Europeiska kommissionens

EXPERTGRUPP PÅ HÖG NIVÅ FÖR AI-FRÅGOR

UTKAST TILL ETISKA RIKTLINJER FÖR TILLFÖRLITLIG AI

SAMMANFATTNING

Arbetsdokument för samråd med intressenter

Bryssel den 18 december 2018

UTKAST TILL ETISKA RIKTLINJER

FÖR TILLFÖRLITLIG AI

Kommissionens expertgrupp på hög nivå för AI-frågor
Utkast till etiska riktlinjer för tillförlitlig AI

EUROPEISKA KOMMISSIONEN
Generaldirektoratet för kommunikation

Kontakt: Nathalie Smuha – Samordnare AI HLEG
E-post: CNECT-HLG-AI@ec.europa.eu

Europeiska kommissionen
B-1049 Bryssel

Dokumentet offentliggjordes den 18 december 2018, på engelska.

Det här arbetsdokumentet har utarbetats av kommissionens expertgrupp på hög nivå för AI-frågor (AI HLEG). Det påverkar inte individuella ställningstaganden från gruppens medlemmar i specifika frågor och inte heller den slutliga versionen av dokumentet. Dokumentet kommer att vidareutvecklas och en slutlig version kommer att presenteras i mars 2019 efter samrådet med intressenter via Europeiska AI-alliansen.

Varken Europeiska kommissionen eller någon person som agerar för kommissionens räkning ansvarar för hur nedanstående information kan komma att användas. Expertgruppen på hög nivå för AI-frågor ansvarar ensam för innehållet i detta arbetsdokument. Personal vid kommissionens enheter har bidragit till utarbetandet av dessa riktlinjer, men de synpunkter som framförs återspeglar expertgruppens ståndpunkter och kan inte under några omständigheter anses vara ett uttryck för Europeiska kommissionens officiella ställningstagande. Detta är ett utkast till expertgruppens första leverabel. En slutlig version av riktlinjerna kommer att läggas fram för kommissionen i mars 2019. En slutlig version av arbetsgruppens nästa leverabel – en slutlig version av rekommendationer om AI-politik och investeringar – kommer att läggas fram i mitten av 2019.

Mer information om expertgruppen på hög nivå för AI-frågor finns här: (https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence).

SAMMANFATTNING

Det här arbetsdokumentet är ett utkast till de etiska riktlinjer för AI-frågor som utarbetas av Europeiska kommissionens expertgrupp på hög nivå för AI-frågor (AI HLEG). En slutlig version ska presenteras i mars 2019.

Artificiell intelligens (AI) är en av de mest omvälvande krafterna i vår tid, och den kommer att påverka hela samhället i grunden. Den innebär fantastiska möjligheter att öka välfärden och tillväxten, vilket är något Europa måste sträva efter. Under det senaste årtiondet har stora framsteg gjorts tack vare tillgången till enorma mängder digitala uppgifter, kraftfull dataarkitektur och framsteg inom AI-tekniker som maskininlärning. Betydande AI-understödd utveckling inom områden som förarlösa fordon, hälso- och sjukvård, hem- och servicerobotar, utbildning och cybersäkerhet förbättrar vår livskvalitet i vardagen. Dessutom är AI en viktig faktor för att lösa många av de utmaningar världen står inför, som global hälsa och välfärd, klimatförändringar, tillförlitliga rättssystem och demokratiska system, liksom andra av FN:s mål för hållbar utveckling.

Artificiell intelligens har kapacitet att generera enorma vinster för individer och samhälle, men innebär också vissa risker som måste hanteras på rätt sätt. På det hela taget är vinsterna med AI större än riskerna, och vi måste därför se till att följa den väg som **maximerar fördelarna med AI och samtidigt minimerar riskerna**. För att se till att vi håller oss på rätt spår behövs en **människocentrerad strategi för AI**, som tvingar oss att ständigt ha i åtanke att utveckling och användning av AI inte ska ses som ett självändamål utan som ett sätt att öka människors välbefinnande. **Tillförlitlig AI ska vara vår ledstjärna**, eftersom människor måste kunna lita på tekniken för att med tillförsikt kunna dra full nytta av AI.

Tillförlitlig AI har **två komponenter:** 1) Den ska respektera grundläggande rättigheter, tillämplig lagstiftning och grundläggande principer och värderingar, för att garantera att den har ett ”**etiskt syfte**”, och 2) den ska vara **tekniskt robust** och tillförlitlig, eftersom bristande teknisk kompetens även med goda avsikter kan orsaka oavsiktlig skada.

I riktlinjerna anges därför en **ram för tillförlitlig AI:**

- **Kapitel I** handlar om att **säkerställa AI:s etiska syfte** genom att fastställa de grundläggande rättigheter, principer och värderingar den ska respektera.
- Utifrån dessa principer beskrivs i **kapitel II riktlinjer för förverkligandet** av tillförlitlig AI, som har ett etiskt syfte och som är tekniskt robust. Detta görs genom att lista kraven på tillförlitlig AI och ge en överblick över de tekniska och icke-tekniska metoder som kan användas för att genomföra detta.
- I **kapitel III** görs sedan kraven **operativa** med hjälp av en konkret – men inte uttömmande – lista för bedömning av tillförlitlig AI. Listan ska sedan anpassas till specifika användningsområden.

Till skillnad från andra dokument om etisk AI syftar dessa riktlinjer alltså inte till att skapa ytterligare en förteckning över kärnvärderingar och principer för AI, utan snarare till att ge vägledning för att konkret implementera dem och göra dem operativa i AI-system. Vägledningen ges i tre nivåer av abstraktion, från den mest abstrakta i kapitel I (grundläggande rättigheter, principer, värderingar) till den mest konkreta i kapitel III (bedömningslista).

Riktlinjerna är avsedda för **alla relevanta berörda parter som utformar, utvecklar eller använder AI**, däribland företag, organisationer, forskare, offentliga tjänsteleverantörer, institutioner, enskilda och andra organ eller enheter. I den slutliga versionen av dessa riktlinjer kommer vi att presentera en mekanism genom vilken intressenterna frivilligt kan godkänna dem.

Det är viktigt att poängtera att riktlinjerna inte är avsedda att ersätta eller avråda från någon form av beslutsfattande eller reglering (vilket kommer att tas upp i expertgruppens nästa leverabel,

rekommendationer för politik och investering, i maj 2019). Riktlinjerna bör dessutom ses som ett levande dokument som regelbundet behöver uppdateras för att fortsätta vara relevant, efterhand som tekniken och vår kunskap om den utvecklas. Det här dokumentet bör därför ses som en utgångspunkt för diskussionerna om ”**tillförlitlig europeisk AI**”.

Europas enda möjlighet att sprida sin etiska AI-strategi är genom att vara konkurrenskraftigt på global nivå. Samtidigt är **en etisk AI-strategi nyckeln till en ansvarsfull konkurrens**, då den kommer att skapa förtroende hos användarna och underlätta en bredare användning av AI. Riktlinjerna är inte avsedda att hämma AI-innovation i Europa utan syftar i stället till att använda etiken som inspiration för att utveckla en unik sorts AI som kan skydda och vara till nytta för såväl individen som det allmänna. Det skulle göra det möjligt för Europa att ta en ledarposition när det gäller etisk och säker AI av spetskvalitet. För att människor i Europa fullt ut ska dra nytta av fördelarna med AI krävs att tillförlitligheten garanteras.

Slutligen syftar riktlinjerna till att även utanför Europa **främja reflektion och diskussion** kring en etisk ram för AI på **global nivå**.

SAMMANFATTNING AV VÄGLEDNINGEN

Varje kapitel i riktlinjerna ger vägledning om hur tillförlitlig AI ska kunna åstadkommas, och riktar sig till alla relevanta berörda parter som på något sätt deltar i utformning, utveckling och användning av AI. Nedan följer en sammanfattning.

Kapitel I: Viktiga råd för att garantera etiskt syfte:

- Se till att AI är **människocentrerad**: AI bör utformas, utvecklas och användas med ett ”**etiskt syfte**” som grundas på och återspeglar de grundläggande rättigheterna, samhällsvärderingarna och de etiska principerna *godhetsprincipen* (göra gott), *icke skada-principen* (inte skada), *autonomiprincipen* (människans självbestämmande), *rättviseprincipen* och *begriplighetsprincipen*. Dessa principer är avgörande för en **tillförlitlig AI**.
- Framtidsutvärdering av AI:s eventuella påverkan på människor och det allmännas bästa bör baseras på grundläggande rättigheter, etiska principer och värderingar. **Särskild uppmärksamhet** bör ägnas situationer som gäller mer **sårbara grupper**, som barn, personer med funktionsnedsättningar och minoriteter, eller situationer där det råder **obalans i fråga om makt eller information**, som förhållandet mellan arbetsgivare och arbetstagare eller mellan företag och konsumenter.
- Det är viktig att erkänna och vara medveten om att samtidigt som AI kan medföra betydande fördelar för individ och samhälle kan den också ha negativa effekter. Det krävs uppmärksamhet på viktiga problemområden.

Kapitel II: Viktiga råd för att skapa tillförlitlig AI:

- Införliva **kraven på tillförlitlig AI redan i de tidigaste faserna av utformningen**: Ansvarsskyldighet, datastyrning, design för alla, styrning av AI:s självständighet (mänsklig översyn), icke-diskriminering, respekt för människans autonomi, respekt för personlig integritet, robusthet, säkerhet samt öppenhet och insyn.
- Överväg både tekniska och icke-tekniska metoder för att garantera att dessa krav byggs in i AI-systemet. Kraven bör också hållas i minnet när man bygger upp de team som ska arbeta med systemet, själva systemet, testmiljön och systemets potentiella applikationer.
- Ge tydlig och proaktiv **information till berörda parter** (kunder, anställda osv.) om AI-systemets kapacitet och begränsningar, så att alla kan skapa sig rimliga förväntningar. I detta sammanhang är det av central vikt att **spårbarhet** byggs in i AI-systemet.

- Gör tillförlitlig AI till **en del av organisationens kultur** och informera alla berörda om hur begreppet tillförlitlig AI tillämpas vid utformning och användning av AI-systemen. Tillförlitlig AI kan också inkluderas i organisationens yrkesetiska regler och uppförandekoder.
- Se till att de **berörda parterna är delaktiga och inkluderade** vid utformningen och utvecklingen av AI-system. Det är också viktig att försäkra sig om **mångfald** när man bygger upp de team som ska utveckla, implementera och testa produkten.
- Sträva efter att **underlätta möjligheter till revision** av AI-system, särskilt i kritiska kontexter eller situationer. I möjligaste mån bör systemen utformas så att enskilda beslut kan spåras till olika indata, såsom data, färdigtränade modeller osv. Dessutom bör AI-systemets **förklaringsmetoder** definieras.
- Säkerställ en specifik process för **styrning av ansvarsskyldigheten**.
- Planera **utbildning** och se till att chefer, utvecklare, användare och arbetsgivare är medvetna om och utbildade i tillförlitlig AI.
- Det är viktigt att vara medveten om att det kan finnas grundläggande motsättningar mellan olika mål (öppenhet kan öppna dörrar för missbruk; identifiering och korrigering av bias (vinkling) kan strida mot integritetsskydd). Kommunicera och dokumentera kompromisser som görs i dessa sammanhang.
- Främja forskning och innovation som kan bidra till att uppnå kraven för tillförlitlig AI.

Kapitel III: Viktiga råd för att bedöma tillförlitlig AI:

- Upprätta en **lista för bedömning av** tillförlitlig AI under utformning, utveckling och användning av AI och anpassa den till det specifika område där systemet används.
- Kom ihåg att en sådan bedömningslista **aldrig kan bli uttömmande** och att tillförlitlig AI aldrig får handla om att kryssa i checkrutor, utan är en kontinuerlig process för att identifiera krav, utvärdera lösningar och försäkra sig om förbättrade resultat under hela AI-systemets livscykel.

Denna vägledning är en del av en vision om en människocentrerad strategi för artificiell intelligens som kommer att göra det möjligt för Europa att bli en världsledande innovatör inom säker och etisk AI med spetskvalitet. Målet är att underlätta och möjliggöra en **”tillförlitlig europeisk AI”** som kommer att öka välbefinnandet för människor i Europa.