

Izvirni znanstveni članek/Article (1.01)

Bogoslovni vestnik/Theological Quarterly 84 (2024) 4, 783—795

Besedilo prejeto/Received:09/2024; sprejeto/Accepted:12/2024

UDK/UDC: 004.021:364.63:343

DOI: 10.34291/BV2024/04/Strahovnik

© 2024 Strahovnik, CC BY 4.0

Vojko Strahovnik

Algoritemsko nasilje *Algorithmic Violence*

Povzetek: Razvoj najnovejših sistemov in orodij na področju umetne inteligence ter poudarek na etičnih vprašanjih, ki jih to prinaša, je v literaturi sprožil pojav novega pojma – algoritemsko nasilje. Vezan je na nastanek, uporabo in posledice rabe sistemov umetne inteligence, ki temeljijo na algoritmih. Naš cilj v tem prispevku bo analiza in kritično ovrednotenje tega pojma. Ugotovitve bodo pretežno skeptične. Trdili bomo, da ta pojem pomembne etične dileme v zvezi z uporabo umetne inteligence do neke mere sicer izpostavlja, vendar pa jih precej bolj smiselno in natančno označujejo drugi izrazi, kot so pristranskost, diskriminacija, izključevanje, stereotipi, marginalizacija, nepravičnost, netransparentnost ipd. Uporaba pojma algoritemsko nasilje tako ni dobro utemeljena in ne prinaša posebnih teoretskih ali praktičnih prednosti ali koristi.

Gljučne besede: umetna inteligenca, algoritmi, nasilje, algoritemsko nasilje, spoznavna nepravičnost, pristranskost

Abstract: With the development of the latest systems and tools in the field of artificial intelligence and the emphasis on the ethical issues it raises, a new concept has emerged in the literature. This is the concept of algorithmic violence. It refers to the development, use, and consequences of AI systems based on algorithms. Our aim in this paper will be to analyse and critically evaluate this notion. The conclusions will be mainly sceptical, but they carry significant implications for the field. It will be argued that while the term does, to some extent, highlight important ethical problems in the use of AI, these are much better characterized by other terms such as bias, discrimination, exclusion, stereotyping, marginalization, unfairness, non-transparency, etc. The use of the term algorithmic violence is thus not well founded and does not bring any particular theoretical or practical advantages or benefits.

Keywords: artificial intelligence, algorithms, violence, algorithmic violence, epistemic injustice, bias

1. Uvod

Z razvojem najnovejših tehnologij na področju umetne inteligence, obenem pa tudi poudarkom na etičnih in širših filozofskih vprašanjih, ki jih te sprožajo (Strahovnik 2023; Žalec 2023), se je v literaturi začel pojavljati nov pojem.¹ To je pojem algoritemskega nasilja (Onuoha 2018; Huws 2024). Moj cilj v tem prispevku bo ta pojem analizirati in kritično ovrednotiti. Moje ugotovitve bodo pretežno skeptične. Trdil bom, da ta pojem sicer izpostavlja pomembne etične dileme, vendar pa te dejansko bolje označujejo drugi izrazi, kot so pristranskost, diskriminacija, izključevanje, stereotipizacija, marginalizacija, nepravilnost, netransparentnost itd. Tovrstni pojmi stara in nova etična vprašanja ter izzive, ki jih je najnovejši napredek na področju umetne inteligence sprožil, zajamejo na bolj celosten in informativen način. Poleg tega so uporabnejši pri iskanju rešitev in oblikovanju strategij za izogibanje omenjenim negativnim učinkom ali tveganjem pri uporabi algoritmov umetne inteligence.

Prispevek v drugem razdelku predstavlja pojem algoritemskega nasilja, v treh podrazdelkih pa še posebej očituje in analizira tri idejne podlage oz. izhodišča, iz katerih uvedba in raba tega pojma izhajata. V tretjem razdelku predstavljam skeptične premisleke, ki govorijo proti rabi pojma algoritemsko nasilje. Pri tem so izpostavljene tudi različne strategije obrambe uporabe tega pojma kot ugovor mojim tezam; ta je oblikovan na podlagi tega, da razprava, vključno s tezo o neustreznosti rabe pojma algoritemsko nasilje, predpostavlja, da moramo najprej opredeliti samo nasilje – in šele nato presojati algoritemsko nasilje.² V četrtem, zadnjem razdelku je izpostavljeno, kako lahko uporaba bolj določenih pojmov tipa pristranskost, netransparentnost, diskriminacija in podobnih tveganja pri rabi umetne inteligence izpostavi bolje. Prav tako se uporaba teh pojmov bolje vpenja v prizadevanja za oblikovanje mreže odgovornosti ter z njimi povezanih dolžnosti za čim večjo omilitev teh tveganj.

2. O pojmu algoritemskega nasilja

Pojem algoritemskega nasilja je razmeroma nov in se je v razpravah pojavil šele v zadnjih letih. Ena prvih omemb je iz leta 2018. Pojavila se je na spletu, v objavi na platformi za razvijalce GitHub; njena avtorica je Mimi Onuoha, raziskovalka in umetnica. Onuoha v tem zapisu algoritemsko nasilje opredeljuje takole:

»Algoritemsko nasilje se nanaša na nasilje, ki ga algoritem ali sistem za samodejno odločanje povzroča tako, da posameznikom preprečuje, da bi

¹ Publikacija je nastala v okviru raziskovalnega programa P6-0269 „Religija, etika, izobraževanje in izzivi sodobne družbe“ in raziskovalnega projekta J6-4626 „Teologija, digitalna kultura in izzivi na človeka osredičene umetne inteligence“, ki ju finančno podpira Agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS), ter v okviru raziskovalnega projekta „Epistemic Identity and Epistemic Virtue: Human Mind and Artificial Intelligence“ s podporo fundacije John Templeton (krovni projekt „New Horizons for Science and Religion in Central and Eastern Europe“, Univerza v Oxfordu)

² Za jasno izpostavitev tega vidika oz. ugovora se zahvaljujem Bojanu Žalcu.

zadovoljili svoje osnovne potrebe. Izvira iz izkoriščevalskih družbenih, političnih in ekonomskih sistemov ter se z njimi krepí, lahko pa je tudi tesno povezano s prostorsko in fizično prenesenimi učinki.« (Onuoha 2018)

Vidimo, da Onuoha algoritemsko nasilje v okviru te opredelitve povezuje prvenstveno s priporočilnimi oz. odločitvenimi sistemi, a opozoriti je treba, da ga v nadaljevanju prispevka ter tudi v spremljajoči razpravi ne omejuje zgolj in samo na takšne sisteme umetne inteligence. To njeno začetno zamisel o vpeljavi novega pojma so nato prevzeli tudi drugi. Safransky (2020) v kontekstu prostorskih študij algoritemsko nasilje opredeljuje podobno in uvaja besedno zvezo »geografije algoritemskega nasilja«, ki označuje »ponavljajočo se in standardizirano obliko nasilja, ki prispeva k rasizaciji prostora in prostorski določenosti revščine« (Safransky 2020, 200) – to je zanj posledica podatkovno vodenih tehnologij načrtovanja in algoritemskega odločanja pri urejanju prostora. Eden od vidikov oz. posebnosti, ki jih omenja, je ta, da je algoritemsko nasilje bolj skrito ali manj očitno kot druge oblike nasilja, ki so vidne ali ‚spektakularne‘. Huws in sodelavci (2023) pojem algoritemskega nasilja uvajajo v kontekstu analize visokošolskega oz. akademskega prostora ter v njem navzočega nadzora, upravljanja in nasilja, ki spremlja opravljanje dela. Tega tesno povezujejo s pojmom birokratskega nasilja – to se lahko porodi na temelju etične brezbriznosti, ki sledi razčlovečenju v kontekstu birokratskega upravljanja (Bauman 2006, 163–165). Mbalaka (2023) algoritemsko nasilje navezuje na pojem spoznavnega nasilja in ga tako pravzaprav razume kot vrsto spoznavnega nasilja, kjer ključno vlogo pri produkciji oz. implementaciji tega nasilja igrajo algoritmi umetne inteligence.

Nekoliko širše lahko rečemo, da razprava o algoritemskem nasilju zadeva »izpostavitev, razkritje in kritično proučitev, kako algoritemski sistemi spodbujajo določene oblike nasilja, kako nasilna dejanja opravičujejo ali na novo določajo, kakšno nasilje velja za legitimno« (Bellanova et al. 2021, 123). Kar naj bi algoritemsko nasilje oddvajalo od preostalih oblik nasilja, je torej to, da naj bi bili pri takem nasilju v njegovo porajanje, uresničevanje ali upravičevanje v bistveni meri vpeti algoritmi. Pri tem je treba izpostaviti, da je v večini primerov in razprav sam pojem algoritma razumljen na širok način in kot tesno povezan s tehnologijami umetne inteligence. Ne gre torej le za zaporedje natančno določenih in računalniško izvedljivih korakov oz. napotkov, ki privede do želenega rezultata ali rešitve, pač pa za vsakršne algoritme, ki so vključeni v procese učenja in delovanja sistemov umetne inteligence – ki torej strojem omogočajo, da s pomočjo predhodnega procesa učenja posnemajo človeško inteligenco in samostojno izvajajo zapletene naloge. Etični pomisleki in tveganja, ki jih delovanje takšnih sistemov odpira, so že bili jasno izpostavljeni (Weidinger et al. 2022; Strahovnik, Miklavčič in Centa Strahovnik 2020). Transparentnost, razločljivost, razumljivost, sledljivost, preglednost, poštenost, nediskriminatornost, predvidljivost in odgovornost so ključna merila in družbene zahteve, ki veljajo tako za ljudi kot tudi za ustanove, ki opravljajo družbene funkcije. Ko algoritmi umetne inteligence prevzamejo naloge oz. delo, ki so jih prej izvajali ljudje ali institucije, prevzamejo tudi družbene zahteve,

vkjučno z zgoraj omenjenimi merili (Bostrom in Yudkowsky 2014). Zdi se, da želi razprava o algoritemskem nasilju k zgoraj omenjenim premislekom dodati še vidik nasilja ter zahtevo po nenasilnosti.

Konceptualizacijo algoritemskega nasilja ter utemeljitev tega pojma lahko tako obravnavamo z več vidikov. V nadaljevanju na kratko izpostavljam tri najbolj razširjene, ki se medsebojno sicer ne izključujejo. Mednje spadajo okvir strukturnega nasilja, okvir birokratizacije z upravljavskimi strukturami ter epistemičnega oz. spoznavnega nasilja. Ne gre zgolj za vidike obravnave pojma algoritemsko nasilje, ampak tudi za izhodišča, na podlagi katerih zgoraj omenjeni avtorji sam pojem algoritemskega nasilja sploh zasnujejo.³ Glede na ta izhodišča bomo obravnavani pojem v nadaljevanju tudi lažje ovrednotili.

2.1 Strukturno nasilje

Prvi okvir, ki predstavlja podlago za vpeljavo pojma algoritemskega nasilja, je okvir strukturnega nasilja. Strukturno nasilje se nanaša na načine, kako družbene strukture in institucije povzročajo škodo in ohranjajo neenakost. Kot izraz ga je prvi skoval norveški sociolog Johan Galtung (1969), ki ga je v prvi vrsti zanimalo zlasti, kako družbene institucije, strukture in sistemi vplivajo na marginalizirane skupnosti. Glavni namen uvedbe pojma strukturno nasilje je bil poudariti posredne, sistemske in pogosto neopažene mehanizme, ki določene skupine ali posameznike pogosto prikrajšajo za izpolnitev osnovnih človekovih potreb, priložnosti in pravic. Galtung je ta koncept uvedel v zvezi z vprašanjem delovanja nasilja in njegovih povzročiteljev. Izhodiščna vprašanja so bila zlasti naslednja:

»ali obstaja subjekt (oseba), ki deluje, ali ne. Ponovno se lahko vprašamo: ali o nasilju lahko govorimo, če nihče ne izvaja neposrednega nasilja, ni dejaven? /.../ Vrsto nasilja, kjer akter, ki izvaja nasilje, obstaja, bomo poimenovali osebno ali neposredno; nasilje, kjer takega akterja ni, pa strukturno ali posredno. V obeh primerih so lahko posamezniki ubiti ali poškodovani, prizadeti ali poškodovani v obeh pomenih teh besed ter manipulirani s strategijami palice ali korenčka. Toda medtem ko je v prvem primeru te posledice mogoče pripisati konkretnim osebam kot akterjem, v drugem primeru to ni več smiselno. V strukturi morda ni osebe, ki bi drugi osebi škodovala neposredno. Nasilje je vgrajeno v strukturo in se kaže kot neenaka moč – in posledično kot neenake življenjske možnosti.« (170–171)

³ Opozoriti moramo, da se s tem izrazom v navezavi na umetno inteligenco običajno povezuje dva pomenska odtenska. Prvi je prevladujoč in se nanaša na nasilje ter z njim povezano škodo, povzročeno osebi, skupini ali skupnosti, ki jo prizadene zasnutje, razvoj ali uporaba algoritmov, temelječih na umetni inteligenci. Standardni primer bi bila škoda, povzročena skupini, ki je ob implementaciji algoritma umetne inteligence sistematično diskriminirana (npr. generativni veliki jezikovni model ali generator slik, ki temnopolte osebe povezuje z negativnimi vlogami ali jih kot del tvorbe slik prikazuje neustrezno). Drugi pomenski odtенок je nekoliko drugačen in se nanaša na nasilje, ki je vgrajeno v sam razvoj oz. trening globokih nevronske mreže z uporabo algoritmov – v tem primeru je razumljeno kot nasilje do nevronske mreže (Kockleman 2020, 326). V tem prispevku se omejujemo na prvi, bolj običajen pomen, drugega pa puščamo ob strani.

Strukturno nasilje je oblika nasilja, do katerega prihaja pri delovanju nepravilnih sistemov, npr. institucionaliziranega razizma ali kastnega sistema, ko ti sistemi in strukture nesorazmerno vzpostavljajo, ohranjajo ali poglobljajo trpljenje in prikrajšanost določenih skupin prebivalstva. Drugače kot neposredno nasilje, ki med drugim lahko vključuje fizične poškodbe, smrt ali prisilo (v izvedbi delovalca), strukturno nasilje deluje prek družbenih struktur, sistemov, politik, norm in ustvarja ali ohranja pogoje, zaradi katerih so določene skupine prebivalstva prikrajšane.

Vidik strukturnega nasilja je v tem smislu tudi temeljna podmena za razvoj pojma algoritemsko nasilje, kot ga predlaga in oblikuje Mimi Onuoha, pri čemer navaja nekaj povezav in primerov. Med drugim rabo izraza algoritemsko nasilje utemeljuje s tem, da »potrebujemo izraz, ki obravnava novejšo, pogosto digitalno in s podatki podprto obliko neenakosti. Kot prvi korak pri artikulaciji teh iskanj želim postaviti besedno zvezo algoritemsko nasilje.« (2018) Nadaljuje, da pojem algoritemskega nasilja vključuje pojave, ki smo jim bili priča v zadnjem desetletju in ki so takšni, da slonijo na vse večji dostopnosti obsežnih podatkovnih zbirk, napredku v računski moči ter predvsem hitrem razvoju na področjih umetne inteligence in strojnega učenja in uporabi teh tehnologij v hierarhično neenaki družbi. Algoritmčno nasilje zajema tako mikrodogodke kot obsežne in daljnosežne pojave oz. procese. Onuoha opisuje nekatere takšne pojave:

»To je tisti neprijeten občutek, ki ga dobite, ko si na spletu ogledate majico, nato pa se ta majica do konca dneva oglašuje na vseh spletnih straneh, ki jih obiščete. /.../ To je bistvo frustracij, ki jih čutijo vozniki Uberja, Lyfta in Juna, ko jim njihove aplikacije narekujejo, naj opravljajo navidezno smiselne prevoze ali naj lovijo ponudbe ›Odlična priložnost!‹ ali ›Ravno v pravem času‹, zaradi česar na koncu pravzaprav prejmejo nižjo plačo. /.../ V okvir pojma algoritemsko nasilje bi morali uvrstiti tudi nekatera dejanja Facebooka, na primer ko je podjetje uporabnikom predpisalo, da se morajo v račune prijaviti s praviimi imeni, a je deaktiviralo račune ljudi, katerih imena se niso zdela legitimna. /.../ Omenili bi lahko tudi omejitve za iskalce zaposlitve – življenjepise jim avtomatski sistemi za sledenje vlogam zavrnejo, ker v njih ni ›pravih‹ ključnih besed. /.../ Vse to so oblike algoritemskega nasilja.« (Onuoha 2018)

Algoritemskemu nasilju je s strukturnim nasiljem skupna predvsem poudarjena odsotnost jasno prepoznavnega subjekta ali povzročitelja nasilja ter predhoden obstoj marginaliziranih skupin, ki jih tovrstno nasilje še posebej prizadene.

2.2 Birokratizacija in moderno upravljanje

Druga podmena pobud za uveljavitev pojma algoritemskega nasilja je povezana s področjem birokratizacije in modernega upravljanja (Huws et al. 2023). Avtorji kot del tega okvira pogosto navajajo pojave, kot so organizacijska kultura s standardiziranimi in enoznačne cilje določujočimi pravili, metrikami in rigidnimi postopki nadzora in vrednotenja, strukturami vodenja in predpisi glede vedenja oz. izpol-

njevanja vlog. V tem smislu so algoritmizacija, avtomatizacija in digitalizacija del istega procesa, ki najprej zajame samo delo, potem pa tudi celotno posameznikovo življenje. S tega vidika je algoritemsko nasilje razumljeno kot »praksa usmerjanja delavca ali skupine delavcev s spodkopavanjem, izključevanjem, nerazumnimi zahtevami in, kar je ključno, z uporabo birokratskih in algoritmičnih ukrepov nadzora in ocenjevanja« (99), ki privede do nezaželenih, pogosto lahko tudi zelo težkih posledic za zaposlene. V tem smislu je algoritemsko nasilje povezano z algoritemskim upravljanjem, imenovanim tudi ,algokracija': algoritem v postopku ali procesu predpiše niz korakov, ki izključujejo vse odklone ali drugačne posege, s tem pa so povezane tudi prakse kaznovanja in izključevanja (103). Gre za upravljanje področij, ki posameznikovo življenje bistveno določajo (npr. vrednotenje dela, promet, upravni postopki idr.).

Ta teoretski okvir, ki je v ozadju pojma algoritmično nasilje, se opira predvsem na dve vrsti nasilja, in sicer na birokratsko nasilje oz. tehnokratsko nasilje (Bauman 2006; Arendt 1969) in počasno nasilje (Nixon 2006). Arendt (1969) je birokracijo oz. njeno prevlado izpostavila kot prefinjen sistem dominacije, kjer se odgovornost, ki je ni več moč pripisati nikomur, porazgubi, zato pa tudi gre za obliko vlade nikogar. V tem smislu gre tudi za obliko tiranije, saj slednjo določa prav to, da manjka vsakršna odgovornost vladajočega. Izpostavila je še, da večja kot je birokratizacija javnega življenja, večja je težnja po nasilju. »V popolnoma razviti birokraciji ni več nikogar, s komer bi se lahko prepirali, komur bi lahko podali svoje pritožbe, na kogar bi lahko izvajali pritiske moči. Birokracija je oblika vladavine, v kateri so vsi prikrajšani za politično svobodo, za moč delovanja; kajti vladavina Nikogar ni ne-vladavina – in kjer so vsi enako nemočni, imamo tiranijo brez tirana.« (81) Počasno nasilje pa je – drugače kot običajno nasilje, ki je vidno oz. lahko prepoznavno ter povezano z dogodkom ali dejanjem – razumljeno kot nasilje, ki se dogaja postopoma in neopazno, ki je razpršeno v času in prostoru, njegove posledice pa so določljive šele z zamudo – torej nasilje, ki ga običajno sploh ne obravnavamo kot nasilje (Nixon 2006, 2). Algoritemsko nasilje v tem smislu deli značilnosti tovrstnega nasilja. Novost je, da se plasti birokratskega nasilja in počasnega nasilja dodaja plast delovanja algoritmov umetne inteligence.

2.3 Spoznavno oz. epistemično nasilje

Tretja podmena zagovora uvedbe pojma algoritemsko nasilje je vidik epistemičnega nasilja. Eden izmed načinov, kako je algoritemsko nasilje povezano s spoznavnim nasiljem, so vidiki algoritmov, ki igrajo vlogo popolnoma natančnih opisov, objektivnih ,rešitev' ali racionalnih interpretacij realnosti, ki je predvidljiva, vključno s »posebnimi oblikami znanja in nevednosti, ki jih ti sistemi ustvarjajo« (Bellanova et al. 2021, 123). Preden se vrnemo k algoritemskemu nasilju, si najprej približajmo pojem spoznavnega nasilja. Spivak (1998, Dotson 2011) spoznavno nasilje opredeljuje kot vrsto nasilja, ki poskuša odpraviti znanje oz. vednost, ki jo imajo marginalni subjekti. Ena od vrst takšnega spoznavnega nasilja se na primer pojavlja na področju začenja. Dotson (2011, 238) tovrstno nasilje označuje kot nasilje, ko občinstvo/poslušalstvo (namerno ali nenamerno) zavrača vzajemnost

v jezikovni, komunikacijski izmenjavi, ta zavrnitev pa temelji na ‚škodljivi nevednosti‘, ki je zanesljiva nevednost, nastajajoča zaradi spoznavnih vrzeli. Lahko nastopa v obliki utišanja ali podrejanja. Takšna zastavitev ima izhodišče v razpravah o spoznavni nepravčnosti.

Spoznavna nepravčnost je celostno uvedena in utemeljena v knjigi Mirande Fricker z naslovom *Epistemic Injustice: Power and the Ethics of Knowing* (Fricker 2007; Strahovnik 2018). Da bi lahko spoznavno nepravčnost ustrezno razumeli, se moramo osredotočiti na spoznavne prakse spoznavnih akterjev, ki so umeščene v okvir družbe, družbenih odnosov, odnosov moči in njihove družbene identitete. Ob tem Fricker izpostavlja metodološko primarnost spoznavne nepravčnosti, ki misli o (spoznavni) pravičnosti predhodi – z osredotočanjem na pojavne oblike spoznavne nepravčnosti njeno bistvo razumemo lažje, kot pa če bi pozornost usmerjali le na vrzeli v pravičnosti oziroma jo razumeli kot manko pravičnosti. Fricker razlikuje med dvema vrstama spoznavne nepravčnosti, in sicer med začetvalno nepravčnostjo in hermenevtično nepravčnostjo. Jasen primer prve je denimo nezaupanje ali neupoštevanje koga zgolj na podlagi njegove rase, spola, družbenega položaja ipd. Takšen posameznik s tega vidika trpi spoznavno nepravčnost, saj je izključen iz družbenih spoznavnih praks, obenem pa se mu odreka vloga nosilca vednosti. Hermenevtična nepravčnost pa se vzpostavlja tam, kjer prihaja do vrzeli v hermenevtičnih virih, ki so na voljo za vednost in razumevanje – kjer na primer posamezniki ali družbene skupine nimajo orodij za doseganje vednosti in ustrezno razumevanje npr. lastnega izkustva (Fricker 2007 9–29; 147–175).

Zelo pomembno je, da izpostavimo značilno spoznavno naravo take nepravčnosti, ki sega onkraj vprašanj o samem dostopu do informacij, vednosti ali dostopnosti šolanja za posameznike ali skupine. Pri spoznavni nepravčnosti gre za pristno spoznavno nepravčnost, ki je bistveno povezana s spoznavnimi praksami, kot sta prenašanje vednosti na druge preko pričevanja ter oblikovanje ustreznega razumevanja našega lastnega izkustva (1). Da bi zgoraj omenjeni obliki spoznavne nepravčnosti lažje razumeli, moramo najprej razumeti pojmovanji moči in identitete, s katerima jih lahko opredelimo. Moč je družbeno umeščena zmožnost nadzora in vpliva na dejavnost drugega oziroma drugih. Pri hermenevtični obliki spoznavne nepravčnosti razumevanje onemogoča ali hromi hermenevtična vrzel – običajno je povezana z različnimi družbenimi identitetami, različne skupine v družbi pa prizadene različno. Velja, da se posamezne skupine hermenevtično marginalizira s krnjenjem njihove udeležbe v spoznavnih in družbenih praksah, ki ustvarjajo skupnostne smisle. Spoznavna nepravčnost pa privede do spoznavne škode, ki koga prizadene kot subjekta oziroma nosilca vednosti ali pa katere druge spoznavne vrednote oz. cilja (Strahovnik 2018). Fricker kot rešitev za preseganje spoznavne nepravčnosti predlaga razvijanje in kultiviranje specifičnih spoznavnih kreposti. Začetvalsko občutljivost tako izpostavlja kot obliko razumske občutljivosti. Razvijamo jo lahko s pozornostjo na začetvalske prakse, ustreznim oz. zanesljivim pripisovanjem kredibilnosti, pozorno zaznavo in dojemanjem spoznavnega statusa ipd. Dopolnjuje jo krepost hermenevtične pravičnosti: ta predpostavlja, da po-

skušamo kritično premisliti, smo pozorni in občutljivi na morebitne hermenevitične vrzeli pri govorcu – in svoje razumevanje temu prilagodimo (Fricker 2007, 5–7). V nekaterih okvirih razprave o spoznavni nepravilnosti srečamo tudi govor o spoznavnem nasilju (McKinnon 2016; Dotson 2011). Spoznavno nasilje je to, do česar spoznavna nepravilnost lahko privede oziroma vanj preide. Spoznavno nasilje med drugim vključuje spoznavno zatiranje ali utišanje. In na tej točki se v okvir naše razprave vpne tudi vpeljava pojma algoritemsko nasilje.

Mbalaka (2023) govori o spoznavnonasilnih pristranskostih pri oblikovanju in delovanju sistemov umetne inteligence in opozarja, kako lahko sistemi umetne inteligence vodijo do spoznavnega nasilja. To prikazuje s primerom delovanja algoritmov umetne inteligence, ki so del modelov Dall-E 2 in Starry AI za generiranje slik oz. pretvorbo besedilnega poziva v sliko. Trdi, da sta oba modela zasnovana tako, da odražata implicitne predsodke sodobne družbe in jih s svojim delovanjem ohranjata. Algoritemska pristranskost torej lahko povzroči algoritemsko spoznavno nasilje, ki vključuje pojave, kot so zatiranje različnih stališč, premajhna zastopanost določenih sistemov vednosti in zniževanje njihovega pomena. Kot primere tovrstnega nasilja izpostavlja pogostost pomanjkljivih in netočnih prikazov tipične temnopolte družine, pa tudi generične ali halucinirane upodobitve določenih kulturnih artefaktov (npr. tipičnih vzorcev na oblačilih, drugega okrasja in predmetov). Ker so v te fenomene bistveno vpeti algoritmi umetne inteligence, Mbalaka tovrstno nasilje povezuje z njimi.

3. Skeptični premisleki glede pojma algoritemsko nasilje

Preden preidem k bolj specifičnim skeptičnim premislekom o pojmu algoritemskega nasilja, naj opozorim, da lahko pri pogledu na razumevanje določene vrste nasilja upoštevamo naslednje tri temeljne vidike: (i) kdo ali kaj je povzročitelj nasilja; (ii) način, na katerega delovalec ali povzročitelj nasilje izvaja ali ohranja škodo, neenakost itd.; (iii) vrsto nasilja z vidika posebnih posledic oz. učinka na žrtve nasilja. Ti trije vidiki se medsebojno ne izključujejo in se v več primerih združujejo ali prekrivajo. Vsi trije gradniki so načeloma lahko podlaga za smiselnost uvedbe pojma algoritemsko nasilje.

Izhodiščni skeptični premiselek lahko uokvirimo v zvezi z zelo širokim razumevanjem izraza ‚algoritemsko nasilje‘. Nekateri avtorji algoritemsko nasilje v razpravi uporabljajo v širokem pomenu, ki temelji na ideji, da je vsakršna raba algoritmov in s tem delovanje algoritmov samo po sebi nasilno. Podobno široke so tudi razlage, ki prakse datafikacije in digitalizacije obravnavajo kot že obremenjene z nasiljem. V obeh primerih delovanje algoritmov, ki temelji na umetni inteligenci, sploh ne more biti nenasilno (Bellanova idr. 2021, 126). Algoritemsko nasilje je vseprisotno in neizogibno, posledično pa je koncept algoritemskega nasilja v najboljšem primeru dvoumen ali v najslabšem nesmiseln – v nobenem primeru pa ne more biti primeren za poskus razumevanja in blaženja različnih škodljivih učinkov, ki jih uporaba algoritmov umetne inteligence povzroča ali bi jih lahko povzročila. Hkrati ni povsem jasno, komu ali čemu ter v kakšni obliki je po tem širokem

razumevanju škoda povzročena. Tretja težava takšnega razumevanja nasilja, ki da ga algoritmi povzročajo, je, da pušča odprto vprašanje, kako potem razumeti in upoštevati pozitivno in koristno uporabo algoritmov – npr. pri prepoznavanju, preprečevanju ali zdravljenju bolezni, pri varovanju okolja ipd. (Strahovnik, Miklavčič in Centa Strahovnik 2020), kar se vsaj po nekaterih kriterijih lahko upravičeno zdi zaupanja vredno (Miklavčič 2021).

Zagovorniki izraza bi odgovor lahko oblikovali v naslednji smeri, in sicer, da je »kritična dodana vrednost algoritemskega nasilja (ta), da lahko opozori na prakse, ki niso nasilne v ‚tradicionalnem‘ smislu in jih zato ni mogoče obravnavati kot takšne, če želimo izpolniti to merilo.« (Bellanova et al. 2021, 126) Temu lahko pritrudimo, a vseeno se lahko upravičeno vprašamo o smiselnosti izbire tega izraza, ne da bi bili predstavljeni dobri primeri takšnega algoritemskega nasilja, ki jih ne moremo opisati na drugačen način oz. z rabo drugih izrazov.

Druga možnost obrambe koncepta algoritemskega nasilja je naslednja: algoritmi, ki temeljijo na umetni inteligenci, bi lahko nasilno delovanje spodbujali ali pa se uporabljali kot sredstva zanj. S tem povezana konceptualizacija algoritemskega nasilja je torej nasilje, ki se doseže s temi *sredstvi*. Šlo bi za situacije, ki vključujejo zanašanje na algoritme kot pomoč pri sprejemanju odločitev v kontekstih, v katerih posledice takšnih odločitev posameznikom ali skupnostim nedvomno lahko povzročijo škodo. Primer bi bili sistemi ‚točkovanja‘ v kazenskem pravosodju oz. kot del kaznovanja ali pa uporaba polavtonomnega orožja. Tako algoritmi umetne inteligence omogočajo nasilna dejanja, izvedena na daljavo, kot je to v primeru vojaške tehnologije, ki vključuje daljinsko vodene ali polavtonomne oborožitvene sisteme (Bellanova et al. 2021, 127).

Vendar se ob upoštevanju tega odgovora znova pojavi vprašanje implementacije, upravljanja in zastopanja. Najbolj naravno je razumeti, da algoritmi sami nimajo sposobnosti delovanja ali namena – nasprotno, so orodja, ki jih ustvarijo in uporabljajo ljudje. Zato lahko pripisovanje nasilja algoritmom zamegli vlogo in odgovornost človeških odločevalcev za rezultate ali posledice delovanja algoritemskih sistemov. Tovrsten temelj za določitev algoritemskega nasilja kot posebne vrste nasilja pa tu umanjka. Pomembno je osvetliti tudi pretiran poudarek na nekakšnem tehnološkem determinizmu kot prepričanju, da tehnologija družbene izide določa neodvisno od človekovega delovanja. Osredotočanje le na algoritme kot vir družbenih težav spregleduje širše družbene, gospodarske in politične dejavnike, ki k neenakostim in nepravilnostim vsi prispevajo. Algoritmi umetne inteligence so precej kompleksni in večplastni sistemi, zato je škodo težko pripisati zgolj algoritmom – izid in njegove posledice določa interakcija med algoritmi, zbirkami podatkov, uporabniki in družbenim kontekstom.

Vredno je opozoriti tudi na naslednje razlikovanje, ki ga lahko ponazorimo na primeru utemeljitve pojma algoritemskega nasilja s pojmom spoznavnega nasilja. Za oceno smiselnosti rabe algoritemskega nasilja je koristno razlikovati med dvema glavnima perspektivama ali možnostma. Prva možnost je, da se spoznavno nasilje, ki je pripisano sistemom umetne inteligence, prenese iz prejšnjega spo-

znavnega nasilja, ki ne temelji (nujno) na umetni inteligenci ali njeni rabi, temveč je povezano s pristranskostmi pri zbiranju, skrbništvu in uporabi pristranskih učnih podatkovnih nizov. Pri generativnih sistemih umetne inteligence gre preprosto za to, da so bili naučeni na nizu vhodnih podatkov, ki je privilegiral določene skupine oz. značilnosti. Druga možnost pa je spoznavno algoritemsko nasilje, ki se pojavlja predvsem zaradi samega delovanja algoritmov umetne inteligence (bodisi ker so ali bi bili namerno zasnovani kot pristranski bodisi kot stranski produkt optimizacije algoritmov, npr. za hitrost). Vendar avtorji (Mbalaka 2023) ne navajajo primerov, ki bi ta vidik podpirali.

O vsem zgoraj omenjenem bi lahko imeli tudi naslednji pomislek. Ne moremo biti skeptični oz. pojma algoritemskega nasilja zavračati, če prej sami ne ponudimo opredelitve nasilja in na tej podlagi dokažemo, da algoritemsko nasilje okvirom te opredelitve ne zadosti. Odziv na ta izziv je dvojen oz. ima dva kraka.

Pri prvem kraku izberemo določeno definicijo nasilja in potem posledično ugotavljamo, ali je glede na to definicijo o algoritemskem nasilju upravičeno govoriti. Ta strategija je sicer možna, vendar v pomembnem smislu predstavlja zmoto zahteve po vprašanju (angl. *begging the question*). Ena izmed bolj razširjenih in uporabljenih definicij nasilja je tista, ki jo je podala Svetovna zdravstvena organizacija. Ta (WHO 1996; cf. WHO 2002, 5) nasilje opredeljuje kot »namerno uporabo fizične sile ali moči (v obliki grožnje ali dejansko) proti sebi, drugi osebi, skupini ali skupnosti, ki ima ali bi lahko imela za posledico poškodbe, smrt, psihično škodo, slab razvoj ali prikrajšanje.« Nadalje WHO nasilje deli na tri kategorije glede na to, kdo ga izvaja (samopovzročeno oz. nase usmerjeno, medosebno in kolektivno), in štiri nadaljnje kategorije glede na vrsto nasilja (fizično, spolno, psihološko ali nasilje, ki vključuje prikrajšanje ali zanemarjanje) (WHO 2002, 6; Dwyer 2022). Glede na to razmeroma ozko opredelitev nasilja večine pojavov, ki jih avtorji zgoraj podajajo kot primere algoritemskega nasilja, ne moremo označiti kot nasilje.

Pri drugem kraku izhajamo iz opredelitev oz. pojmovanj nasilja nasploh, ki jih uporabljajo oz. na katere se sklicujejo sami avtorji, ki zagovarjajo vpeljavo pojma algoritemsko nasilje. Če se najprej posvetimo vidiku strukturnega nasilja, potem lahko izpostavimo, da Galtung (1969) ne izhaja iz kakšnega splošnega pojma nasilja, ampak pravzaprav strukturno nasilje in s tem razširjeno definicijo nasilja uvaža prav zato, ker zanj običajni pojem nasilja ni ustrezen – ne zajame namreč celote nasilja (kot nasprotja miru) oz. vseh njegovih razsežnosti. Odpira se vprašanje, ali je to ustrezna podlaga tudi za uvedbo pojma algoritemsko nasilje. Pri strukturnem nasilju gre za nasilje, ki je posredno in ki je vgrajeno v družbeno strukturo. Seveda so lahko del družbene strukture in delovanja institucij družbe tudi algoritmi umetne inteligence, vendar ti do nasilja privedejo le v primeru, ko ima takšne učinke struktura kot celota. Algoritmi umetne inteligence nikoli ne predstavljajo strukture kot take, ampak so kvečjemu njen del oz. služijo kot morebitno sredstvo strukturnega nasilja. Za pojem strukturnega nasilja je Galtung predlagal tudi rabo izraza družbena nepravilnost (Galtung 1969, 171). Podobne pomisleke imamo pri podmeni, ki se pri zagovoru rabe pojma algoritemsko nasilje naslanja na vidik birokratizacije in administracije. Birokratizacija in birokratsko upravljanje seveda lahko

privedeta tudi na nasilja (kakor to pri holokavstu izpostavlja Bauman [2006]), a to ne pomeni, da je vsakršna birokratizacija že nasilna oz. sama po sebi nasilje. In tudi če pogledamo zagovor pojma algoritemsko nasilje поближе (Huws et al. 2023), lahko vidimo, da ključni poudarek v zvezi z negativnimi pojavi, ki so omenjeni (npr. psihološka škoda, izgorelost, depresija, ipd.), ni na sistemih umetne inteligence, temveč na razvoju upravljalvske strukture in tudi širše organizacijske kulture, ki tovrstno nasilje omogoča oz. do teh negativnih pojavov privede.

Za konec se lahko obrnemo še k tretji podmeni, to je polju spoznavne pravičnosti. Kot smo že omenili, je govor o spoznavnem nasilju tesno povezan z razpravo o spoznavni (ne)pravičnosti. Vendar si je smiselno podrobneje ogledati, na kaj se ta izraz sploh nanaša. Kot je izpostavljeno že zgoraj, se z izrazom spoznavno nasilje najpogosteje označuje prakse utišanja, nerazumevanja in zatiranja, ki so povezane s sistemi vednosti in subjekti kot nosilci vednosti. Eden izmed primerov, ki ga omenja Dotson (2011, 240), je primer triletnega otroka, ki zavrne nasvet oz. opozorilo odraslega, da je ogenj nevaren in da se z njim ne sme igrati. Dotson na podlagi različic tega primera trdi, da do spoznavnega nasilja, kadar ga vrši otrok, pride v primerih, ko je odrasli prizadet kot nosilec vednosti ali pa ko zaradi otrokovega ravnanja pride do škode (npr. materialne škode ob požaru). Upravičeno se lahko vprašamo, ali ne gre v teh in podobnih primerih za neustrezno in neupravičeno široko razumevanje pojma spoznavnega nasilja oz. ali ne bi bilo ustrezneje izbrati drug opis. Če denimo skupnost zaničevanju osebe ne verjame zgolj na podlagi njenega spola ali rase, potem je to lahko posledica siceršnje diskriminacije, predsodkov, predhodnega strukturnega nasilja ipd. in v tem smislu vrsta spoznavne nepravičnosti. In če tovrstne situacije označimo kot spoznavno nasilje, ne dodamo nobenih pomembnih določil oz. podlage za ustreznejše razumevanje situacije. Podobno lahko trdimo tudi za pojem algoritemskega nasilja. Če nam algoritem na podlagi majhnega vzorca našega vedenja ali podatkov, ki jih o nas ima/zbira, vztrajno ponuja oglas za določeno stvar ali storitev, je to lahko odraz predsodkov v učnih podatkih ali pridobljenih vzorcev obnašanja potrošnikov – ne pa nujno algoritemsko nasilje.

4. Zaključek: predsodki, nepravičnost in škoda v digitalni družbi

Preidimo k sklepu. Uporaba drugih, bolj specifičnih izrazov v zvezi z razvojem, izvajanjem in delovanjem algoritmov umetne inteligence – v nasprotju s splošnim izrazom algoritemsko nasilje – se pri prihodnjem urejanju algoritmov in določanju odgovornosti za njihovo uporabo kaže kot koristnejša.

„Akt EU o umetni inteligenci“ (angl. *AI Act*; EU 2024) izraza algoritemsko nasilje ne vsebuje; uporablja izraze, kot so manipulacija, izkoriščanje, tveganje (npr. tveganje škode za zdravje in varnost, tveganje napačnih ali pristranskih odločitev na temelju umetne inteligence), škoda, pristranska obravnava, (ne)transparentnost (Miklavčič 2023) itd. Med prakse oz. načine delovanja ter uporabe umetne inteligence, ki jih zakon prepoveduje, spadajo subliminalna manipulacija in načrtno

izkoriščanje človeške ranljivosti (npr. uporaba manipulativnih ali zavajajočih tehnik, ki spodbujajo ali ovirajo posameznikovo avtonomijo, odločanje ali svobodno izbiro tako, da se ljudje teh tehnik ne zavedajo), povzročanje fizične ali psihične škode oz. škode za zdravje in varnost, ki bi negativno vplivala na temeljne pravice, družbeni nadzor ipd. Med drugim Akt v okviru svojih izhodišč določa naslednje: »Raznolikost, nediskriminacija in pravičnost pomeni, da se sistemi UI razvijajo in uporabljajo na način, ki vključuje različne akterje in spodbuja enakopraven dostop, enakost spolov in kulturno raznolikost, hkrati pa se preprečujejo diskriminatorni učinki in nepravilni predsodki, ki jih pravo Unije ali držav članic prepoveduje.« (EU 2004) Taka so torej bolj določna izhodišča o tveganjih in nevarnosti diskriminacije.

V okviru zagovora pojma algoritemskega nasilja je bila ena od podmen vezana na vidik birokratizacije, vključno s postopki vrednotenja obsega oz. učinkov dela – negativni vplivi so označeni kot nasilje. „Akt EU o UI“ je bolj določen in jasno izpostavlja, zakaj gre pri takšni rabi UI za visoka tveganja in kakšna so ta tveganja. Določa, da sistemi

»UI, ki se uporabljajo pri zaposlovanju, upravljanju delavcev in dostopu do samozaposlitve, zlasti za zaposlovanje in izbor oseb, za sprejemanje odločitev, ki vplivajo na delovne pogoje, napredovanje ali prenehanje pogodbenih delovnih razmerij, za dodeljevanje nalog na podlagi individualnega vedenja, osebnih lastnosti ali značilnosti ter za spremljanje ali ocenjevanje oseb v z delom povezanih pogodbenih razmerij.«

predstavljajo sisteme visokega tveganja – še posebej zato, ker lahko takšni sistemi ohranjajo vzorce diskriminacije iz preteklosti, ki so povezani s spolom, starostno skupino, invalidnostjo, rasnim ali etničnim poreklom, spolno usmerjenostjo idr., in njihova uporaba ogroža temeljne pravice do varstva podatkov in zasebnosti. Takšen pristop, ki tveganja, nedovoljene prakse ter ukrepe za njihovo zmanjševanje oz. kaznovanje poimenuje določneje, se zdi bolj smiselni kot govor o algoritemskem nasilju.

Reference

- Arendt, Hannah.** 1969. *On Violence*. New York: Harcourt Brace Jovanovich Publishers.
- Bauman, Zygmund.** 2006. *Moderna in holokavst*. Ljubljana: Študentska založba.
- Bellanova, Rocco, Kristina Irion, Katja Lindskov Jacobsen, Francesco Ragazzi, Rune Saugmann in Lucy Suchman.** 2021. Toward a Critique of Algorithmic Violence. *International Political Sociology* 15:121–150.
- Bostrom, Nick, in Eliezer Yudkowsky.** 2014. The ethics of artificial intelligence. V: Keith Frankish in William Ramsey, ur. *Cambridge Handbook of Artificial Intelligence*, 316–334. New York: Cambridge University Press.
- Dotson, Kristie.** 2011. Tracking epistemic violence, tracking patterns of silencing. *Hypatia* 26, št. 2:236–257.
- Dwyer, Philip.** 2022. *Violence*. New York: Oxford University Press.
- EU.** 2004. Akt o umetni inteligenci. Lex Europa. <https://eur-lex.europa.eu/legal-content/SL/TXT/HTML/?uri=CELEX:52021PC0206> (pridobljeno 23. 5. 2024)
- Fricker, Miranda.** 2003. Epistemic Injustice and a Role for Virtue in the Politics of Knowing, *Metaphilosophy*, 34 št. 1-2:54–73.

- — —. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.
- Galtung, Johan.** 1969. Violence, Peace, and Peace Research. *Journal of Peace Research* 6, št. 3:167–191.
- Huws, Ursula, in trije anonimni soavtorji_ice.** 2024. Algorithmic violence: Towards an interdisciplinary understanding of bullying in academia. *Work Organisation, Labour & Globalisation* 18, št. 1: 97–122. <https://doi.org/10.13169/workorglaboglob.18.1.0097>
- Kockelman, Paul.** 2020. The Epistemic and Performative Dynamics of Machine Learning Praxis. *Signs and Society* 8, št. 2:319–355. <https://doi.org/10.1086/708249>
- Mbalaka, Blessing.** 2023. Epistemically violent biases in artificial intelligence design: the case of DALL-E 2 and Starry AI. *Digital Transformation and Society* 2, št. 4:376–402.
- McKinnon, Rachel.** 2016. Epistemic Injustice. *Philosophy Compass* 11, št. 8: 437–446.
- Miklavčič, Jonas.** 2021. Zaupanje in uspešnost umetne inteligence v medicini. *Bogoslovni vestnik* 81, št. 4:935–946.
- — —. 2023. Ideal transparentnosti v digitalni dobi. *Bogoslovni vestnik* 83, št. 4:825–838.
- Nixon, Rob.** 2011. *Slow Violence and the Environmentalism of the Poor*. Cambridge, MA: Harvard University Press.
- Onuoha, Mimi.** 2018. Algorithmic Violence. GitHub. <https://github.com/MimiOnuoha/On-Algorithmic-Violence> (pridobljeno 21. 7. 2024).
- Raaijmakers, Stephan.** 2022. *Deep learning for natural language processing*. Shelter Island, NY: Manning.
- Safransky, Sara.** 2019. Geographies Of Algorithmic Violence: Redlining the Smart City. *International Journal of Urban and Regional Research* 44, št. 2:200–218.
- Spivak, Gayatri.** 1998. Can the subaltern speak? V: Cary Nelson in Lawrence Grossberg, ur. *Marxism and the interpretation of culture*, 271–313. Urbana: University of Illinois Pres
- Strahovnik, Vojko.** 2018. Spoznavna (ne)pravičnost, krepost spoznavne ponižnosti in monoteizem. *Bogoslovni vestnik* 78, št. 2:299–311.
- — —. 2023. Etični in teološki izzivi velikih jezikovnih modelov. *Bogoslovni vestnik* 83, št. 4:839–852. <https://doi.org/10.34291/BV2023/04/Strahovnik>
- Strahovnik, Vojko, Jonas Miklavčič in Mateja Centa Strahovnik.** 2020. Etični vidiki uporabe algoritemskega odločanja in ostalih sistemov UI v času pandemij oz. izrednih razmer. *Bogoslovni vestnik*, 80, št. 2:321–334. <https://doi.org/10.34291/BV2020/02/Strahovnik>
- Weidinger, Laura, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving in Iason Gabriel.** 2022. Taxonomy of Risks posed by Language Models. V: *2022 ACM Conference on Fairness, Accountability, and Transparency (FACCT '22)*, 214–229. New York, NY: Association for Computing Machinery.
- WHO.** 1996. *WHO Global Consultation on Violence and Health. Violence: A Public Health Priority*. Ženeva: World Health Organization.
- — —. 2002. *The world report on violence and health*. Uredili Etienne G.Krug, Linda L. Dahlberg, James A. Mercy, Anthony B. Zwi in Rafael Lozano. Ženeva: World Health Organization.
- Žalec, Bojan.** 2023. Ali je umetna inteligenca inteligenca v pravem pomenu besede? Vprašanje psihičnih značilnosti in splošnost. *Bogoslovni vestnik* 83, št. 4:813–823. <https://doi.org/10.34291/BV2023/04/Zalec>