Evropska komisija

STROKOVNA SKUPINA NA VISOKI RAVNI ZA UMETNO INTELIGENCO

OSNUTEK
ETIČNIH SMERNIC ZA ZAUPANJA VREDNO UMETNO INTELIGENCO

POVZETEK

Delovni dokument za posvetovanje z deležniki

OSNUTEK ETIČNIH SMERNIC

ZA ZAUPANJA VREDNO UMETNO INTELIGENCO

Strokovna skupina na visoki ravni za umetno inteligenco
Osnutek etičnih smernic za zaupanja vredno umetno inteligenco

Evropska komisija
Generalni direktorat za komuniciranje

Stik Nathalie Smuha – koordinatorka strokovne skupine na visoki ravni za umetno inteligenco
E-pošta CNECT-HLG-AI@ec.europa.eu

Evropska komisija
B-1049 Bruselj

Dokument je bil objavljen 18. decembra 2018 v angleščini.

Strokovna skupina na visoki ravni za umetno inteligenco je ta delovni dokument pripravila brez poseganja v stališča posameznih članov skupine o posameznih točkah in brez poseganja v končno različico dokumenta. Skupina bo ta dokument dodatno razvijala in njegovo končno različico predstavila marca 2019 po posvetovanju z deležniki prek evropskega zavezništva za umetno inteligenco.

 To je osnutek prvega prispevka strokovne skupine na visoki ravni za umetno inteligenco. Njegova končna različica bo Komisiji predstavljena marca 2019. Končna različica drugega prispevka – priporočil glede politike in naložb v zvezi z umetno inteligenco – bo predstavljena sredi leta 2019.

Več informacij o strokovni skupini na visoki ravni za umetno inteligenco je na voljo na spletu (https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence).

POVZETEK

Ta delovni dokument je osnutek etičnih smernic za umetno inteligenco, ki jih pripravlja strokovna skupina Evropske komisije na visoki ravni za umetno inteligenco in katerih končna različica bo predstavljena marca 2019.

Umetna inteligenca je ena najbolj preobrazbenih sil našega časa in bo gotovo spremenila tkivo družbe. Pomeni veliko priložnost za povečanje blaginje in rasti, za doseganje česar si mora Evropa prizadevati. V zadnjem desetletju je bil dosežen velik napredek zaradi razpoložljivosti obsežnih količin digitalnih podatkov, zmogljivih računalniških arhitektur in napredka tehnik umetne inteligence, kot je strojno učenje. Pomembne izboljšave, ki jih je omogočila umetna inteligenca glede avtonomnih vozil, zdravstva, hišnih/storitvenih robotov, izobraževanja ali kibernetske varnosti, vsak dan izboljšujejo kakovost našega življenja. Poleg tega je umetna inteligenca ključna pri obravnavi številnih velikih izzivov, s katerimi se spopada svet, kot so globalno zdravje in dobrobit, podnebne spremembe, zanesljivi pravni in demokratični sistemi ter drugi izzivi, ki so navedeni v ciljih trajnostnega razvoja Združenih narodov.

Umetna inteligenca je zmožna ustvariti izjemne koristi za posameznike in družbo, vendar tudi povzroča nekatera tveganja, ki bi jih bilo treba ustrezno obvladovati. Ker na splošno koristi umetne inteligence odtehtajo njena tveganja, moramo zagotoviti, da stopamo po poti **čim večjega povečevanja koristi umetne inteligence ob hkratnem zmanjševanju njenih tveganj**. Da ostanemo na pravi poti, **je potreben na človeka osredotočen pristop k umetni inteligenci,** zaradi katerega imamo ves čas v mislih, da razvoj in uporaba umetne inteligence nista sama sebi namen, ampak je njun cilj povečanje dobrobiti ljudi. **Naša zvezda vodnica bo zaupanja vredna umetna inteligenca**, saj bodo lahko imeli ljudje od umetne inteligence zanesljivo in v celoti koristi le, če bodo tehnologiji lahko zaupali.

Zaupanja vredna umetna inteligenca ima **dva elementa:** (1) spoštovati mora temeljne pravice, veljavne predpise ter osnovna načela in vrednote, da se zagotovi „**etični namen**“, ter (2) biti mora **tehnično trdna** in zanesljiva, saj utegne neobvladovanje tehnologije celo ob dobrih namenih povzročiti nenamerno škodo.

Te smernice torej določajo **okvir za zaupanja vredno umetno inteligenco:**

- **Poglavje I** obravnava **zagotavljanje etičnega namena umetne inteligence** z določitvijo temeljnih pravic, načel in vrednot, ki jih mora spoštovati.
- Iz teh načel so v **poglavju II** izpeljana **navodila za uresničevanje** zaupanja vredne umetne inteligence, ki obravnavajo etični namen in tehnično trdnost. Pri tem so navedene zahteve za zaupanja vredno umetno inteligenco, podan pa je tudi pregled tehničnih in netehničnih metod, ki se lahko uporabijo za njeno uvajanje.
- **Poglavje III** nato **operacionalizira** zahteve z navajanjem konkretnega, toda neizčrpnega seznama za ocenjevanje zaupanja vredne umetne inteligence. Ta seznam se nato prilagodi posameznim primerom uporabe.

V nasprotju z drugimi dokumenti, ki obravnavajo etično umetno inteligenco, namen smernic ni podati še enega seznama osnovnih vrednot in načel za umetno inteligenco, temveč predstaviti smernice za njihovo dejansko uvajanje in operacionalizacijo v sistemih umetne inteligence. Takšna navodila so podana na treh ravneh abstrakcije, od najabstraktnejših v poglavju I (temeljne pravice, načela in vrednote) do najkonkretnejših v poglavju III (seznam za ocenjevanje).

Te smernice so namenjene vsem **pomembnim deležnikom, ki razvijajo, uvajajo ali uporabljajo umetno inteligenco**, kar zajema podjetja, organizacije, raziskovalce, javne službe, institucije, posameznike in druge subjekte. V končni različici teh smernic bo predstavljen mehanizem, ki bo deležnikom omogočal, da jih prostovoljno podprejo.

Pomembno je, da te smernice niso namenjene kot nadomestek oblikovanja politik ali pravnega urejanja

(to bo obravnaval drugi prispevek strokovne skupine na visoki ravni za umetno inteligenco, tj. Priporočila glede politike in naložb, ki bodo objavljena maja 2019) in naj ne bi odvračale od njihove uvedbe. Poleg tega bi morali te smernice videti kot živ dokument, ki ga je treba skozi čas redno posodabljati, da se zagotovi stalna relevantnost ob razvoju tehnologije in našega razumevanja. Ta dokument bi moral biti torej izhodišče za razpravo o **„zaupanja vredni umetni inteligenci, izdelani v Evropi“**.

Evropa lahko svoj etični pristop k umetni inteligenci razširja samo, če je konkurenčna na globalni ravni, **etični pristop k umetni inteligenci pa je ključen, da se omogoči odgovorna konkurenčnost**, saj bo ustvarjal zaupanje uporabnikov in olajšal širše sprejemanje umetne inteligence. Namen teh smernic ni dušiti inovacije na področju umetne inteligence v Evropi, temveč uporabljati etiko kot navdih za razvoj edinstvene vrste umetne inteligence, ki bo varovala posameznike in skupno dobro ter jim koristila. Tako lahko Evropa doseže vodilni položaj na področju najnaprednejše, varne in etične umetne inteligence. Samo ob zagotovljenem zaupanju bodo evropski državljani deležni vseh koristi umetne inteligence.

Končno je cilj teh smernic, ki presega meja Evrope, **spodbujati premislek in razpravo** o etičnem okviru za umetno inteligenco na **globalni ravni**.

NAVODILA

Vsako poglavje smernic vsebuje navodila glede doseganja zaupanja vredne umetne inteligence, namenjena vsem pomembnim deležnikom, ki razvijajo, uvajajo ali uporabljajo umetno inteligenco, povzeta pa so spodaj:

Poglavje I: ključna navodila za zagotavljanje etičnega namena:

- Treba je zagotoviti, da je umetna inteligenca **osredotočena na človeka**: umetna inteligenca bi se morala razvijati, uvajati in uporabljati z **„etičnim namenom“** na osnovi temeljnih pravic, družbenih vrednot in etičnih načel *dobronamernosti* (delaj dobro), *neškodovanja* (ne škoduj), *avtonomije ljudi, pravičnosti* in *razložljivosti* ter jih tudi odražati. To je ključno pri prizadevanjih za **zaupanja vredno umetno inteligenco**.
- Opreti se je treba na temeljne pravice, etična načela in vrednote, da se za naprej ocenijo morebitni vplivi umetne inteligence na ljudi in skupno dobro. **Posebno pozornost** je treba nameniti situacijam, ki vključujejo **ranljivejše skupine**, na primer otroke, invalide ali manjšine, ali situacij z **asimetrijo moči ali informacij**, na primer med delodajalci in delojemalci ali podjetji in potrošniki.
- Priznavati in zavedati se je treba dejstva, da ima umetna inteligenca, ki prinaša bistvene koristi za posameznike in družbo, lahko tudi negativne učinke. Treba je budno spremljati kritična področja.

Poglavje II: ključna navodila za uresničevanje zaupanja vredne umetne inteligence:

- **Zahteve za zaupanja vredno** umetno inteligenco morajo biti vključene **od najzgodnejše faze načrtovanja**: odgovornost, upravljanje podatkov, načrtovanje za vse, upravljanje avtonomije umetne inteligence (človeški nadzor), nediskriminacija, spoštovanje človeške avtonomije, spoštovanje zasebnosti, trdnost, varnost, preglednost.
- Upoštevati je treba tehnične in netehnične metode za zagotavljanje uvedbe teh zahtev v sistem umetne inteligence. Poleg tega je treba te zahteve upoštevati pri sestavljanju ekipe za gradnjo sistema, samem sistemu, v preizkusnem okolju in potencialnih aplikacijah sistema.
- Jasno in proaktivno je treba posredovati **informacije deležnikom** (strankam, zaposlenim itd.) o zmogljivostih in omejitvah sistema umetne inteligence, kar omogoča realna pričakovanja. V zvezi s tem je ključno zagotavljanje **sledljivosti** sistema umetne inteligence.
- Treba je poskrbeti, da zaupanja vredna inteligenca postane **del kulture organizacije**, ter deležnikom posredovati informacije o tem, kako se zaupanja vredna umetna inteligenca uvede v načrtovanje in uporabo sistemov umetne inteligence. Zaupanja vredna inteligenca se lahko vključi tudi v

deontološke listine ali kodekse ravnanja organizacij.

- Treba je zagotoviti sodelovanje in **vključevanje deležnikov** v načrtovanje in razvoj sistemov umetne inteligence. Poleg tega je treba zagotavljati **raznolikost** pri sestavi ekip, ki izdelek razvijajo, uvajajo in preizkušajo.
- Prizadevati si je treba za **omogočanje revidiranja** sistemov umetne inteligence, zlasti v kritičnih okoljih ali situacijah. Sistem je treba v največji možni meri načrtovati tako, da je omogočeno sledenje posameznih odločitev po različnih vnosih: podatkih, naučenih modelih itd. Poleg tega je treba opredeliti **metode za razlaganje** sistema umetne inteligence.
- Zagotoviti je treba poseben postopek za **upravljanje odgovornosti**.
- Treba je predvideti **usposabljanje in izobraževanje** ter zagotoviti, da so vodje, razvijalci, uporabniki in delodajalci seznanjeni z zaupanja vredno umetno inteligenco in usposobljeni zanjo.
- Zavedati se je treba, da lahko obstajajo temeljna trenja med različnimi cilji (preglednost lahko odpre vrata zlorabi; prepoznavanje in odpravljanje pristranskosti je lahko v nasprotju z varstvom zasebnosti). O teh kompromisih je treba poročati in jih dokumentirati.
- Spodbujati je treba raziskave in inovacije za pospešitev razvoja zahtev za zaupanja vredno umetno inteligenco.

<u>Poglavje III: ključna navodila za ocenjevanje zaupanja vredne umetne inteligence</u>

- Sprejeti je treba **seznam za ocenjevanje** zaupanja vredne umetne inteligence pri njenem razvoju, uvajanju ali uporabi ter ga prilagoditi posameznemu primeru uporabe, v katerem se sistem uporablja.
- Zavedati se je treba, da seznam za ocenjevanje **ne bo nikoli izčrpen** ter da zagotavljanje zaupanja vredne umetne inteligence ni zgolj odkljukanje postavk na seznamu, temveč stalni proces opredeljevanja zahtev, ocenjevanja rešitev in zagotavljanja boljših izidov v celotnem življenjskem ciklu sistema umetne inteligence.

Ta navodila so del vizije, ki sprejema na človeka osredotočen pristop k umetni inteligenci, kar bo Evropi omogočilo, da postane globalno vodilni inovator na področju etične, varne in najnaprednejše umetne inteligence. Spodbudila in omogočila naj bi **„zaupanja vredno umetno inteligenco, izdelano v Evropi“**, kar bo povečalo dobrobit evropskih državljanov.