Europska komisija

STRUČNA SKUPINA NA VISOKOJ RAZINI O UMJETNOJ INTELIGENCIJI

NACRT ETIČKIH SMJERNICA ZA POUZDANU UMJETNU INTELIGENCIJU

SAŽETAK

Radni dokument za savjetovanje dionika

NACRT ETIČKIH SMJERNICA

ZA POUZDANU UMJETNU INTELIGENCIJU

Stručna skupina na visokoj razini o umjetnoj inteligenciji
Nacrt etičkih smjernica za pouzdanu umjetnu inteligenciju

Europska komisija
Glavna uprava za komunikaciju

Kontakt Nathalie Smuha - koordinatorica Stručne skupine na visokoj razini o umjetnoj inteligenciji
elektronička pošta CNECT-HLG-AI@ec.europa.eu

Europska komisija
B-1049 Bruxelles

Dokument objavljen 18. prosinca 2018. na engleskom jeziku.

Ovaj radni dokument pripremila je Stručna skupina na visokoj razini o umjetnoj inteligenciji ne dovodeći u pitanje pojedinačna stajališta svojih članova o određenim točkama, i ne dovodeći u pitanje konačnu verziju dokumenta. Rad na ovom dokumentu će se nastaviti, a njegova konačna verzija bit će predstavljena u ožujku 2019., nakon savjetovanja dionika kroz Europski savez za umjetnu inteligenciju.

Više informacija o Stručnoj skupini na visokoj razini o umjetnoj inteligenciji dostupno je na internetu (https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence).

SAŽETAK

Ovaj je radni dokument nacrt etičkih smjernica za umjetnu inteligenciju koje je sastavila Stručna skupina na visokoj razini o umjetnoj inteligenciji Europske komisije (AI HLEG), konačna verzija kojih treba biti pripremljena do ožujka 2019.

Umjetna inteligencija jedna je od najsnažnijih transformativnih silnica našeg vremena te će sigurno izmijeniti strukturu društva. Velika je prigoda za povećanje blagostanja i rasta, čemu Europa mora težiti. U proteklom desetljeću, ostvaren je značajan napredak zahvaljujući dostupnosti golemih količina digitalnih podataka, moćnim računalnim arhitekturama te napretku u razvoju tehnika umjetne inteligencije kao što je strojno učenje. Znatna postignuća ostvarena zahvaljujući umjetnoj inteligenciji na području autonomnih vozila, zdravstvene zaštite, kućanskih/uslužnih robota, obrazovanja i kibersigurnosti, svakodnevno unaprjeđuju kvalitetu naših života. Nadalje, umjetna inteligencija ključna je za suočavanje s mnogim velikim izazovima pred kojima se svijet nalazi, primjerice globalno zdravlje i dobrobit, klimatske promjene, pouzdani pravni i demokratski sustavi, te drugi navedeni u ciljevima održivog razvoja Ujedinjenih naroda.

Umjetna inteligencija može donijeti velike koristi pojedincima i društvu, no otvara prostor i za određene rizike kojima treba primjereno upravljati. Budući da, u cjelini, koristi od umjetne inteligencije nadmašuju njezine rizike, moramo slijediti put koji će **maksimizirati koristi umjetne inteligencije, a istovremeno minimizirati rizike. Kako bismo ostali na pravom putu, potreban je pristup umjetnoj inteligenciji usmjeren na ljude, što će nas prisiliti da na razvoj i uporabu umjetne inteligencije ne gledamo kao na nešto što je samo sebi svrha, nego kao na nešto što za cilj ima unaprijediti dobrobit ljudi. Pouzdana umjetna inteligencija bit će naš putokaz,** s obzirom na to da će ljudi moći sigurno i u potpunosti ostvariti koristi umjetne inteligencije samo ako će tu tehnologiju moći smatrati pouzdanom.

Pouzdana umjetna inteligencija ima **dvije sastavnice:** 1. treba poštovati temeljna prava, relevantne propise i osnovna načela i vrijednosti, jamčeći na taj način „**etičnu svrhu**“ i 2. treba biti **tehnički pouzdana i sigurna jer, čak i s dobrim namjerama, nedostatak tehnoloških znanja može izazvati nenamjernu štetu.**

Ove smjernice stoga utvrđuju okvir za **pouzdanu umjetnu inteligenciju:**

- **Poglavlje I.** objašnjava kako **osigurati etičnu svrhu umjetne inteligencije**, utvrđivanjem temeljnih prava, načela i vrijednosti s kojima treba biti usklađena.
- Na temelju tih načela, **Poglavlje II.** razrađuje **smjernice za ostvarenje** pouzdane umjetne inteligencije, u pogledu i etične svrhe i tehničke otpornosti. To je ostvareno navođenjem popisa zahtjeva koje pouzdana umjetna inteligencija treba ispuniti te pregledom tehničkih i netehničkih metoda koje se mogu rabiti u njezinoj provedbi.
- **Poglavlje III.** zatim **operacionalizira** te zahtjeve i nudi popis nekih od konkretnih mjerila za ocjenu pouzdanosti umjetne inteligencije. Taj se popis zatim prilagođava posebnim slučajevima upotrebe.

Za razliku od drugih dokumenata o etičnoj umjetnoj inteligenciji, ove smjernice nemaju za cilj ponuditi još jedan popis temeljnih vrijednosti i načela za umjetnu inteligenciju, nego radije ponuditi smjernice za njihovu konkretnu provedbu i operacionalizaciju u sustavima umjetne inteligencije. Smjernice su dane u tri razine apstrakcije, od najapstraktnije u Poglavlju I. (temeljna prava, načela i vrijednosti), do najkonkretnije u Poglavlju III. (popis mjerila za ocjenjivanje).

Ove su smjernice upućene svim **relevantnim dionicima koji razvijaju, uvode ili koriste umjetnu inteligenciju**, uključujući poduzeća, organizacije, istraživače, javne službe, institucije, pojedince i druge subjekte. U konačnoj verziji ovih smjernica ponudit će se mehanizam koji će dionicima omogućiti da ih

dobrovoljno podrže.

Važno je napomenuti da se ovim smjernicama ne želi nadomjestiti bilo koja vrsta oblikovanja politika ili reguliranja (čime će se baviti drugi dokument AI HLEG-a: Preporuke za politiku i ulaganja na području umjetne inteligencije, koje trebaju biti pripremljene do svibnja 2019.), niti im je cilj odvratiti od njihovog uvođenja. Nadalje, smjernice treba smatrati živim dokumentom koji je potrebno redovito ažurirati kako bi se zajamčila njegova trajna relevantnost u skladu s razvojem tehnologije i naših znanja o tehnologiji. Ovaj bi dokument stoga trebao biti polazište za raspravu o **„Pouzdanoj umjetnoj inteligenciji proizvedenoj u Europi“**.

Bez obzira na to što Europa svoj etični pristup umjetnoj inteligenciji može širiti samo ako je konkurentna na globalnoj razini, **etični pristup umjetnoj tehnologiji ključan je za omogućavanje odgovorne konkurentnosti** jer će stvoriti povjerenje kod korisnika i olakšati šire korištenje umjetne inteligencije. Cilj ovih smjernica nije gušiti inovacije na području umjetne inteligencije u Europi, nego iskoristiti etičnost kao nadahnuće za razvoj jedinstvene marke umjetne inteligencije čiji je cilj štititi i donijeti koristi i pojedincima i zajedničkom dobru. To će Europi omogućiti da se pozicionira kao predvodnik na području napredne, sigurne i etične umjetne inteligencije. Europski će građani moći u potpunosti ostvariti koristi umjetne inteligencije samo ako ona bude pouzdana.

Naposljetku, cilj je ovih smjernica izvan Europe **poticati promišljanja i rasprave** o etičnom okviru za umjetnu inteligenciju na **globalnoj razini**.

IZVRŠNE SMJERNICE

Svako poglavlje smjernica nudi naputke za ostvarivanje pouzdane umjetne inteligencije, upućene svim relevantnim dionicima koji razvijaju, uvode ili koriste umjetnu inteligenciju, kako je sažeto navedeno u nastavku:

Poglavlje I.: Ključne smjernice za osiguranje etične svrhe:

- Osigurati da je umjetna inteligencija **usmjerena na ljude**: umjetnu inteligenciju treba razvijati, uvoditi i koristiti s „**etičnom svrhom**“, koja će se temeljiti na osnovnim pravima, društvenim vrijednostima i etičkim načelima *dobročinstva* (činiti dobro), *neškodljivosti* (ne činiti loše), *poštivanja osobnosti ljudi, pravednosti* i *objašnjivosti*, te ih odražavati. To je ključno za rad na ostvarenju **pouzdane umjetne inteligencije**.
- Uzdati se u osnovna prava, etička načela i vrijednosti kako bi se vrednovali budući mogući učinci umjetne inteligencije na ljude i zajedničko dobro. Posvećivati **posebnu pozornost** situacijama koje uključuju **ranjive skupine** kao što su djeca, osobe s invaliditetom ili manjine te situacijama s **asimetričnošću moći ili informacija**, primjerice između poslodavaca i zaposlenika ili poduzeća i potrošača.
- Priznavati i biti svjestan činjenice da umjetna inteligencija, premda pojedincima i društvu donosi značajne koristi, može imati i negativan učinak. Ostati oprezan u odnosu na posebno problematična područja.

Poglavlje II.: Ključne smjernice za ostvarenje pouzdane umjetne inteligencije:

- Uključivati **zahtjeve za pouzdanu umjetnu inteligenciju od samog početka projektiranja**: odgovornost, upravljanje podacima, projektiranje za sve, upravljanje autonomijom umjetne inteligencije (ljudski nadzor), zabrana diskriminacije, poštovanje ljudske osobnosti, poštovanje privatnosti, otpornost, sigurnost, transparentnost.
- Razmotriti tehničke i netehničke metode kako bi se osigurala provedba tih zahtjeva u sustavu

umjetne inteligencije. Nadalje, uzimati te zahtjeve u obzir prilikom stvaranja tima koji će raditi na sustavu, samog sustava, okruženja za ispitivanje i potencijalnih primjena sustava.

- Pružiti, na jasan i proaktivan način, **informacije dionicima** (potrošačima, zaposlenicima, itd.) o sposobnostima i ograničenjima sustava umjetne inteligencije, što će im omogućiti da imaju realistična očekivanja. U tom je smislu ključno osigurati **sljedivost** sustava umjetne inteligencije.
- Učiniti pouzdanu umjetnu inteligenciju **dijelom organizacijske kulture** i pružiti informacije dionicima o tome na koji je način pouzdana umjetna inteligencija ugrađena u projektiranje i korištenje sustava umjetne inteligencije. Pouzdana umjetna inteligencija može se uključiti i u deontološke povelje ili kodekse ponašanja organizacija.
- Osigurati sudjelovanje i **uključenost dionika** u projektiranje i razvoj sustava umjetne inteligencije. Nadalje, osigurati **raznolikost** prilikom formiranja timova koji razvijaju, provode i ispituju proizvode.
- Težiti **olakšavanju mogućnosti provjere** sustava umjetne inteligencije, osobito u kritičnim kontekstima ili situacijama. Koliko je moguće projektirati sustav kako bi se omogućilo praćenje pojedinih odluka do različitih ulaznih čimbenika, podataka, prethodno obučenih modela, itd. Nadalje, definirati **metode obrazlaganja** sustava umjetne inteligencije.
- Osigurati poseban proces za **upravljanje odgovornošću**.
- Predvidjeti **osposobljavanje i obrazovanje**, te osigurati da upravitelji, razvojni programeri, korisnici i poslodavci budu osviješteni u pogledu pouzdane umjetne inteligencije i da su osposobljeni za nju.
- Imati na umu da mogu postojati bitne napetosti između različitih ciljeva (transparentnost može otvoriti vrata zloupotrebi; identificiranje i ispravljanje pristranosti mogu ugroziti zaštitu privatnosti). Komunicirati i dokumentirati te kompromise.
- Poticati istraživanja i inovacije kako bi se promicalo ostvarenje zahtjeva za pouzdanu umjetnu inteligenciju.

Poglavlje III.: Ključne smjernice za ocjenjivanje pouzdane umjetne inteligencije

- Donijeti **popis mjerila za ocjenjivanje** pouzdane umjetne inteligencije pri razvoju, uvođenju i upotrebi umjetne inteligencije i prilagoditi ga posebnim slučajevima upotrebe u kojima se sustav koristi.
- Imati na umu da popis mjerila za ocjenjivanje **nikad nije sveobuhvatan**, i da kod osiguranja pouzdane umjetne inteligencije nije riječ o ispunjavanju formalnih zahtjeva, nego o kontinuiranom procesu identificiranja zahtjeva, vrednovanja rješenja i osiguranja boljih ishoda kroz cijeli životni ciklus sustava umjetne inteligencije.

Ove smjernice dio su vizije koja obuhvaća pristup umjetnoj inteligenciji usmjeren na ljude, što će Europi omogućiti da postane globalno vodeći inovator u etičnoj, sigurnoj i naprednoj umjetnoj inteligenciji. Cilj im je olakšati i omogućiti **„pouzdanu umjetnu inteligenciju proizvedenu u Europi“**, što će unaprijediti dobrobit europskih građana.