

Uundgåelige kunstig intelligens-hallucinationer kan sende borgere på vildspor

INDLÆG

FEBRUAR 2026

Af Daniel Juhász Vigild, ph.d.-studerende,
ROCKWOOL Fonden

Lige inden jul præsenterede regeringen en aftale om tre store projekter med kunstig intelligens. Et af projekterne går ud på at erstatte de nuværende digitale selvbetjeningsløsninger med chatbotbaserede digitale assistenter, der kan hjælpe borgere og virksomheder på rette vej, når de skal navigere rundt i den digitale offentlige sektor, eksempelvis på borger.dk eller sundhed.dk. Det være en sygdomsramt borger, der søger viden om, hvad han eller hun har ret til af støtte, eller hvad jobcenteret kræver for, at man opretholder kontanthjælp.

Problemet er dog, at en digital assistent risikerer at sende borgere og virksomheder i en forkert retning. Dette skyldes den underlæggende teknologi: De store sprogmodeller, som eksempelvis OpenAIs ChatGPT, erotorisk kendt for at hallucinere. En "hallucination" er udtryk for en falsk information, modellen påstår - ofte på temmelig overbevisende vis - er sand.

Som vi har set helt konkret på det seneste, kan det være en bog, som en forfatter ikke har skrevet, eller et citat en politiker ikke har sagt. Som eksemplerne viser, kan chatbotten overbevise selv nogle af de mest kompetente og veluddannede mennesker i Danmark.

Det er denne teknologi, regeringen ønsker skal vejlede borgere og virksomheder, når de skal forstå kompleks lovgivning og navigere i snørklede offentlige systemer.

Selvom man godt kan forestille sig, at teknologien kan give hurtigere sagsbehandling og en bedre oplevelse for mange borgere, følges det af en betydelig risiko for, at fejlbehæftede vejledninger kan have negative konsekvenser for især ressourcetsvage borgere i udsatte positioner. Det kunne være tidsfrister, der ikke blev overholdt, misforståede rettigheder eller tabte muligheder, eksempelvis at man ikke får den ydelse, man ellers er berettiget til.

Problemet med hallucinationer er ikke begrænset til enkeltstående tilfælde med stor mediefokus, som eksemplet med Dansk Folkepartis fejlagtige kronik: Det er systematisk og uundgåeligt på grund af den måde, matematikken i maven på chatbotten virker.

Her bliver gætteeri nemlig belønnet, mens usikkerhed bliver straffet, og den udfordring forsvinder ikke, så længe chatbots er skruet sammen på den måde, de er i dag.

Det er svært at spå om, hvor tit disse hallucinationer vil opstå, når først de digitale assistenter bliver taget i brug i den offentlige sektor. To offentligt tilgængelige oversigter forsøger dog at måle, hvor ofte de mest populære chatbots hallucinerer, og de fleste modeller giver falske svar omkring 1-2 ud af 10 gange.

Derfor er det centralt, at man kan opdage, når chatbotten indimellem hallucinerer. For borgere med komplicerede sager og spørgsmål bliver højst sandsynligt den gruppe, der får sværest ved at bekræfte, at den vejledning de modtager, er faktisk korrekt.

Man kan derfor frygte, at hallucinationerne vil ramme socialt skævt, og lede netop de borgere, der har allermest brug for en hjælpende menneskelig hånd på digitalt vildspor.

Dermed ikke sagt at teknologien er værdiløs. Som regeringens aftale også lægger op til, kan chatbots være gode assistenter til at vejlede den erfarne sagsbehandler og derved give et nyttigt produktivitetsløft.

Måske det også på længere sigt kan være ansvarligt at få den til at hjælpe borgere, hvis man kan give chatbotten lidt mindre "selvtillid", så den bliver bedre til at indrømme: "det her ved jeg ikke nok om, måske burde du tale med et menneske?".

