. https://doi.org/10.1016/j.jbi.2020.103655
- Martens, D., Baesens, B., Van Gestel, T., & Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. European Journal of Operational Research, 183(3), 1466–1476. https://doi.org/10.1016/j.ejor.2006.04.051
- Martinelli, V., Lumer, E. L. L., Chiappedi, M., Politi, P., Gregorini, M., Rampino, T., Peri, A., Pietrabissa, A., & Fusar-Poli, L. (2024). Ethical Issues in Living Donor Kidney Transplantation: An Update from a Psychosocial Perspective. Healthcare (Basel, Switzerland), 12(18), 1832. https://doi.org/10.3390/healthcare12181832
- Martínez, N., Agudo, U., Matute, H., (2022). Human cognitive biases present in Artificial Intelligence, Revista Internacional de los Estudios Vascos. RIEV, 67, 2
- Marx, K. (1976). Capital: A Critique of Political Economy. Vol. I. Harmondsworth: Penguin.
- Mathew, D. E., Ebem, D. U., Ikegwu, A. C., Ukeoma, P. E., & Dibiaezue, N. F. (2025). Recent emerging techniques in explainable artificial intelligence to enhance the interpretability and understanding of AI models for humans. Neural Processing Letters, 57(1), 16. https://doi.org/10.1007/s11063-025-11732-2
- Mathews, S. M. (2019). Explainable artificial intelligence applications in NLP, biomedical, and malware classification: A literature review. In Intelligent Computing: Proceedings of the Computing Conference (1269–1292).
- Matiasek, J., & Wynia, M. K. (2008). Reconceptualizing the informed consent process at eight innovative hospitals. Joint Commission journal on quality and patient safety, 34(3), 127–137. https://doi.org/10.1016/s1553-7250(08)34015-x

- Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and Information Technology, 6(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1
- Matthias, A. (2011). Algorithmic moral control of war robots: Philosophical questions. Law, Innovation and Technology, 3(2), 279–301.
- Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of Management Review, 20(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335
- Mazur, D. J. (1988). Why the goals of informed consent are not realized: Treatise on informed consent for the primary care physician. Journal of General Internal Medicine, 3(4), 370–380. https://doi.org/10.1007/BF02595796
- McCarthy, J., (1958). Programs with common sense. In Proceedings of the Teddington Conference on the Mechanisation of Thought Processes (77–84).
- McCarthy, J., (2007). From here to human-level AI. Artificial Intelligence, 171(18), 1174–1182. https://doi.org/10.1016/j.artint.2007.10.009
- McCarthy, J., (1956). What is artificial intelligence?, https://www-formal.stanford.edu/jmc/whatisai.pdf, pristup: 1. travnja 2025.
- McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. E. (1955). A proposal for the Dartmouth summer research project on artificial intelligence, dostupno: https://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf, pristup, 7. studenog 2025.
- McCoy, M. (2008). Autonomy, consent, and medical paternalism: Legal issues in medical intervention. Journal of Alternative and Complementary Medicine, 14(6), 785–792. https://doi.org/10.1089/acm.2007.0803
- McCracken, J., Bajwah, S., & Shaw, C. (2025). Factors Affecting Healthcare Professionals' Decision-Making Around Parenteral Nutrition in Advanced Cancer: A Systematic Review. Journal of human nutrition and dietetics : the official journal of the British Dietetic Association, 38(1), e70009. https://doi.org/10.1111/jhn.70009
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics, 5(4), 115–133. https://doi.org/10.1007/BF02478259
- McDougall, R., J., (2019). Computer knows best? The need for value-flexibility in medical AI. Journal of medical ethics, 45(3), 156–160. https://doi.org/10.1136/medethics-2018-105118

- McDuff, D., Schaekermann, M., Tu, T., Palepu, A., Wang, A., Garrison, J., Singhal, K., Sharma, Y., Azizi, S., Kulkarni, K., Hou, L., Cheng, Y., Liu, Y., Mahdavi, S. S., Prakash, S., Pathak, A., Semturs, C., Patel, S., Webster, D. R., Dominowska, E., Gottweis, J., Barral, J., Chou, K., Corrado, G., Natarajan, V. (2025). Towards accurate differential diagnosis with large language models. Nature. https://doi.org/10.1038/s41586-025-08869-4
- McKendrick, J., & Thurai, A. (2022). AI isn't ready to make unsupervised decisions. Harvard Business Review. https://hbr.org/2022/09/ai-isnt-ready-to-make-unsupervised-decisions, pristup: 1. travnja 2025.
- Melegati, J., & Wang, X. (2020). Hypotheses Elicitation in Early-Stage Software Startups Based on Cognitive Mapping. Agile Processes in Software Engineering and Extreme Programming: 21st International Conference on Agile Software Development, XP 2020, Copenhagen, Denmark, June 8–12, 2020, Proceedings, 383, 211–220. https://doi.org/10.1007/978-3-030-49392-9_14
- Mercado, J. E., Rupp, M. A., Chen, J. Y. C., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent agent transparency in human–agent teaming for multi-UxV management. Human Factors, 58(3), 401–415. https://doi.org/10.1177/0018720815621206
- Merchant, C. (2008). Secrets of Nature: The Bacon Debates Revisited. Journal of the History of Ideas, 69(1), 147–162. http://www.jstor.org/stable/30139672
- Merton, R. K. (1943). The normative structure of science. In N. W. Storer (Ur.), The sociology of science: Theoretical and empirical investigations (267–278). Chicago, IL: University of Chicago Press.
- Mienye, I. D., Swart, T. G., Obaido, G., Jordan, M., & Ilono, P. (2025). Deep Convolutional Neural Networks in Medical Image Analysis: A Review. Information, 16(3), 195. https://doi.org/10.3390/info16030195
- Miguel, I., Sanz, B., & Lazcoz, G. (2020). Machine learning in the EU health care context: Exploring the ethical, legal and social issues. Information, Communication & Society, 23(8), 1139–1153. https://doi.org/10.1080/1369118X.2019.1658300
- Mihan, A., Pandey, A., & Van Spall, H. G. (2024). Mitigating the risk of artificial intelligence bias in cardiovascular care. The Lancet. Digital health, 6(10), e749–e754. https://doi.org/10.1016/S2589-7500(24)00155-9

- Mikulić, D., & Kocman, B. (2020). Eurotransplant i mjesto Republike Hrvatske u njemu / Eurotransplant and the role of Croatia in Eurotransplant. Medicina Fluminensis, 56(4), 498-503. https://doi.org/10.21860/medflum2020_245224
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007
- Millgram, E. (2015). The great endarkenment: Philosophy for an age of hyperspecialization. Oxford University Press.
- Milligan, E., & Jones, J. (2016). Rethinking Autonomy and Consent in Healthcare Ethics. InTech. doi: 10.5772/65765
- Ministarstvo zdravstva Republike Hrvatske. (2016). Nacionalne smjernice za obradu i procjenu primatelja i darivatelja bubrega (2. izd.). https://zdravlje.gov.hr/UserDocsImages/dokumenti/Tekstovi%20razni/Nacionalne%20smjernice%20za%20obradu%20i%20odabir%20primatelja%20i%20darivatelja%20bubrega.pdf, pristup: 21. studenog 2025.
- Ministarstvo zdravstva Republike Hrvatske. (2025). Prikaz rezultata Nacionalnog transplantacijskog programa 2024 https://zdravlje.gov.hr/UserDocsImages/2025_Objave/Prikaz%20rezultata%20nacionalnog%20transplantacijskog%20programa%202024..pdf, pristup: 21. studenog 2025.
- Minsky, M. (1968). Semantic Information Processing. MIT, Press, Cambridge, MA.
- Minsky, M. (1991). "Logical Versus Analogical or Symbolic versus Connectionist or rsus Scruffy," AI Magazine, Vol. 12:2, (35-51).
- Minsky, M., Papert, S. (1969). Perceptrons: An introduction to computational geometry. Cambridge, MA: MIT Press.
- Misselhorn, C. (2022). Artificial Moral Agents: Conceptual Issues and Ethical Controversy. In S. Voeneky, P. Kellmeyer, O. Mueller, & W. Burgard (Ur.), The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives (31–49). Cambridge: Cambridge University Press.
- Mitani, A., Huang, A., Venugopalan, S., Sidibé, D., & Lu, C. (2020). Detection of anaemia from retinal fundus images via deep learning. Nature Biomedical Engineering, 4(1), 18–27. https://doi.org/10.1038/s41551-019-0487-z
- Mitchell, T. (2025). Trust and transparency in artificial intelligence. Philosophy & Technology, 38(87). https://doi.org/10.1007/s13347-025-00916-2

- Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In Proceedings of the Conference on Fairness, Accountability, and Transparency, (279–288). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287574
- Mohale, V. Z., Obagbuwa, I. C. (2025). A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. Frontiers in artificial intelligence, 8, 1526221. https://doi.org/10.3389/frai.2025.1526221
- Mohapatra, R. K., Jolly, L., & Dakua, S. P. (2025). Advancing explainable AI in healthcare: Necessity, progress, and future directions. Computational biology and chemistry, 119, 108599. https://doi.org/10.1016/j.compbiolchem.2025.108599
- Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (2. izd.). Christoph Molnar. https://christophm.github.io/interpretable-ml-book/
- Monett, D., Lewis, C.W.P. (2018). Getting Clarity by Defining Artificial Intelligence—A Survey. In: Müller, V. (Ur.) Philosophy and Theory of Artificial Intelligence 2017. PT-AI 2017. Studies in Applied Philosophy, Epistemology and Rational Ethics, Vol 44. Springer, Cham. https://doi.org/10.1007/978-3-319-96448-5_21
- Monett, D., Lewis, C.W.P.., Thórisson, K.R.., Bach, J., Baldassarre, G., Granato, G., Berkeley, I.S.N.., Chollet, F., Crosby, M., Shevlin, H., Fox, J., Laird, J.E.., Legg, S., Lindes, P., Mikolov, T., Rapaport, W.J.., Rojas, R., Rosa, M., Stone, P., Sutton, R.S.., Yampolskiy, R.V.., Wang, P., Schank, R., Sloman, A. & Winfield, A. (2020). Special Issue "On Defining Artificial Intelligence"—Commentaries and Author's Response. Journal of Artificial General Intelligence, 11(2), 2020. 1-100. https://doi.org/10.2478/jagi-2020-0003
- Mooghali, M., Stroud, A. M., Yoo, D. W., Barry, B. A., Grimshaw, A. A., Ross, J. S., Zhu, X., & Miller, J. E. (2024). Trustworthy and ethical AI-enabled cardiovascular care: A rapid review. BMC Medical Informatics and Decision Making, 24, 247. https://doi.org/10.1186/s12911-024-02653-6
- Moor, J. H. (1979). Are there decisions computers should never make? Nature and System, 1(4), 217–229.
- Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. IEEE Intelligent Systems, 21(4), 18–21. https://doi.org/10.1109/MIS.2006.80

- Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: A mapping review. Social science & medicine (1982), 260, 113172. https://doi.org/10.1016/j.socscimed.2020.113172
- Morrow, E., Zidaru, T., Ross, F., Mason, C., Patel, K. D., Ream, M., & Stockley, R. (2023). Artificial intelligence technologies and compassion in healthcare: A systematic scoping review. Frontiers in psychology, 13, 971044. https://doi.org/10.3389/fpsyg.2022.971044
- Mosakas, K. (2021). On the moral status of social robots: Considering the consciousness criterion. AI & Society, 36, 429–443. https://doi.org/10.1007/s00146-020-01002-1
- Muehlhauser, L., & Helm, L. (2012). Intelligence explosion and machine ethics. In A. Eden, J. Søraker, J. H. Moor, & E. Steinhart (Ur.), Singularity hypotheses: A scientific and philosophical assessment (205–238). Springer. https://doi.org/10.1007/978-3-642-14708-9_11
- Mühlhoff, R. (2025). The ethics of AI: Power, critique, responsibility. Bristol University Press. https://doi.org/10.51952/9781529249262
- Mukherjee, S. (2018). This cat sensed death. What if computers could, too? The New York Times Magazine, https://www.nytimes.com/2018/01/03/magazine/the-dying-algorithm.html, pristup: 2. prosinca 2025.
- Muller, E., (2025). The future of transplantation: Not a privilege, but a fundamental right: Presidential address at the 30th International Congress of The Transplantation Society, Istanbul, Turkey, September 22–25, 2024. Transplantation, 109(6), 901–904. https://doi.org/10.1097/TP.0000000000005340
- Müller, H., Mayrhofer, M. T., van Veen, E.-B., & Holzinger, A. (2021). The ten commandments of ethical medical AI. Computer, 54(7), 119–123. https://doi.org/10.1109/MC.2021.3074263
- Murphy, R., & Woods, D. D. (2009). Beyond Asimov: The three laws of responsible robotics. IEEE Intelligent Systems, 24(4), 14–20. https://doi.org/10.1109/MIS.2009.69
- Muyskens, K., Ma, Y., Menikoff, J., Li, H., Chen, R., Patel, S., & Thompson, L. (2025). When can we kick (some) humans “out of the loop”? An examination of the use of AI in medical imaging for lumbar spinal stenosis. Asian Bioethics Review, 17(3), 207–223. https://doi.org/10.1007/s41649-024-00290-9

- Nakamura, Y., Sonoda, Y., Yamagishi, Y., Kikuchi, T., Nakao, T., Miki, S., Hanaoka, S., Yoshikawa, T., & Abe, O. (2024). It is Not Time to Kick Out Radiologists. Asian bioethics review, 17(1), 9–15. https://doi.org/10.1007/s41649-024-00325-1
- Nancy, J.-L. (2017). The Banality of Heidegger. Fordham Univ Press.
- National Transportation Safety Board. (2019). Highway accident report: Collision between vehicle controlled by developmental automated driving system and pedestrian (Report No. NTSB/HAR-19/03, PB2019-101402). U.S. Government Publishing Office.
- Nature Editorial. (2016). Digital intuition. Nature, 529(7587), 437. https://doi.org/10.1038/529437a
- Nazer, L. H., Zatarah, R., Waldrip, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., Hicklen, R. S., Moukheiber, L., Moukheiber, D., Ma, H., & Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS digital health, 2(6), e0000278. https://doi.org/10.1371/journal.pdig.0000278
- Newport, C. (2021). A world without email: Find Focus and Transform the Way You Work Forever (from the NYT bestselling productivity expert). Penguin UK.
- Newton, K. (2001). Trust, social capital, civil society, and democracy. International Political Science Review, 22(2), 201–214. https://www.jstor.org/stable/1601186
- Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (427–436). IEEE. https://doi.org/10.1109/CVPR.2015.7298640
- Nickel, P. J., Franssen, M., & Kroes, P. (2010). Can we make sense of the notion of trustworthy technology? Knowledge, Technology & Policy, 23(3-4), 429-444. https://doi.org/10.1007/s12130-010-9124-6
- Nikiforidis, K., Kyrtsoglou, A., Vafeiadis, T., Kotsiopoulos, T., Nizamis, A., Ioannidis, D., Votis, K., Tzovaras, D., & Sarigiannidis, P. (2025). Enhancing transparency and trust in AI-powered manufacturing: A survey of explainable AI (XAI) applications in smart manufacturing in the era of Industry 4.0/5.0. ICT Express, 11(1), 135–148. https://doi.org/10.1016/j.icte.2025.XX.XXX
- Norgeot, B., Glicksberg, B. S., & Butte, A. J. (2019). A call for deep-learning healthcare. Nature medicine, 25(1), 14–15. https://doi.org/10.1038/s41591-018-0320-3

- Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. Patterns (New York, N.Y.), 2(10), 100347. https://doi.org/10.1016/j.patter.2021.100347
- Norvig, P. (2017). On Chomsky and the two cultures of statistical learning. In W. Pietsch, J. Wernecke, & M. Ott (Ur.), Berechenbarkeit der Welt? https://doi.org/10.1007/978-3-658-12153-2_3
- Novacescu, D., Latcu, S. C., Raica, M., Baderca, F., Dumitru, C.-S., Daminescu, L., Bardan, R., Dema, V., Croitor, A., Cut, T. G., & Cumpanas, A. A. (2024). Surgical Strategies for Renal Transplantation: A Pictorial Essay. Journal of Clinical Medicine, 13(14), 4188. https://doi.org/10.3390/jcm13144188
- Nowak, B., Meier, Y., & Krämer, N. (2025). Challenges in defining and measuring trust and distrust in science. In A. Fage-Butler, L. Ledderer, & K. H. Nielsen (Ur.), Science communication and trust (365–384). Palgrave Macmillan. https://doi.org/10.1007/978-981-96-1289-5_18
- Nyholm, S. (2019). Other minds, other intelligences: The problem of attributing agency to machines. Cambridge Quarterly of Healthcare Ethics, 28(4), 592–598. https://doi.org/10.1017/S0963180119000537
- Nyholm, S. (2020). Humans and robots: Ethics, agency, and anthropomorphism. Rowman & Littlefield International.
- Nyholm, S. (2023A). A new control problem? Humanoid robots, artificial intelligence, and the value of control. AI and Ethics, 3, 1229–1239. https://doi.org/10.1007/s43681-022-00231-y
- Nyholm, S. (2023B). Responsibility gaps, value alignment, and meaningful human control over artificial intelligence. In Risk and responsibility in context, (1. izd.), Vol. 1, (191–213). Routledge. https://doi.org/10.4324/9781003276029-14
- O'Neil, C., (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown Publishing Group.
- O'Neill, O. (2002). Autonomy and trust in bioethics. Cambridge University Press.
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science (New York, N.Y.), 366(6464), 447–453. https://doi.org/10.1126/science.aax2342
- Oedingen, C., Bartling, T., Dierks, M. L., Mühlbacher, A. C., Schrem, H., & Krauth, C. (2020). Public preferences for the allocation of donor organs for transplantation:

Focus group discussions. Health expectations : an international journal of public participation in health care and health policy, 23(3), 670–680. https://doi.org/10.1111/hex.13047

- Olawade, D. B., Marinze, S., Qureshi, N., Weerasinghe, K., & Teke, J. (2025). Transforming organ donation and transplantation: Strategies for increasing donor participation and system efficiency. European journal of internal medicine, 133, 14–24. https://doi.org/10.1016/j.ejim.2024.11.010
- Orbán, F., & Stefkovics, Á. (2025). Trust in artificial intelligence: A survey experiment to assess trust in algorithmic decision-making. AI & Society, 40, 4955–4969. https://doi.org/10.1007/s00146-025-02237-6
- Ord, T. (2021). The Precipice: Existential risk and the future of humanity. Grand Central Publishing.
- Orlowski, E. J., Norhashim, H., & Wey, T. K. (2025). 'Too much alignment; not enough culture': Re-balancing cultural alignment practices in LLMs. arXiv. https://arxiv.org/abs/2509.26167
- Ortale, R., Costa, G., & Forestiero, A. (Ur.). (2025). Explainable machine intelligence in healthcare. Springer Cham. https://doi.org/10.1007/978-3-031-91379-2
- Oxford Essential Quotations. (2014). In Oxford University Press eBooks. https://doi.org/10.1093/acref/9780191794230.001.0001
- Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). Minds and Machines, 29(3), 441–459. https://doi.org/10.1007/s11023-019-09502-w
- Paliwal, G., Kumar, A., Sharma, K. P., Bhargava, D., & Shrimal, V. M. (2025). Transformative impact of explainable artificial intelligence: Bridging complexity and trust. Discover Artificial Intelligence, 5, 51. https://doi.org/10.1007/s44163-025-00281-1
- Panigutti, C., Hamon, R., Hupont, I., Fernandez Llorca, D., Fano Yela, D., Junklewitz, H., Scalzo, S., Mazzini, G., Sanchez, I., Soler Garrido, J., & Gomez, E. (2023). The role of explainable AI in the context of the AI Act. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23) (1139–1150). Association for Computing Machinery. https://doi.org/10.1145/3593013.3594069
- Park, H. J. (2024). Patient perspectives on informed consent for medical AI: A web-based experiment. Digital health, 10, 20552076241247938. https://doi.org/10.1177/20552076241247938

- Parker, W. F., & Chin, M. H. (2020). Epistemic Authority and Trust in Shared Decision Making About Organ Transplantation. AMA journal of ethics, 22(5), E408–E415. https://doi.org/10.1001/amajethics.2020.408
- Parvin, N., Joo, S. W., Jung, J. H., & Mandal, T. K. (2025). Multimodal AI in Biomedicine: Pioneering the Future of Biomaterials, Diagnostics, and Personalized Healthcare. Nanomaterials (Basel, Switzerland), 15(12), 895. https://doi.org/10.3390/nano15120895
- Patel, K., Huang, S., Rashid, A., Varghese, B., & Gholamrezanezhad, A. (2023). A Narrative Review of the Use of Artificial Intelligence in Breast, Lung, and Prostate Cancer. Life (Basel, Switzerland), 13(10), 2011. https://doi.org/10.3390/life13102011
- Pearson, H., Ledford, H., Hutson, M. & Van Noorden, R. (2025). Exclusive: the most-cited papers of the twenty-first century, Nature, 640, https://www.nature.com/articles/d41586-025-01125-9/, pristup: 15. travnja 2025.
- Pellegrino, E. D., & Thomasma, D. C. (1987). The conflict between autonomy and beneficence in medical ethics: proposal for a resolution. The Journal of contemporary health law and policy, 3, 23–46.
- Perni, S., Lehmann, L. S., & Bitterman, D. S. (2023). Patients should be informed when AI systems are used in clinical trials. Nature Medicine, 29, 1890–1891. https://doi.org/10.1038/s41591-023-02367-8
- Perrigo, B., (2025). Google DeepMind CEO Demis Hassabis on AI in the Military and What AGI Could Mean for Humanity, Time, https://time.com/7280740/demis-hassabis-interview/, pristup: 2. prosinca 2025.
- Perry, R. B. (1926). The general theory of value. Longmans, Green and Company.
- Persijn, G. G. (2006). Allocation of organs, particularly kidneys, within Eurotransplant. Human Immunology, 67(6), 419–423. https://doi.org/10.1016/j.humimm.2006.03.008
- Petersen, E., Potdevin, Y., Mohammadi, E., Zidowitz, S., Breyer, S., Nowotka, D., Henn, S., Pechmann, L., Leucker, M., Rostalski, P., & Herzog, C. (2022). Responsible and regulatory conform machine learning for medicine: A survey of challenges and solutions. IEEE Access, 10, 58375–58398. https://doi.org/10.1109/ACCESS.2022.3180567

- Peterson, M. (2007). Should the precautionary principle guide our actions or our beliefs? Journal of Medical Ethics, 33(1), 5–10.
- Petkovic, D. (2022). It is not “accuracy vs. explainability” — we need both for trustworthy AI systems (arXiv No. 2212.11136v2). arXiv. https://doi.org/10.48550/arXiv.2212.11136
- Petrović, S., (1959). Može li stroj prevoditi poeziju "Strojno prevođenje i statistika u jeziku"
- Pham, T. (2025). Ethical and legal considerations in healthcare AI: innovation and policy for safe and fair use. Royal Society open science, 12(5), 241873. https://doi.org/10.1098/rsos.241873
- Pickering, B. (2021). Trust, but Verify: Informed Consent, AI Technologies, and Public Health Emergencies. Future Internet, 13(5), 132. https://doi.org/10.3390/fi13050132
- Pieters, W. (2011). Explanation and trust: What to tell the user in security and AI? Ethics and Information Technology, 13(1), 53–64. https://doi.org/10.1007/s10676-010-9253-3
- Pietrzykowski, T., & Smilowska, K. (2021). The reality of informed consent: Empirical studies on patient comprehension—Systematic review. Trials, 22, 57. https://doi.org/10.1186/s13063-020-04969-w
- Placani, A. (2024). Anthropomorphism in AI: Hype and fallacy. AI and Ethics, 4(3), 691–698. https://doi.org/10.1007/s43681-024-00419-4
- Poibeau, T. (2017). Machine translation. MIT Press.
- Poole, D. L., Mackworth, A. K., (2010). Artificial Intelligence: Foundations of Computational Agents. Cambridge: Cambridge University Press.
- Poslon, L. (2025). Osnaživanje povjerenja osiguravanjem pravednosti u sustavima umjetne inteligencije u zdravstvenoj skrbi. Nova prisutnost, 23(2), 277–294. https://doi.org/10.31192/np.23.2.4
- Poslon, L., & Čartolovni, A. (2025). Improving explanation measures inside the decision-making process to increase the potential of AI applications in healthcare. In R. Ortale, G. Costa, & A. Forestiero (Ur.), Explainable machine intelligence in healthcare (41–55). Springer, Cham. https://doi.org/10.1007/978-3-031-91379-2_3
- Potter, N. N. (2002). How can I be trusted? A virtue theory of trustworthiness. Rowman & Littlefield Publishers.

- Price, W., N., II. (2018). Medical malpractice and black-box medicine. U I. Cohen, H. Lynch, E. Vayena, & U. Gasser (Ur.), Big data, health law, and bioethics (pp. 295–306). Cambridge University Press.
- Price, W., N., II. (2015). Black-Box Medicine. Harvard Journal of Law & Technology, 28(2), 419. https://repository.law.umich.edu/articles/2275/
- Pruski, M. (2024). AI-enhanced healthcare: Not a new paradigm for informed consent. Bioethical Inquiry, 21(3), 475–489. https://doi.org/10.1007/s11673-023-10320-0
- Purves, D., Jenkins, R., & Strawser, B. J. (2015). Autonomous machines, moral judgment, and acting for the right reasons. Ethical Theory and Moral Practice, 18(5), 851–872. https://doi.org/10.1007/s10677-015-9563-y
- Putnam, R. D. (2000). Bowling alone: The collapse and revival of American community. Simon & Schuster.
- Puzio, A. (2025). Robot, let us pray! Can and should robots have religious functions? An ethical exploration of religious robots. AI & Society, 40, 1019–1035. https://doi.org/10.1007/s00146-023-01812-z
- Quinn, T. P., Senadeera, M., Jacobs, S., Coghlan, S., & Le, V. (2021). Trust and medical AI: the challenges we face and the expertise needed to overcome them. Journal of the American Medical Informatics Association : JAMIA, 28(4), 890–894. https://doi.org/10.1093/jamia/ocaa268
- Rajkomar, A., Kannan, A., Chen, K., Vardoulakis, L., Chou, K., Cui, C., & Dean, J. (2019). Automatically Charting Symptoms From Patient-Physician Conversations Using Machine Learning. JAMA internal medicine, 179(6), 836–838. https://doi.org/10.1001/jamainternmed.2018.8558
- Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature medicine, 28(1), 31–38. https://doi.org/10.1038/s41591-021-01614-0
- Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. https://arxiv.org/abs/1711.05225
- Ramirez, C. G. B., & McCauley, J. (Ur.). (2018). Contemporary Kidney Transplantation. Springer Cham.

- Ramsey, P. (2002). The patient as person: Explorations in medical ethics. Yale University Press.
- Rao, V. M., Hla, M., Moor, M., Adithan, S., Kwak, S., Topol, E. J., & Rajpurkar, P. (2025). Multimodal generative AI for medical image interpretation. Nature, 639(8056), 888–896. https://doi.org/10.1038/s41586-025-08675-y
- Rapaport, W. J. (2023). Philosophy of computer science. Cambridge University Press.
- Ratti, E., & Graves, M. (2022). Explainable machine learning practices: Opening another black box for reliable medical AI. AI and Ethics, 2(4), 801–814. https://doi.org/10.1007/s43681-022-00141-z
- Rawls, J. (1999). A theory of justice. Belknap Press.
- Rebollo Mateos, M. I., & López Castro, J. (2025). Ethical dilemmas of living donor kidney transplantation: A systematic review. Nefrología. Advance online publication. https://doi.org/10.1016/j.nefro.2025.501398
- Reinhardt, K. (2023). Trust and trustworthiness in AI ethics. AI Ethics, 3(4), 735–744. https://doi.org/10.1007/s43681-022-00200-5
- Reis Pereira, P., Almeida, M., Ribeiro, B., Oliveira, J., Costa, L., Pedroso, S., Martins, S., Dias, L., & Malheiro, J. (2022). Biologically Unrelated Living Donor for Kidney Transplantation Associated With Higher Acute Rejection Rate but With Similar Graft Survival of a Related Living Donor. Cureus, 14(10), e30189. https://doi.org/10.7759/cureus.30189
- Renftle, M., Trittenbach, H., Poznic, M., Heil, R. (2024). What do algorithms explain? The issue of the goals and capabilities of Explainable Artificial Intelligence (XAI). Humanities and Social Sciences Communications, 11, Article 760. https://doi.org/10.1057/s41599-024-03277-x
- Resnik, D. B., & Hosseini, M. (2025). The ethics of using artificial intelligence in scientific research: New guidance needed for a new tool. AI Ethics, 5, 1499–1521. https://doi.org/10.1007/s43681-024-00493-8
- Reuter, M., Kirchhoff, B. M., Franke, T., Radüntz, T., & Peifer, C. (2025). To trust or not to trust a human(-like) AI—A scoping review and conjoint analyses on factors influencing anthropomorphism and trust. Zeitschrift für Arbeitswissenschaft, 79, 402-432. https://doi.org/10.1007/s41449-025-00481-6
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD

International Conference on Knowledge Discovery and Data Mining (1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

- Rich, E., (1983). Artificial Intelligence. New York: McGraw-Hill
- Richmond, K. M., Muddamsetty, S. M., Gammeltoft-Hansen, T., Olsen, H. P., & Moeslund, T. B. (2024). Explainable AI and law: An evidential survey. Digital Society, 3(1). https://doi.org/10.1007/s44206-023-00081-z
- Riedl, R., Hogeterp, S. A., & Reuter, M. (2024). Do patients prefer a human doctor, artificial intelligence, or a blend, and is this preference dependent on medical discipline? Empirical evidence and implications for medical practice. Frontiers in psychology, 15, 1422177. https://doi.org/10.3389/fpsyg.2024.1422177
- Rittenberg, B. S. P., Holland, C. W., Barnhart, G. E., Gaudreau, S. M., & Neyedli, H. F. (2024). Trust with increasing and decreasing reliability. Human factors, 66(12), 2569–2589. https://doi.org/10.1177/00187208241228636
- Roberts, S. M. (2014). Bioethics in practice - a quarterly column about medical ethics: ethics and end-stage renal disease. Ochsner journal, 14(2), 154.
- Roberts, J. P., Wolfe, R. A., Bragg-Gresham, J. L., Rush, S. H., Wynn, J. J., Distant, D. A., Ashby, V. B., Held, P. J., & Port, F. K. (2004). Effect of changing the priority for HLA matching on the rates and outcomes of kidney transplantation in minority groups. The New England journal of medicine, 350(6), 545–551. https://doi.org/10.1056/NEJMoa025056
- Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A. I., Etmann, C., McCague, C., Beer, L., Weir-McCall, J. R., Teng, Z., Gkrania-Klotsas, E., AIX-COVNET, Rudd, J. H. F., Sala, E., & Schönlieb, C.-B. (2021). Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence, 3(3), 199–217. https://doi.org/10.1038/s42256-021-00307-0
- Robinson, D. G. (2022). Voices in the Code: A Story About People, Their Values, and the Algorithm They Made. Russell Sage Foundation.
- Rochlin, P. G. I. (1997). Trapped in the net: The Unanticipated Consequences of Computerization.
- Rokeach, M. (1973). The nature of human values. Free Press.
- Romanes, G. J. (1882). Animal intelligence. D. Appleton & Company.

- Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. IEEE Access, 8, 42200–42216. https://doi.org/10.1109/ACCESS.2020.2976199
- Rosenbacke, R., Melhus, Å., & Stuckler, D. (2024). False conflict and false confirmation errors are crucial components of AI accuracy in medical decision making. Nature communications, 15(1), 6896. https://doi.org/10.1038/s41467-024-50952-3
- Ross, L., Woodward, J. (2023). Causal approaches to scientific explanation. U E. N. Zalta & U. Nodelman (Ur.), The Stanford encyclopedia of philosophy. Stanford University. https://plato.stanford.edu/archives/spr2023/entries/causal-explanation-science/, pristup: 2. prosinca 2025.
- Roth, S., Baur, L., Boer, D., & Kramer, S. (2025). Enhancing symbolic machine learning by subsymbolic representations. arXiv. https://arxiv.org/abs/2506.14569
- Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. Journal of Personality, 35(4), 651–665. https://doi.org/10.1111/j.1467-6494.1967.tb01454.x
- Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Introduction to special topic forum: Not so different after all: A cross-discipline view of trust. Academy of Management Review, 23(3), 393–404. https://doi.org/10.5465/amr.1998.926617
- Rubin, D. L. (2019). Artificial intelligence in imaging: The radiologist's role. Journal of the American College of Radiology, 16, 1309–1317. https://doi.org/10.1016/j.jacr.2019.05.036
- Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nature machine intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x
- Rudner, R. (1953). The scientist qua scientist makes value judgments. Philosophy of Science, 20(1), 1–6. https://www.jstor.org/stable/185617
- Russell, S. (2022). Artificial intelligence and the problem of control. In H. Werthner, E. Prem, E. A. Lee, & C. Ghezzi (Ur.), Perspectives on digital humanism (25–36). Springer. https://doi.org/10.1007/978-3-030-86144-5_3
- Russell, S. J. (2019). Human compatible: Artificial intelligence and the problem of control. Penguin Random House.
- Russell, S. J., & Norvig, P. (2021). Artificial intelligence: A modern approach (4. izd.). Pearson.

- Russo, F., Schliesser, E., & Wagemans, J. (2024). Connecting ethics and epistemology of AI. AI & Society, 39, 1585–1603. https://doi.org/10.1007/s00146-022-01617-6
- Ryan, M. (2020). In AI we trust: Ethics, artificial intelligence, and reliability. Science and Engineering Ethics, 26(5), 2749-2767. https://doi.org/10.1007/s11948-020-00228-y
- Ryo, M. (2022). Explainable artificial intelligence and interpretable machine learning for agricultural data analysis (SSRN Electronic Journal). SSRN. https://doi.org/10.2139/ssrn.4230728
- Sagona, M., Dai, T., Macis, M., & Darden, M. (2025). Trust in AI-assisted health systems and AI's trust in humans. npj Health Systems, 2(10). https://doi.org/10.1038/s44401-025-00016-5
- Saidi, R. F., & Hejazii Kenari, S. K. (2014). Challenges of organ shortage for transplantation: solutions and opportunities. International journal of organ transplantation medicine, 5(3), 87–96.
- Salaün, A., Knight, S., Wingfield, L., Zhu, T., (2024). Predicting graft and patient outcomes following kidney transplantation using interpretable machine learning models. Scientific Reports, 14, 17356. https://doi.org/10.1038/s41598-024-66976-0
- Salzburg Global Seminar (2011). Salzburg statement on shared decision making. BMJ (Clinical research), 342, d1745. https://doi.org/10.1136/bmj.d1745
- Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. Philosophy & Technology, 34(4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x
- Saranya, A., & Subhashini, R. (2023). A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. Decision Analytics Journal, 7, 100230. https://doi.org/10.1016/j.dajour.2023.100230
- Sarvamangala, D. R., & Kulkarni, R. V. (2022). Convolutional neural networks in medical image understanding: a survey. Evolutionary intelligence, 15(1), 1–22. https://doi.org/10.1007/s12065-020-00540-3
- Sasseville, M., Ouellet, S., Rhéaume, C., Sahlia, M., Couture, V., Després, P., Paquette, J. S., Darmon, D., Bergeron, F., & Gagnon, M. P. (2025). Bias mitigation in primary health care artificial intelligence models: Scoping review. Journal of Medical Internet Research, 27, e60269. https://doi.org/10.2196/60269

- Savage, N. (2022). Breaking into the black box of artificial intelligence. Nature. https://doi.org/10.1038/d41586-022-00858-1
- Schidlt, H. (1987). Artificial Intelligence Using C. New York: McGraw-Hill.
- Schiff, D., & Borenstein, J. (2019). How Should Clinicians Communicate With Patients About the Roles of Artificially Intelligent Team Members?. AMA journal of ethics, 21(2), E138–E145. https://doi.org/10.1001/amajethics.2019.138
- Schloendorff v. Society of New York Hospital, 105 N.E. 92 (N.Y. 1914). https://cite.case.law/ne/105/92/
- Schneeberger, D., Stöger, K., & Holzinger, A. (2020). The European legal framework for medical AI. In C. Gütl, M. Hüllermeier, A. Holzinger, & K. Kunze (Ur.), Machine learning and knowledge extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 international cross-domain conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, proceedings (209–226). Springer. https://doi.org/10.1007/978-3-030-57321-8_12
- Schouten, D., Nicoletti, G., Dille, B., Chia, C., Vendittelli, P., Schuurmans, M., Litjens, G., & Khalili, N. (2025). Navigating the landscape of multimodal AI in medicine: A scoping review on technical challenges and clinical applications. Medical image analysis, 105, 103621. https://doi.org/10.1016/j.media.2025.103621
- Schuler, K., Jung, I.-C., Zerlik, M., Hahn, W., Sedlmayr, M., & Sedlmayr, B. (2025). Context factors in clinical decision-making: A scoping review. BMC Medical Informatics and Decision Making, 25, 133. https://doi.org/10.1186/s12911-025-02965-1
- Schur, E., Brouns, A., Lee, P. (2025). Ethical analysis of the responsibility gap in artificial intelligence. International Journal of Ethics and Society, 6(4), 1–10. http://ijethics.com/article-1-356-en.html
- Schuster, N., & Kilov, D. (2025). Moral disagreement and the limits of AI value alignment: A dual challenge of epistemic justification and political legitimacy. AI & Society. https://doi.org/10.1007/s00146-025-02427-2
- Schütze, P. (2024). The impacts of AI futurism: An unfiltered look at AI's true effects on the climate crisis. Ethics and Information Technology, 26(23). https://doi.org/10.1007/s10676-024-09758-6
- Schwalbe, G., & Finzel, B. (2024). A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. Data

Mining and Knowledge Discovery, 38(1), 3043–3101. https://doi.org/10.1007/s10618-022-00867-8

- Schwerzmann, K., & Campolo, A. (2025). “Desired behaviors”: Alignment and the emergence of a machine learning ethics. AI & Society, 40, 5181–5194. https://doi.org/10.1007/s00146-025-02272-3
- Scott, I. A., Doust, J. A., Keijzers, G. B., & Wallis, K. A. (2023). Coping with uncertainty in clinical practice: A narrative review. Medical Journal of Australia, 218(8), 418–425. https://doi.org/10.5694/mja2.51925
- Scriven, M. (1974). The exact role of value judgments in science. In K. F. Schaffner & R. S. Cohen (Ur.), PSA 1972 (219–247). D. Reidel Publishing Company. https://doi.org/10.1007/978-94-010-2140-1_15
- Searle, J. (2010). Making the social world: The structure of human civilization. Oxford University Press.
- Sepinwall, A. J. (2024). Artificial moral agents: Corporations and AI. In S. Hormio & B. Wringe (Ur.), Collective responsibility, 27-48. Springer. https://doi.org/10.1007/978-3-031-68718-1_3
- Serafimova, S. (2020). Whose morality? Which rationality? Challenging artificial intelligence as a remedy for the lack of moral enhancement. Humanities and Social Sciences Communications, 7(119). https://doi.org/10.1057/s41599-020-00614-8
- Serramia, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J. A., Rodriguez, M., Wooldridge, M., Morales, J., & Ansotegui, C. (2018). Moral values in norm decision making. In Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS ’18) (1294–1302). International Foundation for Autonomous Agents and Multiagent Systems.
- Serramia, M., Rodriguez-Soto, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J. A., Morales, J., Wooldridge, M., & Ansotegui, C. (2023). Encoding ethics to compute value-aligned norms. Minds & Machines, 33, 761–790. https://doi.org/10.1007/s11023-023-09649-7
- Sezgin, E. (2023). Artificial intelligence in healthcare: Complementing, not replacing, doctors and healthcare providers. Digital health, 9, 20552076231186520. https://doi.org/10.1177/20552076231186520

- Shah, P., Thornton, I., Kopitnik, N. L., & Hipskind, J. E. (2024). Informed consent. In StatPearls. StatPearls Publishing. pristup: 1. listopada 2025. https://www.ncbi.nlm.nih.gov/books/NBK430827/
- Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27(3), 379–423.
- Sharkey, A. (2020). Can we program or train robots to be good? Ethics and Information Technology, 22(3), 283–295. https://doi.org/10.1007/s10676-017-9425-5
- Sharma, C., Sharma, S., Sharma, K. P., Kumar, V., & Kumar, R. (2024). Exploring explainable AI: A bibliometric analysis. Discover Applied Sciences, 6, 615. https://doi.org/10.1007/s42452-024-06324-z
- Shatz, D. (1986). Autonomy, Beneficence, and Informed Consent: Rethinking the Connections. Cancer Investigation, 4(3), 257–269. https://doi.org/10.3109/07357908609018455
- Shavers, V. L., Lynch, C. F., & Burmeister, L. F. (2000). Knowledge of the Tuskegee study and its impact on the willingness to participate in medical research studies. Journal of the National Medical Association, 92(12), 563–572.
- Sheldon, K. M., & Kasser, T. (2001). Getting older, getting better? Personal strivings and personality maturity across the life span. Developmental Psychology, 37(4), 491–501. https://doi.org/10.1037/0012-1649.37.4.491
- Shermer, M., (2011). The Believing Brain. New York: Times Books Henry Holt And Company
- Shevtsova, D., Ahmed, A., Boot, I. W. A., Sanges, C., Hudecek, M., Jacobs, J. J. L., Hort, S., & Vrijhoef, H. J. M. (2024). Trust in and Acceptance of Artificial Intelligence Applications in Medicine: Mixed Methods Study. JMIR human factors, 11, e47031. https://doi.org/10.2196/47031
- Shew, A. (2017). Animal constructions and technological knowledge. Lexington Books.
- Shickel, B., Loftus, T. J., Adhikari, L., Ozrazgat-Baslanti, T., Bihorac, A., & Rashidi, P. (2019). DeepSOFA: A Continuous Acuity Score for Critically Ill Patients using Clinically Interpretable Deep Learning. Scientific reports, 9(1), 1879. https://doi.org/10.1038/s41598-019-38491-0
- Shin, D. (2025). Automating epistemology: How AI reconfigures truth, authority, and verification. AI & Society. https://doi.org/10.1007/s00146-025-02560-y

- Shmatko, A., Ghaffari Laleh, N., Gerstung, M., & Kather, J. N. (2022). Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. Nature cancer, 3(9), 1026–1038. https://doi.org/10.1038/s43018-022-00436-4
- Shrestha, P., Van Pilsum Rasmussen, S. E., King, E. A., Gordon, E. J., Faden, R. R., Segev, D. L., Humbyrd, C. J., & McAdams-DeMarco, M. (2022). Defining the ethical considerations surrounding kidney transplantation for frail and cognitively impaired patients: a Delphi study of geriatric transplant experts. BMC geriatrics, 22(1), 566. https://doi.org/10.1186/s12877-022-03209-x
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T. P., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484–489. https://doi.org/10.1038/nature16961
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., Van Den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. Nature, 550(7676), 354–359. https://doi.org/10.1038/nature24270
- Simmel, G. (1908). The Sociology of Georg Simmel (K. Wolff, Trans.). New York: Free Press.
- Simon, B. D., Ozyoruk, K. B., Gelikman, D. G., Harmon, S. A., & Türkbey, B. (2025). The future of multimodal artificial intelligence models for integrating imaging and clinical metadata: a narrative review. Diagnostic and interventional radiology (Ankara, Turkey), 31(4), 303–312. https://doi.org/10.4274/dir.2024.242631
- Simon, H. A. (1960). The Corporation: Will It Be Managed by Machines? In M. Anshen & G. L. Bach (Ur.), Management and Corporations (str. xx-xx). New York: McGraw-Hill.
- Simon, H. A. (1992). What is an "explanation" of behavior? Psychological Science, 3(3), 150–161. https://doi.org/10.1111/j.1467-9280.1992.tb00017.x
- Simonite, T., Google's AI guru wants computers to think more like brains. (2018). https://www.wired.com/story/googles-ai-guru-computers-think-more-like-brains/, pristup: 10. ožujka 2025.

- Singh, G., Stefenon, S. F., & Yow, K. C. (2025). The shallowest transparent and interpretable deep neural network for image recognition. Scientific Reports, 15, 13940. https://doi.org/10.1038/s41598-025-92945-2
- Sipser, M. (2012). Introduction to the theory of computation (3. izd.). Cengage Learning.
- Sison, A. J. G., & Redín, D. M. (2023). A neo-Aristotelian perspective on the need for artificial moral agents (AMAs). AI & Society, 38(1), 47–65. https://doi.org/10.1007/s00146-021-01283-0
- Siwek, Z., Oleszkowicz, A., & Słowińska, A. (2017). Values realized in personal strivings and motivation, and meaning in life in Polish university students. Journal of Happiness Studies, 18, 549–573. https://doi.org/10.1007/s10902-016-9737-x
- Skansi, S. (2018). From Logic to Cognitive Science. In: Introduction to Deep Learning. Undergraduate Topics in Computer Science. Springer, Cham. https://doi.org/10.1007/978-3-319-73004-2_1
- Skulmowski, A. (2025). Informed consent in educational AI research needs to be transparent, flexible, and dynamic. Minds and Machines. Advance online publication. https://doi.org/10.1111/mbe.12435
- Smith, H. (2021). Clinical AI: Opacity, accountability, responsibility and liability. AI & Society, 36(2), 535–545. https://doi.org/10.1007/s00146-020-01019-6
- Smith, V. L. (1998). The two faces of Adam Smith. Southern Economic Journal, 65(1), 1–19. https://doi.org/10.1002/j.2325-8012.1998.tb00125.x
- Smolenski, J. (2024). The foundations of informed consent and bodily self-sovereignty: A positive suggestion. Monash Bioethics Review, 42(2), 115–136. https://doi.org/10.1007/s40592-024-00203-4
- Sofge, E. (2015). Bill Gates fears A.I., but A.I. researchers know better. Popular Science. https://www.popsci.com/bill-gates-fears-ai-ai-researchers-know-better/, pristup: 26. rujna 2025.
- Søgaard, A. (2023). On the Opacity of Deep Neural Networks. Canadian Journal of Philosophy, 53(3), 224–239. doi:10.1017/can.2024.1
- Søgaard, A. (2025). Can machines be trustworthy? AI and Ethics, 5(3), 313–321. https://doi.org/10.1007/s43681-023-00351-z

- Solanki, P., Grundy, J., & Hussain, W. (2023). Operationalising ethics in artificial intelligence for healthcare: A framework for AI developers. AI Ethics, 3(2), 223–240. https://doi.org/10.1007/s43681-022-00195-z
- Soltanzadeh, S. (2025). A metaphysical account of agency for technology governance. AI & Society, 40(6), 1723–1734. https://doi.org/10.1007/s00146-024-01941-z
- Sotala, K. (2016). Defining human values for value learners. In Proceedings of the Workshops of the 30th AAAI Conference on Artificial Intelligence: AI, Ethics, and Society (113–123). AAAI Press.
- Sparrow, R. (2007). Killer robots. Journal of Applied Philosophy, 24(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x
- Sparrow, R., & Flenady, G. (2025). The testimony gap: Machines and reasons. Minds & Machines, 35, 12. https://doi.org/10.1007/s11023-025-09712-5
- Spatz, E. S., Krumholz, H. M., & Moulton, B. W. (2016). The New Era of Informed Consent: Getting to a Reasonable-Patient Standard Through Shared Decision Making. JAMA, 315(19), 2063–2064. https://doi.org/10.1001/jama.2016.3070
- Speith, T. (2022). A review of taxonomies of explainable artificial intelligence (XAI) methods. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22) (2239–2250). Association for Computing Machinery. https://doi.org/10.1145/3531146.3534639
- Speith, T., Crook, B., Mann, S., Schomäcker, A., & Langer, M. (2024). Conceptualizing understanding in explainable artificial intelligence (XAI): An abilities-based approach. Ethics and Information Technology, 26, 40. https://doi.org/10.1007/s10676-024-09769-3
- Starke, G., & Ienca, M. (2024). Misplaced Trust and Distrust: How Not to Engage with Medical Artificial Intelligence. Cambridge Quarterly of Healthcare Ethics, 33(3), 360–369. doi:10.1017/S0963180122000445
- Staugaard, A. C., (1987). Robotics and AI: An Introduction to Applied Machine Intelligence, Englewood Cliffs, NJ: Prentice Hall.
- Sterlin, E., (2024), World Economic Forum, https://www.weforum.org/stories/2024/08/healthcare-costs-digital-tech/, pristup: 1. travnja 2025.

- Stilgoe, J. (2018). Machine learning, social learning and the governance of self-driving cars. Social Studies of Science, 48(1), 25–56. https://doi.org/10.1177/0306312717741687
- Stöger, K., & Schneeberger, D. (2021). Medical artificial intelligence: The European legal perspective. Communications of the ACM, 64(11), 28–30. https://doi.org/10.1145/3458652
- Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650. https://doi.org/10.18653/v1/P19-1355
- Sudore, R. L., Landefeld, C. S., Williams, B. A., Barnes, D. E., Lindquist, K., & Schillinger, D. (2006). Use of a modified informed consent process among vulnerable patients: a descriptive study. Journal of general internal medicine, 21(8), 867–873. https://doi.org/10.1111/j.1525-1497.2006.00535.x
- Suh, A., Hurley, I., Smith, N., & Siu, H. C. (2025). Fewer than 1% of explainable AI papers validate explainability with humans. In Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25) (1–7). Association for Computing Machinery. https://doi.org/10.1145/3706599.3719964
- Sullins, J. P. III. (2006). When is a robot a moral agent? The International Review of Information Ethics, 6, 23–30. https://doi.org/10.29173/irie136
- Sunyaev, A., Benlian, A., Pfeiffer, J., Jussupow, E., Thiebes, S., Maedche, A., & Gawlitza, J. (2025). High-risk artificial intelligence. Business & Information Systems Engineering. https://doi.org/10.1007/s12599-025-00942-6
- Suresh, H., Gomez, S. R., Nam, K. K., & Satyanarayan, A. (2021). Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (1–16). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445088
- Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2. izd.). The MIT Press.
- Swartout, W. R. (1981). Explaining and justifying expert consulting programs. In Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI-81), Vol. 2, (815–823). Vancouver, Canada: Morgan Kaufmann

- Sylvestre, P., Orr Gaucher, N., Perez, T., & Drouin, O. (2021). Rethinking informed consent in the age of behavioural sciences and relational autonomy. Ethics, Medicine and Public Health, 19, 100708. https://doi.org/10.1016/j.jemep.2021.100708
- Sze, V., Chen, Y., Yang, T., & Emer, J. S. (2017). Efficient Processing of deep Neural Networks: A tutorial and survey. Proceedings of the IEEE, 105(12), 2295–2329. https://doi.org/10.1109/jproc.2017.2761740
- Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. Research in Science Education, 48(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2
- Taleb, N. N. (2010). The Black Swan. New York, NY: Random House.
- Tam, N. T., Huy, N. T., Thoa, leT. B., Long, N. P., Trang, N. T., Hirayama, K., & Karbwang, J. (2015). Participants' understanding of informed consent in clinical trials over three decades: systematic review and meta-analysis. Bulletin of the World Health Organization, 93(3), 186–98H. https://doi.org/10.2471/BLT.14.141390
- Tamburrini, G. (2023). Nuclear weapons and the militarization of AI. In P. Cotta-Ramusino, M. Lowenthal, L. Maiani, & E. Pellecchia (Ur.), Nuclear risks and arms control - Problems and progresses in the time of pandemics and war (EAC 2022, Springer Proceedings in Physics, Vol. 291). Springer. https://doi.org/10.1007/978-3-031-29708-3_17
- Tang, X., Jin, Q., Zhu, K., Yuan, T., Zhang, Y., Zhou, W., Qu, M., Zhao, Y., Tang, J., Zhang, Z., Cohan, A., Greenbaum, D., Lu, Z., & Gerstein, M. (2025). Risks of AI scientists: Prioritizing safeguarding over autonomy. Nature Communications, 16, 8317. https://doi.org/10.1038/s41467-025-63913-1
- Tanida, T., Müller, P., Kaissis, G., & Rueckert, D. (2023). Interactive and Explainable Region-guided Radiology Report Generation. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7433-7442.
- Tantisattamo, E., & Maggiore, U. (2024). Revisiting pre-transplant preparation to optimize long-term kidney transplant outcomes. Journal of Nephrology, 37, 1425–1428. https://doi.org/10.1007/s40620-024-02108-1
- Tartaro, A. (2025). When things go wrong: The recall of AI systems as a last resort for ethical and lawful AI. AI Ethics, 5, 253–262. https://doi.org/10.1007/s43681-023-00327-z

- Taylor, H. M., & Karlin, S. (1998). An introduction to stochastic modeling. Gulf Professional Publishing.
- Taylor, I. (2025). Is explainable AI responsible AI? AI & Society, 40(3), 1695–1704. https://doi.org/10.1007/s00146-024-01939-7
- Tenner, E. (1997). Why things bite back : technology and the revenge of unintended consequences. Vintage Books.
- Terziyan, V., Tiihonen, T., Shukla, A., Gaiduk, M., & Golivets, I. (2025). Towards ethical evolution: Responsible autonomy of artificial intelligence across generations. AI and Ethics, 5, 5165–5190. https://doi.org/10.1007/s43681-025-00759-9
- Thampi, A. (2022). Interpretable AI: Building explainable machine learning systems. Manning Publications.
- The New York Times, (1958), New navy device learns by doing, https://www.nytimes.com/1958/07/08/archives/new-navy-device-learns-by-doing-psychologist-shows-embryo-of.html, pristup: 1. travnja 2025.
- The White House. (2023). Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. The White House. https://bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/, pristup: 2. listopada 2025.
- Theunissen, M., & Browning, J. (2022). Putting explainable AI in context: Institutional explanations for medical AI. Ethics and Information Technology, 24(1), 23. https://doi.org/10.1007/s10676-022-09649-8
- Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. Electronic Markets, 31(2), 447–464. https://doi.org/10.1007/s12525-020-00441-4
- Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F. (2022). The computational limits of deep learning (arXiv:2007.05558). arXiv. https://doi.org/10.48550/arXiv.2007.05558
- Toapanta, N., Comas, J., Revuelta, I., Manonelles, A., Facundo, C., Pérez-Saez, M. J., Vila, A., Arcos, E., Tort, J., Giral, M., Naesens, M., Kuypers, D., Asberg, A., Moreso, F., Bestard, O., & EKITE consortium (2024). Benefits of Living Over Deceased Donor Kidney Transplantation in Elderly Recipients. A Propensity Score Matched Analysis of a Large European Registry Cohort. Transplant international : official

journal of the European Society for Organ Transplantation, 37, 13452. https://doi.org/10.3389/ti.2024.13452

- Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G., & Kaplan, L. (2020). Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI. Patterns (New York, N.Y.), 1(4), 100049. https://doi.org/10.1016/j.patter.2020.100049
- Tong, A., Jan, S., Wong, G., Craig, J. C., Irving, M., Chadban, S., Cass, A., Marren, N., & Howard, K. (2012). Patient preferences for the allocation of deceased donor kidneys for transplantation: a mixed methods study. BMC nephrology, 13, 18. https://doi.org/10.1186/1471-2369-13-18
- Topol, E. J. (2019A). Deep medicine: How artificial intelligence can make healthcare human again. Basic Books.
- Topol E. J. (2019B). High-performance medicine: the convergence of human and artificial intelligence. Nature medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7
- Trench, B. (2025). Trust and mistrust in science: Beyond the binary. In A. Fage-Butler, L. Ledderer, & K. H. Nielsen (Ur.), Science communication and trust (323–343). Palgrave Macmillan. https://doi.org/10.1007/978-981-96-1289-5_16
- Tuncdogan, A., Acar, O. A., & Car, J. (2025). Reframing the AI doctor debate: The comparison should not be with doctors but doctorlessness. Discover Health Systems, 4(1), 78. https://doi.org/10.1007/s44250-025-00264-0
- Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., Potapenko, A., Ballard, A. J., Romera-Paredes, B., Nikolov, S., Jain, R., Clancy, E., Reiman, D., Petersen, S., Senior, A. W., Kavukcuoglu, K., Birney, E., Kohli, P., Jumper, J., Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. Nature, 596(7873), 590–596. https://doi.org/10.1038/s41586-021-03828-1
- Turillazzi, E., & Neri, M. (2015). Informed consent and Italian physicians: Change course or abandon ship—from formal authorization to a culture of sharing. Medicine, Health Care and Philosophy, 18(3), 449–453. https://doi.org/10.1007/s11019-015-9637-6

- Turilli, M. (2007). Ethical protocols design. Ethics and Information Technology, 9(1), 49–62. https://doi.org/10.1007/s10676-006-9128-9
- Turing, A. M. (2004). Intelligent machinery: A heretical theory. U B. J. Copeland (Ur.), The essential Turing: Seminal writings in computing, logic, philosophy, artificial intelligence, and artificial life, plus the secrets of Enigma (469-485). Oxford University Press.
- Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. Proceedings of the London Mathematical Society, 2(42), 230–265. https://doi.org/10.1112/plms/s2-42.1.230
- Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59, 433–460. https://doi.org/10.1093/mind/LIX.236.433
- Turing, A. M. (1951). Can digital computers think? In Copeland, B. J., Ur., The Essential Turing, pages 476–486. Oxford University Press (2004), Oxford.
- Upadhyay, S., Lakkaraju, H., & Gajos, K. Z. (2025). Counterfactual explanations may not be the best algorithmic recourse approach. In Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25) (446–462). Association for Computing Machinery. https://doi.org/10.1145/3708359.3712095
- Ursin, F., Timmermann, C., Orzechowski, M., & Steger, F. (2021). Diagnosing Diabetic Retinopathy With Artificial Intelligence: What Information Should Be Included to Ensure Ethical Informed Consent?. Frontiers in medicine, 8, 695217. https://doi.org/10.3389/fmed.2021.695217
- Uslaner, E. M. (2018). The study of trust. U E. M. Uslaner (Ur.), The Oxford handbook of social and political trust. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190274801.013.39
- Uwechue, R., Chandak, P., Ahmed, Z., Gogalniceanu, P., Kessaris, N., & Mamode, N. (2017). Minimally invasive surgical techniques for kidney transplantation. The Cochrane Database of Systematic Reviews, 2017(7), CD012698. https://doi.org/10.1002/14651858.CD012698
- Vallor, S. (Ur.). (2020). The Oxford Handbook of Philosophy of Technology. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190851187.001.0001
- van Assen, M., Beecy, A., Gershon, G., Newsome, J., Trivedi, H., & Gichoya, J. (2024). Implications of bias in artificial intelligence: Considerations for

cardiovascular imaging. Current Atherosclerosis Reports, 26(4), 91–102. https://doi.org/10.1007/s11883-024-01190-x

- Van Cauwenberge, D., Van Biesen, W., Decruyenaere, J., Leune, T., & Sterckx, S. (2022). "Many roads lead to Rome and the Artificial Intelligence only shows me one road": An interview study on physician attitudes regarding the implementation of computerised clinical decision support systems. BMC Medical Ethics, 23, 50. https://doi.org/10.1186/s12910-022-00787-8
- van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. Minds & Machines, 30(3), 385–409. https://doi.org/10.1007/s11023-020-09537-4
- van der Schee, M., Pinheiro, H., & Gaude, E. (2018). Breath biopsy for early detection and precision medicine in cancer. Ecancermedicalscience, 12, ed84. https://doi.org/10.3332/ecancer.2018.ed84
- Van Fraassen, B. C. (1973). Values and the Heart's Command. The Journal of Philosophy, 70(1), 5–19. https://doi.org/10.2307/2024762
- van Hemmen J. L. (2021). Mathematization of nature: how it is done. Biological cybernetics, 115(6), 655–664. https://doi.org/10.1007/s00422-021-00914-5
- van Lent, M., Fisher, W., & Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. In Proceedings of the 2004 National Conference on Artificial Intelligence, 900–907
- van Voorst, R. (2023). Challenges and limitations of human oversight in ethical artificial intelligence implementation in health care: Balancing digital literacy and professional strain. Mayo Clinic Proceedings: Digital Health, 2(4), 559–563. https://doi.org/10.1016/j.mcpdig.2023.07.002
- van Wynsberghe, A., & Robbins, S. (2019). Critiquing the reasons for making artificial moral agents. Science and Engineering Ethics, 25(3), 719–735. https://doi.org/10.1007/s11948-018-0030-8
- Varda, C., Iordanou, K., Antoniou, J., Barrios, M. M. Z., Yalaz, E., Gurzawska, A., Szüdi, G., Bartar, P., & Häberlein, L. (2025). The role of stewards of trust in facilitating trust in science: A multistakeholder view. Journal of Academic Ethics, 23(3), 463–483. https://doi.org/10.1007/s10805-024-09583-y
- Varkey, B. (2021). Principles of Clinical Ethics and Their Application to Practice. Medical principles and practice : international journal of the Kuwait University, Health Science Centre, 30(1), 17–28. https://doi.org/10.1159/000509119

- Vaughn, L. (2020). Paternalism and patient autonomy. In Bioethics: Principles, issues, and cases (4. izd.), (115–140). Oxford University Press.
- Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. PLoS medicine, 15(11), e1002689. https://doi.org/10.1371/journal.pmed.1002689
- Veatch, R. M. (1973). Generalization of Expertise. The Hastings Center Studies, 1(2), 29–40. https://doi.org/10.2307/3527511
- Véliz, C. (2021). Moral zombies: Why algorithms are not moral agents. AI & Society, 36, 487–497. https://doi.org/10.1007/s00146-021-01189-x
- Veluwenkamp, H., & Hindriks, F. (2024). Artificial agents: responsibility & control gaps. Inquiry, 1–25. https://doi.org/10.1080/0020174X.2024.2410995
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., … Zhu, X. (2001). The sequence of the human genome. Science (New York, N.Y.), 291(5507), 1304–1351. https://doi.org/10.1126/science.1058040
- Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion, 76, 89–106. https://doi.org/10.1016/j.inffus.2021.05.009
- Vinge, V., (1993). The coming technological singularity: How to survive in the post-human era. Whole Earth Review.
- Visoka stručna skupina za umjetnu inteligenciju Europske komisije. (2019). Smjernice etike za pouzdanu umjetnu inteligenciju. Europski parlament. https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/JURI/DV/2019/11-06/Ethics-guidelines-AI_HR.pdf
- Vokinger, K. N., Feuerriegel, S., & Kesselheim, A. S. (2021). Mitigating bias in machine learning for medicine. Communications medicine, 1, 25. https://doi.org/10.1038/s43856-021-00028-w
- von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. Philosophy & Technology, 34, 1607–1622. https://doi.org/10.1007/s13347-021-00477-0

- Voorhoeve, A. (2018). Balancing small against large burdens. Behavioural Public Policy, 2(1), 125–142. doi:10.1017/bpp.2017.4
- Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. Science Robotics, 2(6), Article eaan6080. https://doi.org/10.1126/scirobotics.aan6080
- Wadden, J. J. (2021). What kind of artificial intelligence should we want for use in healthcare decision-making applications? Canadian Journal of Bioethics / Revue canadienne de bioéthique, 4(1), 94. https://doi.org/10.7202/1077636ar
- Wałdoch, K. (2024). Informed Consent for the Use of AI in the Process of Providing Medical Services. Review of European and Comparative Law, 57(2), 121–134. https://doi.org/10.31743/recl.17239
- Wall, L. L. (2006). The medical ethics of Dr J Marion Sims: a fresh look at the historical record. Journal of medical ethics, 32(6), 346–350. https://doi.org/10.1136/jme.2005.012559
- Wallach, W., & Allen, C. (2009). Moral machines: Teaching robots right from wrong. Oxford University Press.
- Wang, B., Rau, P. L. P., & Yuan, T. (2022). Measuring user competence in using artificial intelligence: validity and reliability of artificial intelligence literacy scale. Behaviour & Information Technology, 42(9), 1324–1337. https://doi.org/10.1080/0144929X.2022.2072768
- Wang, L. (2025). Can machines think beyond words? A critique of AI's meaning-production process. AI & Society. https://doi.org/10.1007/s00146-025-02485-6
- Wang, P. (1994). On the working definition of intelligence (Technical Report No. 94). Center for Research on Concepts and Cognition, Indiana University.
- Wang, P. (2019). What do you mean by "AI"? Journal of Artificial General Intelligence, 10(2), 1–37. https://doi.org/10.2478/jagi-2019-0002
- Wang, P., (1995). Non-Axiomatic Reasoning System: Exploring the Essence of Intelligence. Ph.D. Dissertation, Indiana University.
- Wartman, S. A., & Combs, C. D. (2018). Medical Education Must Move From the Information Age to the Age of Artificial Intelligence. Academic medicine : journal of the Association of American Medical Colleges, 93(8), 1107–1109. https://doi.org/10.1097/ACM.0000000000002044

- Waterman, D. A., (1986). A Guide to Expert Systems. Reading, MA: Addison-Wesley
- Watters, T. K., Glass, B. D., & Mallett, A. J. (2024). Identifying the barriers to kidney transplantation for patients in rural and remote areas: A scoping review. Journal of Nephrology, 37, 1435–1447. https://doi.org/10.1007/s40620-023-01755-0
- Weber, J. (1993). Exploring the relationship between personal values and moral reasoning. Human Relations, 46(4), 435–463. https://doi.org/10.1177/001872679304600402
- Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., Haas, J., Legassick, S., Irving, G., Gabriel, I. (2022). Taxonomy of Risks posed by Language Models. 2022 ACM Conference on Fairness, Accountability, and Transparency, 214–229. https://doi.org/10.1145/3531146.3533088
- Weintraub, M. (1985). A Theory and Computational Model of Auditory Monaural Sound Separation.
- Weiss, B. D., (2007). Health Literacy and Patient Safety: Help Patients Understand. Manual for Clinicians. (2. izd.), Weiss BD. Chicago, IL: American Medical Association Foundation
- Wenderott, K., Krups, J., Zaruchas, F., & Weigl, M. (2024). Effects of artificial intelligence implementation on efficiency in medical imaging-a systematic literature review and meta-analysis. NPJ digital medicine, 7(1), 265. https://doi.org/10.1038/s41746-024-01248-9
- Whitney, E. (1990). Paradise Restored. The Mechanical Arts from Antiquity through the Thirteenth Century. Transactions of the American Philosophical Society, 80(1), 1–169. https://doi.org/10.2307/1006521
- Wiegreffe, S., & Pinter, Y. (2019). Attention is not not explanation (arXiv:1908.04626v2). arXiv. https://arxiv.org/abs/1908.04626
- Wiener, N. (1960). Some moral and technical consequences of automation. Science, 131(3410), 1355–1358.
- Wijdicks, E., F., M. (2018). How Harvard Defined Irreversible Coma. Neurocritical care, 29(1), 136–141. https://doi.org/10.1007/s12028-018-0579-8

- Wildi, M., & Hadji Misheva, B. (2022). Financial risk management and explainable, trustworthy, responsible AI. Frontiers in Artificial Intelligence, 5, Article 779799. https://doi.org/10.3389/frai.2022.779799
- Winfield, A. F. T., Blum, C., & Liu, W. (2014). Towards an ethical robot: Internal models, consequences and ethical action selection. In M. Mistry, A. Leonardis, M. Witkowski, & C. Melhuish (Ur.), Advances in autonomous robotic systems: TAROS 2014, (85–96). Springer. https://doi.org/10.1007/978-3-319-10401-0_8
- Winkler, P. W., Zsidai, B., Hamrin Senorski, E., Pruneski, J. A., Hirschmann, M. T., Ley, C., Tischer, T., Herbst, E., Pareek, A., Musahl, V., Oeding, J. F., Oettl, F. C., Longo, U. G., Samuelsson, K., Feldt, R., & ESSKA Artificial Intelligence Working Group (2025). A practical guide to the implementation of AI in orthopaedic research-Part 7: Risks, limitations, safety and verification of medical AI systems. Journal of experimental orthopaedics, 12(2), e70247. https://doi.org/10.1002/jeo2.70247
- Wolkenstein, A. (2024). Healthy Mistrust: Medical Black Box Algorithms, Epistemic Authority, and Preemptionism. Cambridge quarterly of healthcare ethics : CQ : the international journal of healthcare ethics committees, 33(3), 370–379. https://doi.org/10.1017/S0963180123000646
- Wong, A., Otles, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, M., Konye, J., Penoza, C., Ghous, M., & Singh, K. (2021). External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. JAMA internal medicine, 181(8), 1065–1070. https://doi.org/10.1001/jamainternmed.2021.2626
- Woods, J., & Walton, D. (1982). The Petitio: Aristotle's Five Ways. Canadian Journal of Philosophy, 12(1), 77–100. doi:10.1080/00455091.1982.10715784
- Wulf, A. J., & Seizov, O. (2024). "Please understand we cannot provide further information": Evaluating content and transparency of GDPR-mandated AI disclosures. AI & Society, 39, 235–256. https://doi.org/10.1007/s00146-022-01424-z
- Xie, Y., Lu, L., Gao, F., He, S. J., Zhao, H. J., Fang, Y., Yang, J. M., An, Y., Ye, Z. W., & Dong, Z. (2021). Integration of Artificial Intelligence, Blockchain, and Wearable Technology for Chronic Disease Management: A New Paradigm in Smart Healthcare. Current medical science, 41(6), 1123–1133. https://doi.org/10.1007/s11596-021-2485-0

- Xu, D., Liu, F., Ding, X., Ma, J., Suo, Y., Peng, Y. Y., Li, J., & Fu, X. (2025). Exploring ICU nurses' response to alarm management and strategies for alleviating alarm fatigue: a meta-synthesis and systematic review. BMC nursing, 24(1), 412. https://doi.org/10.1186/s12912-025-03084-y
- Xu, H., & Shuttleworth, K. M. J. (2024). Medical artificial intelligence and the black box problem: A view based on the ethical principle of "do no harm". Intelligent Medicine, 4(1), 52–57. https://doi.org/10.1016/j.imed.2023.08.001
- Yacoub, R., Nadkarni, G. N., Cravedi, P., He, J. C., Delaney, V. B., Kent, R., Chauhan, K. N., Coca, S. G., Florman, S. S., Heeger, P. S., Murphy, B., & Menon, M. C. (2018). Analysis of OPTN/UNOS registry suggests the number of HLA matches and not mismatches is a stronger independent predictor of kidney transplant survival. Kidney international, 93(2), 482–490. https://doi.org/10.1016/j.kint.2017.07.016
- Yala, A., Schuster, T., Miles, R., Barzilay, R., & Lehman, C. (2019). A Deep Learning Model to Triage Screening Mammograms: A Simulation Study. Radiology, 293(1), 38–46. https://doi.org/10.1148/radiol.2019182908
- Yanagawa, M., & Sato, J. (2024). Seeing Is Not Always Believing: Discrepancies in Saliency Maps. Radiology. Artificial intelligence, 6(1), e230488. https://doi.org/10.1148/ryai.230488
- Yang, G., Ye, Q., & Xia, J. (2022). Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. An international journal on information fusion, 77, 29–52. https://doi.org/10.1016/j.inffus.2021.07.016
- Yao, L., Liu, Y., Wang, T., Han, C., Li, Q., Li, Q., You, X., Ren, T., Wang, Y. (2025). Global trends of big data analytics in health research: a bibliometric study. Frontiers in medicine, 12, 1456286. https://doi.org/10.3389/fmed.2025.1456286
- Yeung, M. Y., Coates, P. T., & Li, P. K. (2022). Kidney Organ Allocation System: How to Be Fair. Seminars in nephrology, 42(4), 151274. https://doi.org/10.1016/j.semnephrol.2022.09.002
- Yoshizawa, T., Ishida, S., Sato, T., Ohta, M., Honma, T., Terayama, K., & Kato, K. (2025). A data-driven generative strategy to avoid reward hacking in multi-objective molecular design. Nature Communications, 16(1), 2409. https://doi.org/10.1038/s41467-025-57582-3

- Yudkowsky, E. (2001). Creating Friendly AI 1.0: The analysis and design of benevolent goal architectures (Tech. Rep.). Machine Intelligence Research Institute. https://intelligence.org/files/CFAI.pdf
- Zadeh, L. A. (1973). Outline of a new approach to the analysis of complex systems and decision processes. IEEE Transactions on Systems, Man, and Cybernetics, SMC-3(1), 28–44. https://doi.org/10.1109/TSMC.1973.5408575
- Zafar, M. (2025). Normativity and AI moral agency. AI and Ethics, 5, 2605–2622. https://doi.org/10.1007/s43681-024-00566-8
- Zambeli-Ljepović, A., Tungsanga, S., Ghimire, A., Bello, A. K., Okpechi, I. G., Siyoum, M. T., Mushi, F. A., Asiimwe, F., Bagasha, P., Okungu, V., Bertozzi, S. M., Hoffmann, T. J., Thomson, D., Muller, E., Scott, J. W., Syed, S., Ascher, N., Stock, P., & Rose, J. (2025). The potential of kidney transplantation to reduce mortality from chronic kidney disease: A global, cross-sectional, modelling study. The Lancet Global Health, 13(10), e1691–e1700.
- Zanotti, G., Chiffi, D., & Schiaffonati, V. (2024). AI-related risk: An epistemological approach. Philosophy & Technology, 37, 66. https://doi.org/10.1007/s13347-024-00755-7
- Zeng, W., Zhu, H., Qin, C., Wu, H., Cheng, Y., Zhang, S., Jin, X., Shen, Y., Wang, Z., Zhong, F., & Xiong, H. (2025). Application-driven value alignment in agentic AI systems: Survey and perspectives. arXiv e-prints. https://doi.org/10.48550/arXiv.2506.09656
- Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence. Patterns (New York, N.Y.), 3(4), 100455. https://doi.org/10.1016/j.patter.2022.100455
- Zhang, H., Figueroa, F. T., & Hermanns, H. (2025). Saliency maps give a false sense of explainability to image classifiers: An empirical evaluation across methods and metrics. Proceedings of the 16th Asian Conference on Machine Learning, Proceedings of Machine Learning Research, 260, 479–494. https://proceedings.mlr.press/v260/zhang25a.html
- Zhang, J., & Zhang, Z. M. (2023). Ethics and governance of trustworthy medical artificial intelligence. BMC Medical Informatics and Decision Making, 23(1), 7. https://doi.org/10.1186/s12911-023-02103-9

- Zhang, M., & Ma, X. (2025). The subtle balance of trust: how employees' expected and perceived trust influence impression management. Frontiers in psychology, 16, 1526860. https://doi.org/10.3389/fpsyg.2025.1526860
- Zhang, X., Gu, Y., Yin, J., Zhang, Y., Jin, C., Wang, W., Li, A. M., Wang, Y., Su, L., Xu, H., Ge, X., Ye, C., Tang, L., Shen, B., Fang, J., Wang, D., & Feng, R. (2023). Development, Reliability, and Structural Validity of the Scale for Knowledge, Attitude, and Practice in Ethics Implementation Among AI Researchers: Cross-Sectional Study. JMIR formative research, 7, e42202. https://doi.org/10.2196/42202
- Zhou, S., Massie, A. B., Luo, X., Ruck, J. M., Chow, E. K. H., Bowring, M. G., Bae, S., Segev, D. L., & Gentry, S. E. (2018). Geographic disparity in kidney transplantation under KAS. American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons, 18(6), 1415–1423. https://doi.org/10.1111/ajt.14622

Životopis

Luka Poslon rođen je 2. srpnja 1992. godine u Bjelovaru. Trenutno je član istraživačke grupe Digit-HeaL laboratorija i asistent na Hrvatskom katoličkom sveučilištu u Zagrebu, zaposlen uz potporu Hrvatske zaklade za znanost na temelju natječaja „Projekt razvoja karijera mladih istraživača – izobrazba novih doktora znanosti“ (DOK-2021-02-9366). Diplomirao je filozofiju na Fakultetu filozofije i religijskih znanosti 2021. godine. Savjetnik je za mlade pri Centru za proučavanje odnosa znanosti i religije na Fakultetu filozofije i religijskih znanosti te zamjenik člana Savjeta za mlade Vlade Republike Hrvatske.

Sudjelovao je na brojnim međunarodnim konferencijama i projektima, uključujući višemjesečni studijski boravak u Stuttgartu. Znanstveno se bavi etičkim i epistemološkim izazovima primjene umjetne inteligencije (UI) u zdravstvenoj skrbi, s posebnim naglaskom na objašnjivu umjetnu inteligenciju (OUI). Analizira kako OUI može pomoći u razumijevanju i praćenju odluka koje donose sustavi UI, što je ključno za izgradnju povjerenja i suradnje između liječnika, pacijenata i UI. Autor je znanstvenih radova o ljudskom dostojanstvu i etičkim izazovima primjene UI-e u zdravstvu.

Njegovi znanstveno-istraživački interesi obuhvaćaju suvremene izazove bioetike, filozofije tehnologije i etike UI-e. U svom radu proučava međuodnos čovjeka i UI-e, osobito u kontekstu zdravstvene skrbi, te analizira kako tehnologije poput OUI utječu na odnos između pacijenta i liječnika te na donošenje odluka u zdravstvu.

Objavljeni radovi

Poslon, L., & Čartolovni, A. (2025). Improving explanation measures inside the decision-making process to increase the potential of AI applications in healthcare. U R. Ortale, G. Costa, & A. Forestiero (Ur.), Explainable machine intelligence in healthcare, 41–55. Springer. https://doi.org/10.1007/978-3-031-91379-2_3

Poslon, L. (2025). Osnaživanje povjerenja osiguravanjem pravednosti u sustavima umjetne inteligencije u zdravstvenoj skrbi. Nova prisutnost, XXIII (2), 277-294. https://doi.org/10.31192/np.23.2.4

Poslon, L. (2025). Jerko VALKOVIĆ, Jakov ŽIŽIĆ (ur.), Demokršćanstvo: izvori, postignuća i perspektive, Zagreb, Glas Koncila i Hrvatsko katoličko sveučilište, 2022, 239 str.. Nova prisutnost, XXIII (1), 215-219.

Čartolovni, A., Malešević, A., & Poslon, L. (2023). Critical analysis of the AI impact on the patient-physician relationship: A multi-stakeholder qualitative study. Digital health, 9, 20552076231220833. https://doi.org/10.1177/20552076231220833

Poslon, L. (2022). Ljudsko dostojanstvo i kraj života. Obnovljeni Život, 77 (2), 235-248. https://doi.org/10.31337/oz.77.2.7

Poslon, L. (2021). Anto ČARTOLOVNI, Ethical and anthropological aspects of the emerging field of neuroprosthetics, Aracne editrice, Rome, 2016., 184 str.. Diacovensia, 29 (4), 644-649.