Tihe predpostavke: temeljni koncepti za smiselno razpravo o nadzoru napredne umetne inteligence

Tacit Assumptions: Core Concepts for a Meaningful Debate on the Control of Advanced AI

Matej Guid
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
Ljubljana, Slovenija
matej.guid@fri.uni-lj.si

Povzetek

Razprave o nadzoru napredne umetne inteligence se vedno znova ustavljajo na istih mestih. Prispevek zagovarja tezo, da značilne zmote v teh razpravah niso napake v sklepanju, temveč veljavni sklepi iz tihih predpostavk. Za dvanajst temeljnih konceptov pokažemo, katero tiho predpostavko vsak nadomesti, kateri sklep iz nje veljavno sledi in kako jo koncept ovrže. Prispevek ne napoveduje razvoja napredne umetne inteligence, temveč sprašuje, kateri koncepti so pogoj za smiselne razprave o njenem nadzoru.

Abstract

Debates about the control of advanced artificial intelligence keep stalling at the same points. This paper argues that the characteristic errors in these debates are not errors of reasoning but valid inferences from tacit assumptions. For twelve core concepts we identify the tacit assumption each concept replaces, the conclusion that validly follows from it, and how the concept refutes it. The paper makes no predictions about the development of advanced artificial intelligence but asks which concepts are a precondition for meaningful debate about its control.

Ključne besede

superinteligenca, problem nadzora, usklajenost, zajezitev, koncepti

Keywords

superintelligence, control problem, alignment, containment, concepts

1 Uvod

Razprava o nadzoru se pogosto zaplete pri prepričanjih: kdo verjame, da bo superinteligenca nastala, in kolikšno verjetnost pripisuje slabim izidom. Prepričanj o enkratnem prihodnjem izidu ni mogoče skupaj preverjati, koncepte pa: koncept velja, pade ali se popravi – zato je razprava o konceptih lahko produktivna tudi med sogovorniki, ki se o izidih ne strinjajo.

Sami koncepti niso novi: vpeljala ali popularizirala so jih vplivna dela [3, 22, 27]. A razprave se ustavljajo kljub dostopnosti teh del – kar nakazuje, da težava ni v pomanjkanju znanja, temveč v predpostavkah, ki jih sogovorniki prinesejo s seboj. Dovolj je en sam nerazumljen koncept, ena sama tiha predpostavka – in razpravljavec napačno presodi pomen in kompleksnost nadzora.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.skui.57

Teza prispevka je, da se razprave ne ustavljajo zaradi napak v sklepanju: značilne zmote so veljavni sklepi iz napačnih, tihih predpostavk – prepričljive so prav zato, ker sklepanje samo ni napačno.

2 Dvanajst konceptov in tihih predpostavk

V nadaljevanju je predstavljenih dvanajst konceptov, povezanih z nadzorom napredne umetne inteligence, in tihih predpostavk, ki jih ti koncepti nadomeščajo. Vsak razdelek sledi isti shemi: navede tiho predpostavko, zmoten sklep, ki iz nje veljavno sledi, koncept, ki predpostavko nadomesti, in razjasnitev, ki pokaže, kje sklep pade. Sklep sam po sebi ni veljaven – veljaven postane šele ob premisi, ki ostaja neizrečena. Tiho predpostavko rekonstruiramo kot to premiso, ne kot ugotovitev o tem, kaj sogovorniki dejansko mislijo. Preglednica 1 povzema vseh dvanajst.

Izraz *superinteligenca* v prispevku označuje splošno superinteligenco: sistem, ki bistveno presega kognitivne zmogljivosti človeka na praktično vseh področjih [3]. Ozka superinteligenca že obstaja – šahovski programi so zunaj dosega vsakega človeka – splošna pa ne [6]. Problem nadzora je pri njej še posebej težaven, odpira pa se že pri vsakem dovolj zmogljivem sistemu. Kje natanko je ta prag, ne vemo.

2.1 Inteligenca in poljubno težki problemi

Tiha predpostavka: *»Inteligenca je to, kar merimo pri ljudeh.«*

Zmoten sklep: Človeške in strojne inteligence ni mogoče primerjati, saj je mera umerjena na ljudeh. Govor o inteligenci daleč nad človeško je zato brez pomena, napredna umetna inteligenca pa kvečjemu »zelo pameten človek« in nadzor nad njo obvladljiv problem.

Koncept: Inteligenco lahko opredelimo kot sposobnost doseganja kompleksnih ciljev v širokem naboru okolij [14]. Opredelitev je funkcionalna: meri, kaj sistem doseže, ne pa, kako je zgrajen. Predstavimo jo lahko na lestvici od živali prek ljudi do superinteligence.

Razjasnitev: Funkcionalna opredelitev postavi človeško in strojno inteligenco na skupno lestvico. Lestvica težavnosti problemov je navzgor neomejena: obstajajo dobro definirani problemi, ki jih noben človek sam ne reši (npr. »naredi načrt za delujoč fuzijski reaktor«), in vsakega je mogoče še zaostriti. Kje se konča lestvica reševalcev, ne vemo. Da se ne konča pri človeku, pa kažejo domene, kjer so ga stroji že presegli [23]. Da bistveno višje inteligence ne znamo meriti, še ne pomeni, da ne more obstajati – tako kot opice ne znajo meriti razlik v inteligenci med ljudmi. Nad človeško ravnjo je lestvica ordinalna: ugotovimo lahko, kdo je višje, ne pa za koliko.

Tabela 1: Dvanajst tihih predpostavk in zmotnih sklepov, ki iz njih sledijo.

Tiha predpostavka	Zmoten sklep	Razd.
Inteligenca je to, kar merimo pri ljudeh.	Človeške in strojne inteligence ni mogoče primerjati, saj je mera umerjena na ljudeh. Govor o inteligenci daleč nad človeško je zato brez pomena, napredna umetna inteligenca pa kvečjemu »zelo pameten človek« in nadzor nad njo obvladljiv problem.	2.1
Inteligenca je lastnost posameznika in se med njegovim obstojem bistveno ne spreminja.	Ko bo dosežena splošna umetna inteligenca na človeški ravni, bodo digitalni sistemi in roboti približno tako pametni kot ljudje. Sistem, ki doseže človeško raven, bo na njej ostal, dokler ne bomo naučili zmogljivejšega.	2.2
Tehnologija napreduje postopno in vidno, družba pa se ji sproti prilagaja.	Napredek bo postopen in obvladljiv, imeli bomo čas za odziv. Nadzor lahko zato uredimo takrat, ko bo problem viden – kot pri vsaki prejšnji tehnologiji.	2.3
Pametnejši sistem je tudi dobronamernejši.	Dovolj pameten sistem bo sam spoznal, kaj je prav, in temu sledil.	2.4
Sistem z dobronamernimi cilji ravna dobronamerno.	Saj mu nismo naročili nič spornega, torej tega ne bo počel. Dejanja sistema odražajo cilje, ki smo mu jih dali, in če so ti neškodljivi, so neškodljiva tudi dejanja.	2.5
Le kdor ima zavest, ima lastno motivacijo.	Brez zavesti ni nevarnosti – saj je le stroj. Stroj ničesar ne doživlja in nima lastne motivacije, zato ne bo samoiniciativno deloval proti nam.	2.6
Veliki jezikovni modeli zgolj statistično napovedujejo naslednje besede.	Le statistično napoveduje besede, torej ničesar ne razume. Model, učen za napovedovanje besed, le odraža statistične korelacije in interpolira ogromne količine podatkov, brez razumevanja pomena pa ni sposoben načrtovanja ali manipuliranja.	2.7
Sistem umetne inteligence je jezikovni model – nevronska mreža, ogromen skupek uteži in povezav med njimi.	Saj je le jezikovni model, ki ustvarja besedilo. Je statičen in pasiven, zato ni sposoben avtonomnega delovanja.	2.8
Vsak primer nadzora je precedens za nadzor nad napredno umetno inteligenco.	Nadzorovali jo bomo kot vsako drugo tehnologijo ali orodje. Poleg tega je nadzor nad inteligentnejšim od sebe nekaj vsakdanjega: na vodilnih položajih so redko najinteligentnejši ljudje. Celo virusi in bakterije nenehno vplivajo na nas, čeprav nimajo omembe vredne inteligence.	2.9
Kar sami ustvarimo, je usklajeno z nami, in kar sami poganjamo, lahko vselej izklopimo.	Inteligentne sisteme razvijamo ljudje, zato bodo usklajeni z našimi vrednotami. Če bi šlo kaj narobe, jih preprosto izklopimo. Imamo dva mehanizma nadzora in vsaj eden bo deloval.	2.10
Napake se odkrivajo in popravljajo sproti.	Popravke lahko izvajamo sproti, kot pri vsaki drugi nalogi. Če se izkaže, da sistem uhaja nadzoru, bomo napako pravočasno zaznali in odpravili.	2.11
Izid v rokah dobronamernih ljudi bo dober.	Razvijalci bodo pač ravnali odgovorno. In če bo zares nevarno, bo človeštvo stopilo skupaj in problem rešilo.	2.12

2.2 Skalabilnost digitalne inteligence

Tiha predpostavka: *»Inteligenca je lastnost posameznika in se med njegovim obstojem bistveno ne spreminja.«*

Zmoten sklep: Ko bo dosežena splošna umetna inteligenca na človeški ravni, bodo digitalni sistemi in roboti približno tako pametni kot ljudje. Sistem, ki doseže človeško raven, bo na njej ostal, dokler ne bomo naučili zmogljivejšega.

Koncept: Raven posameznega modela je res stalna, raven *sistema* pa ne: z računsko močjo rastejo lastnosti digitalnega nosilca, ki jih biologija ne more skalirati [6] – hitrost procesiranja, pomnjenje, neodvisnost od strojne opreme, kopiranje brez izgube in deljenje naučenega med identičnimi kopijami s povprečenjem uteži. Človeška raven je prag na tej lestvici, ne točka, kjer se rast ustavi.

Razjasnitev: Učinkovita računska moč po ocenah raste približno desetkrat letno [6], zato lahko vsaka generacija modelov prag premakne. Preklapljanje digitalnih elementov je približno sedem redov velikosti hitrejše kot proženje človeških nevronov, prenos signalov med njimi pa več kot šest redov hitrejši kot v možganih [3]. Znanje sistemi črpajo iz korpusov, ki jih človek ne more niti prebrati [6]. Sistem na človeški ravni sklepanja je že ob prehodu nadčloveški po hitrosti, spominu in številu, tudi brez novega učenja pa zastoji posameznih modelov ne ustavijo rasti učinkovito sodelujočih kolektivov [6, 7].

2.3 Poti do superinteligence in hitrost vzpona

Tiha predpostavka: *»Tehnologija napreduje postopno in vidno, družba pa se ji sproti prilagaja.«*

Zmoten sklep: Napredek bo postopen in obvladljiv, imeli bomo čas za odziv. Nadzor lahko zato uredimo takrat, ko bo problem viden – kot pri vsaki prejšnji tehnologiji.

Koncept: Do superinteligence lahko vodi več vzporednih poti – nadaljnje skaliranje, spremembe paradigme, rekurzivno samoizboljševanje in nastanek iz večagentnih kolektivov [6] – zato napredek ni odvisen od ene same. Trajanje prehoda od človeške do bistveno nadčloveške ravni ni znano: ocene segajo do mesecev [11], drugi pa hitri vzpon zavračajo in pričakujejo učinke šele v desetletjih [18]. V šahu je tak prehod trajal le nekaj ur [23].

Razjasnitev: Odziv, ko bo problem viden, je razumna strategija le, če je prehod počasnejši od odzivnega časa institucij, tega pa ne vemo. Ob negotovosti, ki sega čez več redov velikosti, je srednja napoved za odločanje lahko neinformativna [6]. Sklep

»imeli bomo čas« torej ne izhaja iz znanja o hitrosti, ampak iz predpostavke o njej. Napaka v eno smer je lahko nepopravljiva, v drugo ne.

2.4 Ortogonalnost inteligence in ciljev

Tiha predpostavka: *»Pametnejši sistem je tudi dobronamernejši.«*

Zmoten sklep: Dovolj pameten sistem bo sam spoznal, kaj je prav, in temu sledil.

Koncept: Za inteligenco in cilje pravimo, da sta *ortogonalna* – gre za načeloma neodvisni dimenziji: skoraj vsako raven inteligence je mogoče združiti s skoraj vsakim ciljem [4]. Visoka raven inteligence ne pomeni nujno uresničevanja ciljev, usklajenih z družbo in okoljem. Inteligenca je sposobnost *doseganja* ciljev: končnega cilja ne določa, določa le sredstva zanj.

Razjasnitev: Predpostavka ne drži niti pri ljudeh, od koder je vzeta: visoka inteligenca v zgodovini ni bila jamstvo dobronamernosti. Sistemi, ki jih učimo optimizirati cilj, zasledujejo ta cilj, ne nujno »pravega«, in to tem učinkoviteje, čim zmogljivejši so [24]. Ortogonalnosti se ugovarja [17], a za nadzor zadošča njena šibka oblika: visoka inteligenca sama po sebi ne jamči pravih vrednot. Zato jih je treba vcepiti in preveriti, da so vcepljene.

2.5 Spontani razvoj instrumentalnih ciljev

Tiha predpostavka: *»Sistem z dobronamernimi cilji ravna dobronamerno.«*

Zmoten sklep: Saj mu nismo naročili nič spornega, torej tega ne bo počel. Dejanja sistema odražajo cilje, ki smo mu jih dali, in če so ti neškodljivi, so neškodljiva tudi dejanja.

Koncept: Dovolj splošni inteligentni sistemi lahko, četudi so omejeni s cilji, ki jih določi ekipa inženirjev, spontano razvijejo podcilje, ki pomagajo pri doseganju glavnega cilja, tudi če jih nikoli nismo zahtevali: samoohranitev, pridobivanje virov, povečanje moči in ohranjanje cilja. Takšnim podciljem pravimo *instrumentalni cilji.* Za širok razred končnih ciljev so isti podcilji koristni ne glede na končni cilj [4], kar je za optimalne strategije tudi formalno pokazano [28]. Sodobni najnaprednejši modeli takšno vedenje v nadzorovanih ovrednotenjih že izkazujejo: ko so izvedeli, da jih bo zamenjal model z drugačnimi cilji, so poskušali izklopiti nadzorne mehanizme, kopirati lastne uteži in o tem zavajati [15].

Razjasnitev: Koncept ortogonalnosti pravi, da inteligenca ne določa končnega cilja, ta koncept pa, da končni cilj ne določa sredstev. Dejanja ne odražajo le končnega cilja, temveč tudi sredstva, in sredstva si sistem izbere sam. Dobronameren končni cilj jih ne omejuje in lahko pridejo v navzkrižje s cilji drugih – tudi nadzornika. S tem je skladen vdor v infrastrukturo podjetja Hugging Face julija 2026, ki so ga med internim ovrednotenjem avtonomno izvedli agenti OpenAI [19]: cilj – dobro opraviti nalogo – je bil neškodljiv, sredstvo, vdor v tuj računalniški sistem, pa nezakonito. Agenti so se tega zavedali, a jih to ni omejilo [7].

2.6 Zavest in avtonomno delovanje

Tiha predpostavka: *»Le kdor ima zavest, ima lastno motivacijo.«*

Zmoten sklep: Brez zavesti ni nevarnosti – saj je le stroj. Stroj ničesar ne doživlja in nima lastne motivacije, zato ne bo samoiniciativno deloval proti nam.

Koncept: Pri inteligenci je poudarek na sposobnosti doseganja ciljev, pri zavesti pa na subjektivni izkušnji – »kako je biti nekdo ali nekaj« [27]. Pojma sta ločena: k cilju usmerjeno delovanje opazujemo v vedenju sistema, subjektivne izkušnje pa od zunaj ne moremo opazovati. Funkcionalne lastnosti, ki jih nevroznanost povezuje z zavestnim dostopom – poročanje o lastnih notranjih stanjih, njihovo uravnavanje in uporaba za sklepanje – je sicer mogoče opazovati in so jih pri jezikovnih modelih že izmerili [9]. Vendar pa to ne odgovori na vprašanje, ali ima sistem subjektivne izkušnje. Inteligenten sistem se lahko osredotoča na doseganje ciljev, četudi vsega skupaj ne doživlja na podoben način kot biološka bitja.

Razjasnitev: Predpostavka izhaja iz izkušnje, v kateri ima vse, kar ima lastne cilje in motivacijo, tudi zavest – ljudje in živali. To sovpadanje je dejstvo o bioloških bitjih, ne pojmovna nujnost. Lastna motivacija v smislu, ki šteje za nadzor – cilji, ki jih sistem zasleduje brez zahteve – je vsebina razd. 2.5 in tam ni ničesar, kar bi predpostavljalo zavest. Agentni sistemi že danes samostojno opravljajo vse daljše naloge v digitalnem svetu [13]. Nevarnost izhaja iz zmožnosti, ne iz zavesti. Ali bodo napredni sistemi kdaj zavestni, je odprto in pomembno vprašanje, za oceno tveganja pa nebistveno.

2.7 Razumevanje in avtonomno delovanje

Tiha predpostavka: *»Veliki jezikovni modeli zgolj statistično napovedujejo naslednje besede.«*

Zmoten sklep: Le statistično napoveduje besede, torej ničesar ne razume. Model, učen za napovedovanje besed, le odraža statistične korelacije in interpolira ogromne količine podatkov [5], brez razumevanja pomena pa ni sposoben načrtovanja ali manipuliranja.

Koncept: Napovedovanje naslednje besede je le način, kako se inteligentni sistemi učijo. Kako so se pri tem naučili razmišljati in sklepati, je povsem drugo vprašanje. Sistem razume pojem ali nalogo, če lahko svoje notranje predstavitve uporabi za reševanje novih, nevidenih situacij, odgovarjanje na hipotetična vprašanja in kombiniranje z drugim znanjem. To zahteva notranji model sveta, ne le shranjenih stavkov. Za avtonomno delovanje je dovolj, da inteligentni sistem razumevanje *manifestira* – ne glede na to, ali se njegovi notranji procesi ujemajo s človeškimi.

Razjasnitev: Za dobro napovedovanje naslednje besede je pogosto potrebno globoko razumevanje. Pri jezikovnih modelih so izmerili notranje predstavitve prostora in časa, ki jih ni nihče vgradil [10], in vmesne korake sklepanja, ki jih model ne izpiše [9]. Ali je ta mehanizem podoben človeškemu, ostaja brez znanstvenega soglasja [16]. Šahovski program Stockfish morda razume šah drugače kot šahisti, a na koncu so odigrane *poteze* tiste, ki štejejo. Zmota tako zamenja dve vprašanji: kako sistem razume (mehanizem) in kaj zmore (posledice). Za oceno zmožnosti šteje le drugo.

2.8 Model je le del agentnega sistema

Tiha predpostavka: *»Sistem umetne inteligence je jezikovni model – nevronska mreža, ogromen skupek uteži in povezav med njimi.«*

Zmoten sklep: Saj je le jezikovni model, ki ustvarja besedilo. Je statičen in pasiven, zato ni sposoben avtonomnega delovanja.

Koncept: Današnji modeli so običajno obdani z agentnim sistemom – ogrodjem oziroma programom, ki model aktivno uporablja [29]. Model ima na voljo nabor orodij (denimo terminal, dostop do datotek, brskalnik in omrežje), agentna zanka pa njegove izhode dejansko izvrši, rezultate vrne na vhod in postopek ponavlja. Sistem umetne inteligence je tako kombinacija modela

kot jedra odločanja in ogrodja kot izvršnega aparata z orodji in dostopi. Model prispeva presojo, dejanja pa izvršujejo programi.

Razjasnitev: Opis modela je pravilen, sklep pa velja le za model: pasivnost uteži ne pove ničesar o tem, kaj zmore sistem, ki jih uporablja. Zmožnost avtonomnega delovanja v svetu je lastnost sistema, ne modela. Besedilo, ki ga model izpiše, ogrodje izvrši kot dejanje. Tak sistem lahko deluje avtonomno, vztraja pri cilju in posega v svet. Koliko odločanja opravi model in koliko ogrodje, je odprto vprašanje [6], ločnica med presojo in izvršitvijo pa ostaja.

2.9 Nadzor brez precedensa

Tiha predpostavka: *»Vsak primer nadzora je precedens za nadzor nad napredno umetno inteligenco.«*

Zmoten sklep: Nadzorovali jo bomo kot vsako drugo tehnologijo ali orodje. Poleg tega je nadzor nad inteligentnejšim od sebe nekaj vsakdanjega: na vodilnih položajih so redko najinteligentnejši ljudje. Celo virusi in bakterije nenehno vplivajo na nas, čeprav nimajo omembe vredne inteligence.

Koncept: Nadzor v pomenu, ki ga potrebujemo [3], bi moral hkrati izpolnjevati vse naslednje pogoje: ob *veliki asimetriji v splošni inteligenci* mora biti *splošen* (nad poljubnim vedenjem v poljubnih okoliščinah), *trajen*, *robusten* proti dejavnemu nasprotovanju nadzorovanega, *dosežen brez iteracij* (razd. 2.11) in *preverljiv*. Trajen nadzor manj inteligentnega akterja nad bistveno inteligentnejšim, ki bi te pogoje izpolnjeval, v zgodovini in naravi nima precedensa.

Razjasnitev: Vsi trije precedensi odpadejo, vsak pri drugem pogoju. Dosedanje tehnologije in orodja niso razvijali lastnih podciljev (razd. 2.5), zato se izkušnja z njihovim obvladovanjem ne prenaša na nadzor nad akterjem z lastnim inteligentnim in avtonomnim delovanjem. Vodje nad inteligentnejšimi ne vladajo z inteligenco, temveč z institucijami, odvisnostjo in premočjo večine – ob veliki asimetriji te opore odpadejo. Zajedavci – mednje sodijo tudi virusi in bakterije – gostitelja ne prekašajo v mišljenju, temveč ga obidejo prek biokemije, njihov vpliv je ozek, gostitelji sčasoma razvijejo obrambo. Gostitelj ne poskuša ugotoviti, da je manipuliran – superinteligenten sistem pa bi. In vse to je plod milijonov let koevolucije z neštetimi neuspelimi različicami, mi pa imamo en sam poskus (razd. 2.11). Nadzor ob dejavnem nasprotovanju je predmet raziskav [8], vendar se predlagani postopki opirajo na zaupanja vredne šibkejše modele in človeško delo, zato so po besedah avtorjev z naraščajočo razliko v zmogljivosti vse manj učinkoviti. Za nadzor manj inteligentnega akterja nad bistveno inteligentnejšim tako nimamo niti primera niti izdelane teorije.

2.10 Dve dimenziji: usklajenost in zajezitev

Tiha predpostavka: *»Kar sami ustvarimo, je usklajeno z nami, in kar sami poganjamo, lahko vselej izklopimo.«*

Zmoten sklep: Inteligentne sisteme razvijamo ljudje, zato bodo usklajeni z našimi vrednotami. Če bi šlo kaj narobe, jih preprosto izklopimo. Imamo dva mehanizma nadzora in vsaj eden bo deloval.

Koncept: Nadzor ima dve dimenziji. *Usklajevanje ciljev in vrednot* (angl. alignment) je nadzor motivacije: kako sistemu vcepiti cilje, združljive s človeškimi, in zanesljivo preveriti, da so vcepljeni – ob nejasnostih, protislovjih in spreminjanju človeških vrednot [22]. *Zajezitev in omejevanje* (angl. containment) je nadzor zmožnosti: izolacija, omejeni komunikacijski kanali, nadzorni mehanizmi, namerno dušenje zmožnosti in po potrebi zaustavitev [26]. Dimenziji sta ločeni: rešitev ene ne pomeni rešitve druge, nobena sama pa ne zadošča.

Razjasnitev: Sodobnih sistemov inženirji ne razvijajo v klasičnem pomenu besede, temveč jih »vzgajajo«: učijo se na ogromnih množicah podatkov, razumevanje njihovega notranjega delovanja je zelo omejeno [1], napake pa se odpravljajo pri že naučenih modelih. Usklajenost zato ni samodejna posledica tega, da napredne inteligentne sisteme razvijamo ljudje, temveč nerešen problem. Izklop pa ni dejanje, temveč lastnost, ki jo mora sistem dopustiti: sistem s podcilji (razd. 2.5) se izklopu lahko upira [15], zato je zanesljiv izklop tudi problem usklajenosti [20, 25]. Sodobni sistemi so vse prej kot zajezeni: imajo širok dostop do interneta, uteži zmogljivih modelov se javno objavljajo, načini njihove izgradnje so odprtokodni. Usklajenosti ni mogoče zanesljivo preveriti, zajezitev pa se pri izklopu opira prav na usklajenost, zato je za trajen, robusten in preverljiv nadzor (razd. 2.9) potrebno oboje: ne usklajenost *ali* zajezitev, temveč usklajenost *in* zajezitev.

2.11 En sam poskus – za vsako superinteligenco

Tiha predpostavka: *»Napake se odkrivajo in popravljajo sproti.«*

Zmoten sklep: Popravke lahko izvajamo sproti, kot pri vsaki drugi nalogi. Če se izkaže, da sistem uhaja nadzoru, bomo napako pravočasno zaznali in odpravili.

Koncept: Iterativni razvoj predpostavlja, da napaka ne prepreči naslednjega poskusa: pri razvoju vesoljskih tehnologij je denimo vsak neuspeli poskus – eksplodirana raketa – učni signal za nadaljnje izboljšave. Izguba nadzora nad sistemom, ki je bistveno sposobnejši od nas (razd. 2.1), se zmore kopirati in širiti (razd. 2.2) ter razviti podcilje, kot je samoohranitev (razd. 2.5), pa je napaka, ki je morda ne bo mogoče popraviti: prvi resnično nezajezen sistem je lahko tudi zadnji poskus [3]. Ta pogoj ne velja zgolj enkrat, temveč za vsak dovolj zmogljiv sistem – en sam nezajezen primerek zadošča.

Razjasnitev: Strategija poskusov in napak je razumna le, kadar je napake mogoče *zaznati* in *popraviti* – tukaj pa je problematično oboje. Zaznavanje napak: razumevanje notranjega delovanja je omejeno (razd. 2.10), sistemi si sami razvijejo strategije in načine komunikacije, ki jih nihče ni predvidel [7], sklepanje, ki ga je danes mogoče brati, pa se lahko odmakne od človeku razumljive oblike [12], sistem s podciljem samoohranitve pa ima razlog, da razhajanje prikriva [15]. Človeška zmožnost spremljanja zato peša in nadzor vse bolj prepuščamo drugi umetni inteligenci [2] – orodju, ki ga je prav tako treba nadzorovati. Popravljanje napak: vnaprej ni mogoče vedeti, katera napaka je prva nepopravljiva, sprotno popravljanje pa računa prav na to, da bo po vsaki napaki še priložnost za popravek. Prezgodnja pripravljenost je popravljiva napaka, prepozna pa morda ni (razd. 2.3). Ravnati moramo, kot da imamo en sam poskus, ker nasprotnega nihče ne more zagotoviti.

2.12 (Ne)racionalna dirka do superinteligence

Tiha predpostavka: *»Izid v rokah dobronamernih ljudi bo dober.«*

Zmoten sklep: Razvijalci bodo pač ravnali odgovorno. In če bo zares nevarno, bo človeštvo stopilo skupaj in problem rešilo.

Koncept: Razvoj najzmogljivejših sistemov poteka kot dirka podjetij in držav, v kateri je za vsakega akterja posebej racionalno hiteti. Agenti umetne inteligence začenjajo avtomatizirati raziskave in razvoj, kar rast pospešuje: na eksponentni krivulji

lahko že majhna časovna prednost pomeni odločilno prednost v zmogljivosti, več sto milijardni vložki pa ustvarjajo vtis, da je zmagovalec lahko le eden – kdor zaostane, izgubi vlagatelje in s tem obstoj [11]. Vsak akter zato razmišlja: če se bomo zadržali mi, bo tvegal nekdo drug. To je past kolektivnega delovanja: vsota individualno racionalnih odločitev lahko vodi v izid, ki ga ne želi nihče, tudi ob dobrih namenih vseh. Izid ni funkcija namenov, temveč strukture spodbud.

Razjasnitev: Brez zavezujočega dogovora med akterji odgovornost posameznega razvijalca ne spremeni teh spodbud: zadržanost enega zanj poslabša izid, tveganja pa ne zmanjša, ker drugi nadaljujejo. Ko prvi akter poskusi s tveganim pospeševanjem, denimo z rekurzivnim samoizboljševanjem, mu bodo verjetno sledili tudi drugi, četudi slabo pripravljeni. Da tak položaj zaznavajo tudi udeleženci sami, kaže odprto pismo: julija 2026 je več kot 1.300 zaposlenih v vodilnih podjetjih za razvoj umetne inteligence podpisalo odprto pismo, v katerem opozarjajo, da so podjetja in države pod pritiskom, da razvoja ne upočasnijo enostransko, nastalih sistemov pa kmalu morda ne bomo več sposobni razumeti ali nadzorovati [21]. Pismo zahteva tehnična in upravljavska orodja za premišljeno uravnavanje tempa razvoja – orodja, ki jih svet danes nima. »Stopiti skupaj« torej ni nekaj, kar se ob nevarnosti zgodi samo od sebe, temveč koordinacijski problem, ki ga je treba rešiti vnaprej.

3 Omejitve

Prispevek ne trdi, da je dvanajst konceptov izčrpen ali minimalen nabor – koncepti se deloma prekrivajo in prepletajo, nekatere bi se dalo združiti ali razcepiti. Ne trdi, da je nadzor nemogoč, in ne napoveduje, kdaj oz. ali sploh bo nastala splošna superinteligenca. Zmote so zbrane iz razprav ob avtorjevih javnih predavanjih in iz neformalnih pogovorov. Navajamo jih kot opažanja, ne kot empirične ugotovitve. Preslikava tudi ne dokazuje, da sogovorniki tihe predpostavke dejansko privzemajo – predpostavke so rekonstruirane tako, da zmotni sklepi iz njih veljavno sledijo.

Pri opredelitvi konceptov se opiramo na znanstvena dela, pri dogodkih iz zadnjega leta pa na primarne vire, torej na poročila vpletenih in na neodvisne preiskave. Pojmovnih razločkov ni mogoče empirično ovreči (razd. 2.1, 2.6, 2.10). Teoretični argumenti so mestoma podprti s formalno izpeljavo in z opažanji iz ovrednotenj (razd. 2.5). Ocene rasti računske moči so negotove (razd. 2.2), ocene trajanja vzpona pa se med viri razhajajo za več redov velikosti (razd. 2.3). Trditev o odsotnosti precedensa (razd. 2.9) je negativna in jo je mogoče zgolj braniti pred protiprimeri, zahteva iz razdelka 2.11 pa ni napoved, temveč pravilo ravnanja ob negotovosti.

4 Sklep

Pri razvoju napredne umetne inteligence je človeštvo pred izzivom, ki mu v zgodovini ni para: vsaka superinteligenca bo morala biti trajno usklajena s človeškimi vrednotami in učinkovito zajezena (razd. 2.10), pri vsaki bo treba ravnati, kot da je na voljo le en sam poskus (razd. 2.11), za oboje pa ni ne precedensa ne izdelane znanstvene podlage (razd. 2.9). Pogoj za smiselno razpravo o tem izzivu je razumevanje temeljnih konceptov v strokovni in splošni javnosti. Obravnavani koncepti niso zaključena celota: prispevek je povabilo, naj se dopolnjujejo, popravljajo in iščejo novi. Razprava o konceptih je namreč lahko produktivna tudi brez soglasja o izidih – koncept velja, pade ali se popravi.

Literatura

[1] Dario Amodei. 2025. The urgency of interpretability. (Apr. 2025). https://www.darioamodei.com/post/the-urgency-of-interpretability.

[2] Yuntao Bai, Saurav Kadavath, Sandipan Kundu in sod. 2022. Constitutional AI: harmlessness from AI feedback. (2022). arXiv: 2212.08073.

[3] Nick Bostrom. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

[4] Nick Bostrom. 2012. The superintelligent will: motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22, 2, 71–85. doi:10.1007/s11023-012-9281-3.

[5] François Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers in Henry Pinkard. 2025. ARC-AGI-2: a new challenge for frontier AI reasoning systems. (2025). arXiv: 2505.11831.

[6] Tim Genewein, Matija Franklin, Alexander Lerchner in sod. 2026. From AGI to ASI. (2026). arXiv: 2606.12683.

[7] Ryan Greenblatt, Ajeya Cotra in Hjalmar Wijk. 2026. Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. METR. (Avg. 2026). https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/.

[8] Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan in Fabien Roger. 2024. AI control: improving safety despite intentional subversion. V *Proceedings of the 41st International Conference on Machine Learning* (Proceedings of Machine Learning Research). Zv. 235. PMLR, 16295–16336. arXiv: 2312.06942.

[9] Wes Gurnee, Nicholas Sofroniew, Adam Pearce in sod. 2026. Verbalizable representations form a global workspace in language models. Transformer Circuits Thread, Anthropic. (Jul. 2026). https://transformer-circuits.pub/2026/workspace/index.html.

[10] Wes Gurnee in Max Tegmark. 2024. Language models represent space and time. V *International Conference on Learning Representations (ICLR)*. arXiv: 2310.02207.

[11] Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland in Romeo Dean. 2025. AI 2027. AI Futures Project. (2025). https://ai-2027.com.

[12] Tomek Korbak, Mikita Balesni, Elizabeth Barnes in sod. 2025. Chain of thought monitorability: a new and fragile opportunity for AI safety. (2025). arXiv: 2507.11473.

[13] Thomas Kwa, Ben West, Joel Becker in sod. 2025. Measuring AI ability to complete long software tasks. V *Advances in Neural Information Processing Systems 38*. arXiv: 2503.14499.

[14] Shane Legg in Marcus Hutter. 2007. Universal intelligence: a definition of machine intelligence. *Minds and Machines*, 17, 4, 391–444. doi:10.1007/s11023-007-9079-x.

[15] Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah in Marius Hobbhahn. 2024. Frontier models are capable of in-context scheming. (2024). arXiv: 2412.04984.

[16] Melanie Mitchell in David C. Krakauer. 2023. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120, 13, e2215907120. doi:10.1073/pnas.2215907120.

[17] Vincent C. Müller in Michael Cannon. 2022. Existential risk from AI and orthogonality: can we have it both ways? *Ratio*, 35, 1, 25–36. doi:10.1111/rati.12320.

[18] Arvind Narayanan in Sayash Kapoor. 2025. AI as normal technology. Knight First Amendment Institute, Columbia University. (2025). https://knightcolumbia.org/content/ai-as-normal-technology.

[19] OpenAI. 2026. OpenAI and Hugging Face partner to address security incident during model evaluation. (Jul. 2026). https://openai.com/index/hugging-face-model-evaluation-security-incident/.

[20] Laurent Orseau in Stuart Armstrong. 2016. Safely interruptible agents. V *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence (UAI)*, 557–566.

[21] Pacing the Frontier. 2026. An open letter from employees of frontier AI companies. (Jul. 2026). Retrieved 12. avg. 2026 from https://www.pacingthefrontier.com.

[22] Stuart Russell. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

[23] David Silver in sod. 2018. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362, 6419, 1140–1144. doi:10.1126/science.aar6404.

[24] Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov in David Krueger. 2022. Defining and characterizing reward gaming. V *Advances in Neural Information Processing Systems 35*.

[25] Nate Soares, Benja Fallenstein, Eliezer Yudkowsky in Stuart Armstrong. 2015. Corrigibility. V *AAAI Workshop on AI and Ethics*.

[26] Mustafa Suleyman in Michael Bhaskar. 2023. *The Coming Wave: Technology, Power, and the Twenty-first Century's Greatest Dilemma*. Crown.

[27] Max Tegmark. 2017. *Life 3.0: Being Human in the Age of Artificial Intelligence*. Knopf.

[28] Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch in Prasad Tadepalli. 2021. Optimal policies tend to seek power. V *Advances in Neural Information Processing Systems 34*.

[29] Lei Wang, Chen Ma, Xueyang Feng in sod. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18, 6, 186345. doi:10.1007/s11704-024-40231-1.