Infrastrukturdepartementet

Yttrande angående remiss av Europeiska kommissionens förslag till förordning om harmoniserade regler för artificiell intelligens, I2021/01304

Chalmers har tagit del av remissen av Europeiska kommissionens förslag till förordning om harmoniserade regler för artificiell intelligens. Nedan återfinns vårt svar och våra kommentarer på förslaget.

Överlag ställer vi oss positiva till förslaget om harmoniserade regler för artificiell intelligens (AI).

AI är en global fråga: effekterna av oansvarig teknikutveckling och/eller användning på en plats begränsas inte nödvändigtvis av nationsgränser, utan spiller lätt över till andra länder. Internationellt samarbete är därför extra viktigt för att bemästra en kapplöpningsdynamik som annars kan bli farlig, då den riskerar leda till att länder, företag och andra aktörer rusar framåt med utveckling och användning av AI utan att tycka sig ha tid till avgörande frågor om etik, sociala konsekvenser och riskhantering. Och från exempelvis skattepolitik känner vi ju till fenomenet ”race to the bottom”, där avsaknad av gemensam reglering lett till att länder tävlar om att locka till sig företag med olika skatte- och regleringslättnader som gynnar dessa företag men i övrigt kan ha negativa sociala konsekvenser. **Vi ser därför positivt på att EU kopplar ett grepp om regleringen av AI i medlemsländerna**, vilket inte bara innebär en harmonisering som gör det lättare för företag att verka på den europeiska marknaden, utan också stävjar destruktiv kapplöpnings- och race to the bottom-dynamik. Allra helst skulle vi såklart vilja se en samordning ännu högre upp än på EU-nivå, men det som nu föreslås är i högsta grad ett steg i rätt riktning, och erfarenheter från GDPR visar att EU-lagstiftning påverkar agerandet hos globala teknikföretag på ett sätt som ger ringar på vattnet även utanför EU. Tack vare fungerande samhällsinstitutioner har EU nu möjlighet att visa positiva exempel på möjligheten att reglera AI, och en klok och välavvägd EU-lagstiftning kan bli vägledande för resten av världen.

Förslagets riskbaserade ramverk – med dess indelning av AI-tillämpningar i kategorierna oacceptabel risk, högrisk och lågrisk, jämte den separata kategori där särskilda transparenskrav skall gälla – är ett utmärkt sätt att navigera den svåra farleden mellan å ena sidan det ansvarslösa laissez-faire-tänkandets Skylla och å andra sidan det naiva all-risk-måste-elimineras-tänkandets Karybdis. **Att delar av den definition av högrisk-tillämpningar som ligger till grund för indelningen hänskjuts till en uppräkning tillämpningssituationer i ett Annex som är tänkt att relativt snabbt kunna uppdateras är också bra**, med tanke på att AI-utvecklingen går så snabbt att lagstiftning på området utan den sortens beredskap för uppdatering löper stor risk att snabbt bli föråldrad.

Den föreslagna regleringen kommer självklart att innebära visst administrativt och annat merarbete för den som vill rulla ut ny AI-teknik för sina kunder, i fall där tekniken ifråga hamnar i högrisk- och/eller transparenskravskategorin. Vi gissar att en del av svaren från andra remissinstanser kommer att peka på denna börda och föreslå lättnader, men vi menar att **med tanke på de enorma risker av olika slag för människor och samhället som oreglerad AI skulle föra med sig så är den administrativa bördan ett ytterst rimligt pris att betala för att bidra till att hantera dessa risker**.

Dokumentet innehåller ett avsnitt om regelverk för regulatoriska sandlådor, men utöver detta rör inte den föreslagna regleringen utvecklingen av ny AI-teknik, utan bara dess användning. En ryggmärgsreaktion från Chalmers skulle därmed kunna vara lättnad över att vår forsknings- och utvecklingsverksamhet inom AI-området på så vis undkommer regulativa pålagor, men för oss innebär detta att ett desto större moraliskt ansvar faller på oss och våra forskare att noggrant överväga risker med den AI-teknik vi bidrar till att utveckla, att vidta mått och steg för att minimera dessa risker, och att avstå från att gå vidare med projekt för vilka sådana överväganden landar i att riskerna är oacceptabelt stora i förhållande till teknikens potential att göra gott. Vikten av sådant ansvarstagande hos oss och andra FoU-aktörer accentueras ytterligare av att den föreslagna regleringen inte kommer åt riskabel AI-teknik utvecklad inom EU för användare utanför EU.

Så långt våra allmänna reflektioner om EU-dokumentet. Avslutningsvis några mer specifika kommentarer:

- Ordet ”intended” (avsedd) förekommer 82 gånger i lagförslagets huvuddokument. Även om det noggrant understryks att ett AI-system kan ha andra effekter än de avsedda (vilket såklart är ett av skälen till att reglering behövs) **finns en risk att en överbetoning av ett AI-systems avsedda effekter kan uppmuntra den farliga men vanligt förekommande attityd där en aktör intalar sig att eftersom dennes avsikter med den planerade AI-tillämpningen är goda är allt i sin ordning**, så att ingen ytterligare analys behövs av etiska och sociala aspekter eller av risk.

- På sidan 22 talas om förbud mot ”the […] use of certain AI-systems intended to distort human behaviour, whereby physical or psychological harm are likely to occur”. Vi noterar att detta låter tillämpligt på en del AI-system som redan är i utbrett bruk, som AI-algoritmer för exempelvis utplacering i sociala medier-flöden av reklam avsedd att få oss att dricka mer läsk, med alla de negativa hälsoeffekter sådan konsumtion för med sig. Vi har **inget självklart svar på den intrikata frågan om hur långt det förbud som här omtalas bör sträcka sig, men vi finner det angeläget att bättre klargöra** huruvida denna typ av tillämpning täcks in. På sidan 43, Article 5.1(a), nämns villkoret ”subliminal”, vilket dock är ett alltför töjbart begrepp för att vara helt klargörande.

- Stor tonvikt läggs på vikten av att undvika diskriminering på basis av exempelvis kön eller etnicitet, men **ingenstans i förslaget definieras den önskade icke-diskrimineringen**. Det som gör denna brist potentiellt allvarlig är att flera olika icke-ekvivalenta men till synes helt rimliga definitioner av icke-diskriminering mellan grupper finns, och under andra än de mest tursamma omständigheter visar det sig **omöjligt att uppfylla alla dessa samtidigt**; detta är

Kleinbergs så kallade omöjlighetsteorem för algoritmisk rättvisa.[1] Detta gör att **man behöver välja: vill vi att AI-algoritmerna skall ge de olika grupperna samma så kallade false positive och false negative rates, eller är det viktigare att de uppfyller den egenskap som benämns kalibrering?** Det kan hända att det valet är något som bäst hänskjuts till den enskilda tillämpningen och som därför tillåts variera från fall till fall, men det är i så fall något som vore värt att nämna i dokumentet. (Att lämna till de olika medlemsländerna att avgöra hur icke-diskriminering skall definieras skulle däremot underminera idén om att medelst harmonisering underlätta utrullningen av AI-teknik inom EU. Olämpligt vore också att med hänvisning till omöjlighetsteoremet förbjuda AI i det slags situationer där teoremet är tillämpligt; detta vore att skjuta bredvid målet, eftersom det är lika tillämpligt oavsett om bedömningarna eller besluten görs av AI eller av människor.)

- På sidan 48, punkt 3, stipuleras bland annat kravet att data skall vara **”free of errors and complete”. Denna formulering behöver nog mjukas upp**, för om den tas bokstavligt torde kravet uppfyllas av ytterst få eller inga alls av dagens maskininlärningssystem, och i praktiken dessutom vara omöjligt att verifiera.

Göteborg, 21 juni 2021

I tjänsten,

Olle Häggström, professor

Anders Palmqvist, vicerektor för forskning och styrkeområden

[1] Se exempelvis Kleinberg, J., Mullainathan, S. och Raghavan, M. (2016) Inherent trade-offs in the fair determination of risk scores, https://arxiv.org/abs/1609.05807; Sumpter, D. (2019) *Uträknad: sanningen om algoritmerna som styr världen*, Volante, Stockholm; och Häggström, O. (2021) *Tänkande maskiner: Den artificiella intelligensens genombrott*, Fri Tanke, Stockholm.