Comisia Europeană

GRUPUL DE EXPERȚI LA NIVEL ÎNALT PRIVIND INTELIGENȚA ARTIFICIALĂ

PROIECT

ORIENTĂRI ÎN MATERIE DE ETICĂ PENTRU O INTELIGENȚĂ ARTIFICIALĂ (IA) FIABILĂ

REZUMAT

Document de lucru pentru consultarea părților interesate

PROIECT DE ORIENTĂRI ÎN MATERIE DE ETICĂ

PENTRU O INTELIGENȚĂ ARTIFICIALĂ (IA) FIABILĂ

Grupul de experți la nivel înalt privind inteligența artificială
Proiect de orientări în materie de etică pentru o inteligență artificială fiabilă

Comisia Europeană
Direcția Generală Comunicare

Contact Nathalie Smuha - coordonator
E-mail CNECT-HLG-AI@ec.europa.eu

Comisia Europeană
B-1049 Bruxelles

Document publicat la 18 decembrie 2018, în limba engleză.

Prezentul document de lucru a fost elaborat de Grupul de experți la nivel înalt privind inteligența artificială (AI HLEG), fără a aduce atingere poziției individuale a membrilor săi cu privire la anumite aspecte și fără a aduce atingere versiunii finale a documentului. Lucrul la document va continua și în perioada următoare, iar versiunea finală a acestuia va fi prezentată în martie 2019, în urma consultării părților interesate prin intermediul Alianței europene în domeniul inteligenței artificiale.

Mai multe informații cu privire la Grupul de experți la nivel înalt privind inteligența artificială sunt disponibile online (https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence).

REZUMAT

Prezentul document de lucru constituie un proiect al orientărilor privind etica în materie de inteligență artificială elaborat de Grupul de experți la nivel înalt privind inteligența artificială (AI HLEG) al Comisiei Europene, a cărui versiune finală urmează să fie prezentată în martie 2019.

Inteligența artificială (IA) este una dintre forțele cu cel mai mare potențial transformator ale vremurilor noastre, care va conduce inevitabil la modificarea structurii societății. Inteligența artificială oferă o ocazie excelentă pentru stimularea prosperității și a creșterii economice, iar Europa trebuie să depună eforturi pentru realizarea acestor obiective. În ultimul deceniu au fost realizate progrese majore datorită disponibilității unui volum mare de date digitale, a unei arhitecturi de calcul puternice și a progreselor privind tehnicile AI, cum ar fi învățarea automată. Principalele progrese facilitate de AI în ceea ce privește dezvoltarea vehiculelor autonome, asistența medicală, roboții casnici/destinații prestării de servicii, educația sau securitatea cibernetică îmbunătățesc în fiecare zi calitatea vieții noastre. În plus, AI este esențială pentru abordarea multora dintre marile provocări cu care se confruntă lumea, cum ar fi sănătatea și bunăstarea la nivel mondial, schimbările climatice, sistemele juridice și democratice fiabile și altele, exprimate în cadrul obiectivelor de dezvoltare durabilă ale Organizației Națiunilor Unite.

AI poate genera beneficii enorme pentru indivizi și societate, dar comportă, totodată, anumite riscuri care ar trebui gestionate în mod corespunzător. Având în vedere că, în ansamblu, beneficiile inteligenței artificiale sunt mai mari decât riscurile, trebuie să ne asigurăm că alegem calea care **maximizează beneficiile oferite de inteligența artificiale, diminuând, în același timp, riscurile acesteia**. Pentru a ne asigura că rămânem pe calea cea bună, **este necesară o abordare a inteligenței artificiale centrată pe om** , care să ne oblige să nu pierdem din vedere faptul că dezvoltarea și utilizarea inteligenței artificiale nu ar trebui să fie considerate un mijloc în sine, ci ca având obiectivul de a spori bunăstarea ființei umane. Dezvoltarea unei **inteligențe artificiale fiabile va fi pentru noi Steaua călăuzitoare**, întrucât ființele umane nu vor putea să profite cu încredere și pe deplin de beneficiile oferite de inteligența artificială decât dacă pot avea încredere în tehnologie.

Inteligența artificială fiabilă are **două componente:** (1) ar trebui să respecte drepturile fundamentale, reglementările aplicabile și principiile și valorile fundamentale, asigurând un „**scop etic**” și (2) ar trebui să fie **solidă din punct de vedere tehnic** și fiabilă, deoarece, chiar și cu bune intenții, lipsa unei bune stăpâniri a tehnologiei poate provoca daune neintenționate.

Prin urmare, prezentele orientări stabilesc un **cadru pentru dezvoltarea unei inteligențe artificiale fiabile:**

- **Capitolul I** are ca obiect **asigurarea scopului etic al inteligenței artificiale**, prin stabilirea drepturilor, principiilor și valorilor fundamentale pe care ar trebui să le respecte aceasta.
- Pe baza acestor principii, **capitolul II** stabilește **orientări privind dezvoltarea** unei inteligențe artificiale fiabile, care abordează atât chestiunea scopului etic, cât și pe cea a solidității tehnice. Pentru aceasta, se enumeră cerințele care trebuie respectate pentru dezvoltarea unei inteligențe artificiale fiabile și se oferă o perspectivă de ansamblu asupra metodelor tehnice și fără caracter tehnic care pot fi utilizate în acest sens.
- În continuare, **capitolul III oferă o dimensiune operațională** cerințelor, prin furnizarea unei liste de evaluare concrete, care însă nu este exhaustivă, pentru inteligența artificială fiabilă. Apoi, această listă este adaptată la cazurile de utilizare specifice.

Prin urmare, spre deosebire de alte documente referitoare la etica în domeniul IA, prezentele orientări nu urmăresc să furnizeze o nouă listă a valorilor și principiilor fundamentale pentru IA, ci, mai degrabă, să ofere orientări privind integrarea acestor valori și principii în sistemele IA și operaționalizarea concretă a acestora. Prezentele orientări sunt structurate pe trei niveluri de abstractizare, de la cele mai abstracte, prevăzute în capitolul I (drepturi fundamentale, principii și valori), până la cele mai concrete,

prevăzute în capitolul III (lista de evaluare).

Orientările se adresează tuturor **părților interesate relevante care dezvoltă, implementează ori utilizează IA**, fie ele întreprinderi, organizații, cercetători, servicii publice, instituții, persoane fizice sau alte entități. În versiunea finală a acestor orientări se va propune un mecanism care să permită părților interesate să le aprobe în mod voluntar.

Este important faptul că aceste orientări nu sunt destinate să înlocuiască vreo formă de elaborare a politicilor sau de reglementare (aspecte care urmează să fie abordate în al doilea document al AI HLEG: recomandările în materie de politică și de investiții, preconizate pentru mai 2019) ori să descurajeze inițierea unor astfel de politici ori reglementări. În plus, orientările ar trebui să fie considerate un document evolutiv, care trebuie să fie actualizat periodic în decursul timpului pentru a se asigura relevanța sa continuă, deoarece atât tehnologia, cât și cunoștințele noastre în acest domeniu evoluează. Prin urmare, prezentul document ar trebui să constituie un punct de plecare pentru discuția privind „**O inteligență artificială fiabilă, creată în Europa**".

Europa își poate promova abordarea etică privind inteligența artificială doar dacă este competitivă la nivel mondial, însă o **abordare etică a inteligenței artificiale este esențială pentru a se permite o competitivitate responsabilă**, deoarece va genera încredere din partea utilizatorilor și va facilita o adoptare pe scară mai largă a inteligenței artificiale. Prezentele orientări nu sunt menite să împiedice inovarea în domeniul inteligenței artificiale în Europa, ci urmăresc să utilizeze etica drept sursă de inspirație pentru a se dezvolta o marcă unică de IA, care vizează să protejeze și să aducă beneficii atât cetățenilor, cât și binelui comun. Acest lucru permite Europei să se poziționeze ca lider în dezvoltarea unei inteligențe artificiale, sigure și etice, bazate pe tehnologie de vârf. Cetățenii europeni vor putea valorifica pe deplin de beneficiile inteligenței artificiale numai dacă se va garanta fiabilitatea acesteia.

Nu în ultimul rând, dincolo de frontierele Europei, aceste orientări vizează, de asemenea, **stimularea reflecției și a discuțiilor** privind un cadru etic pentru inteligența artificială la **nivel mondial**.

ORIENTĂRI DE PUNERE ÎN APLICARE

Fiecare capitol din orientări oferă îndrumări pentru dezvoltarea unei inteligențe artificiale fiabile, care se adresează tuturor părților interesate relevante care dezvoltă, implementează ori utilizează IA, prezentate succint în continuare:

Capitolul I: Orientările esențiale pentru realizarea scopului etic:

- Asigurarea faptului că inteligența artificială este **centrată pe om**: Inteligența artificială ar trebui să fie dezvoltată, implementată și utilizată cu „**un scop etic"**, având la bază și reflectând drepturile fundamentale, valorile societale și principiile etice precum cel al *binefacerii* (să faci bine), al *nefacerii răului* (să nu faci rău), al *autonomiei oamenilor, al dreptății* și al *explicabilității.* Acest lucru este esențial pentru eforturile de dezvoltare a unei **inteligențe artificiale fiabile**.
- Întemeierea pe drepturile fundamentale, pe principiile și valorile etice, în scopul evaluării în perspectivă a efectelor pe care le-ar putea avea inteligența artificială asupra ființelor umane și a binelui comun. Acordarea unei **atenții deosebite** situațiilor care implică **grupuri mai vulnerabile**, cum ar fi copiii, persoanele cu handicap sau minoritățile, ori situațiilor în care există **asimetrii de putere sau de informație**, cum ar fi în relația dintre angajatori și angajați sau între întreprinderi și consumatori.
- Recunoașterea și conștientizarea faptului că, deși aduce beneficii substanțiale persoanelor fizice și societății, IA poate avea, de asemenea, un impact negativ. Menținerea vigilenței cu privire la domeniile de interes critic.

Capitolul II: Orientări esențiale pentru dezvoltarea unei inteligențe artificiale fiabile:

- Integrarea cerințelor referitoare la **inteligența artificială fiabilă din cea mai timpurie etapă de proiectare**: responsabilitatea, guvernanța datelor, proiectarea pentru toți, guvernanța autonomiei IA (supravegherea de către om), nediscriminarea, respectarea autonomiei oamenilor, respectarea vieții private, soliditatea, siguranța, transparența.
- Analizarea metodelor tehnice și fără caracter tehnic pentru a se asigura integrarea acestor cerințe în sistemul IA. În plus, trebuie să se țină seama de aceste cerințe în etapa de formare a echipei care va lucra la sistem, la proiectarea sistemului propriu-zis, a mediului de testare și a aplicațiilor potențiale ale sistemului.
- Să se furnizeze, în mod clar și proactiv, **informații părților interesate** (clienți, angajați etc.) cu privire la capacitățile și limitările sistemului IA, permițându-le să aibă așteptări realiste. Asigurarea **trasabilității** sistemului IA este esențială în acest sens.
- **Integrarea** IA fiabile în **cultura organizației** și furnizarea de informații părților interesate cu privire la modul în care componentele care țin de IA fiabilă sunt puse în aplicare în proiectarea și utilizarea sistemelor IA. De asemenea, IA fiabilă poate fi inclusă în standardele deontologice sau în codurile de conduită ale organizațiilor.
- Asigurarea participării și a **includerii părților interesate** în conceperea și dezvoltarea sistemului IA. În plus, asigurarea **diversității** atunci când se formează echipele care dezvoltă, implementează și testează produsul.
- **Asigurarea posibilității de auditare** a sistemelor IA, în special în contexte sau situații critice. În măsura în care este posibil, configurarea sistemului dumneavoastră în așa fel încât să permită identificarea legăturii dintre deciziile individuale și diferitele date de intrare, date, modele preformate etc. În plus, **definirea unor metode de explicare** a sistemului IA.
- Asigurarea unui proces specific pentru **guvernanța în materie de responsabilitate**.
- Asigurarea **formării și a educării**, precum și a faptului că managerii, dezvoltatorii, utilizatorii și angajatorii sunt informați în ceea ce privește inteligența artificială fiabilă și sunt formați în acest sens.
- Conștientizarea faptului că ar putea exista tensiuni fundamentale între diferitele obiective (transparența poate deschide calea utilizării abuzive; identificarea și corectarea judecăților părtinitoare ar putea să contrasteze cu măsurile de protecție a vieții private). Comunicarea și documentarea acestor compromisuri.
- Stimularea cercetării și a inovării pentru a se avansa în direcția îndeplinirii cerințelor privind dezvoltarea unei inteligențe artificiale fiabile.

Capitolul III: Orientări esențiale pentru evaluarea fiabilității IA:

- Utilizarea unei **liste de evaluare** pentru fiabilitatea IA atunci când se dezvoltă, se implementează ori se utilizează IA și adaptarea acesteia în funcție de utilizarea specifică dată sistemului.
- Luarea în considerare a faptului că o listă de evaluare nu va fi **niciodată exhaustivă** și că asigurarea unei IA fiabile nu înseamnă bifarea unor căsuțe, ci un proces continuu de identificare a cerințelor, de evaluare a soluțiilor și de asigurare a unor rezultate îmbunătățite pe parcursul întregului ciclu de viață al sistemului IA.

Aceste orientări fac parte dintr-o viziune care îmbrățișează o abordare a inteligenței artificiale centrată pe om, care va permite Europei să devină un lider mondial în domeniul inovării în ceea ce privește dezvoltarea unei inteligențe artificiale etice, sigure și bazate pe tehnologie de vârf. Prezentele orientări sunt menite să faciliteze și să creeze un cadru favorabil dezvoltării unei „**IA fiabile, create în Europa**”, ceea ce va spori bunăstarea cetățenilor europeni.