Reactie op Tegenlicht "Digimens"

geschreven op 2021-12-04 door Michiel van der Meer en Sabina van Rooij. Deze tekst valt ook te vinden op Michiel's website.

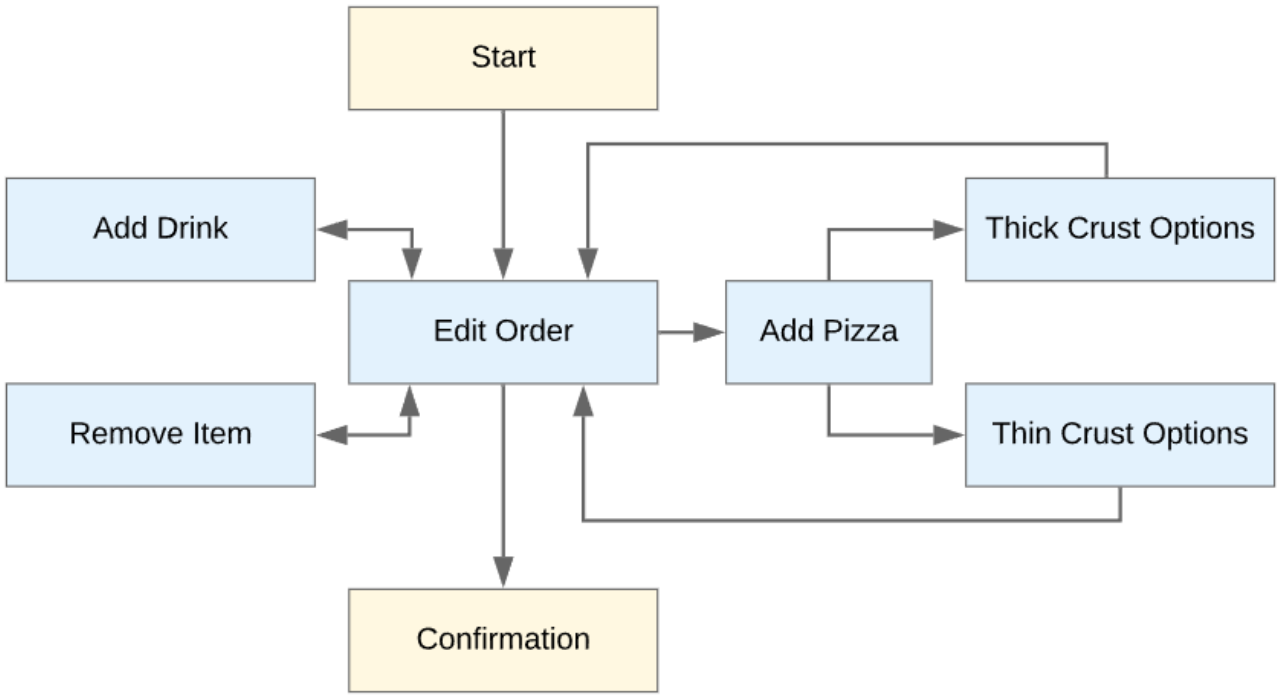
In de Tegenlicht documentaire komen een aantal AI systemen en robots aan bod. Voor deze systemen heb ik wat dieper gegraven naar 1) wat voor beweringen naar voren komen over de vaardigheden van de systemen, 2) in hoeverre deze beweringen waar zijn en 3) of er publiekelijke informatie over te vinden was waarop ik deze inschatting kon baseren.

De genoemde AI systemen

Digimens (Digital Human)

In dit geval is het moeilijk te achterhalen hoe de techniek in elkaar steekt, met name omdat de bedrijven niet hun product zomaar weggeven. Wat hieronder volgt is een korte introductie over hoe gesprekken worden ingeprogrammeerd, en hoe dergelijke system worden gebruikt.

Zowel Sophie van Philadephia als Betsy in Zweden zijn ontworpen door Deloitte, en de achterliggende ontwikkelaar Uneeq. Deze systemen zijn ontwikkeld om een gesprek aan te gaan, waarbij gebruik wordt gemaakt van de lessen uit Conversation Design. Hierin wordt er geprobeerd om computers de menselijke vorm van dialogen te laten volgen.



Een voorbeeld van een dialoogstructuur om een pizza te bestellen.

Hierbij wordt de structuur van een gesprek expliciet genoteerd. Van tevoren staat de vorm, en ook het doel van een gesprek dus vast. Hoewel de mogelijkheid om uit te wijden tot op zekere hoogte kan worden ingeprogrammeerd, zal een dergelijk systeem altijd een vaste volgorde van vragen aflopen, net als een theaterscript, of zelfs als een soort algoritme. Wat een computer begrijpt, is in welk stuk van het gesprek we momenteel zijn, en wat voor soort antwoorden een mens zou kunnen geven. Op het moment dat een gebruiker dan een antwoord geeft, zal (een deel van) het antwoord gebruikt worden om naar een nieuwe sectie van het gesprek te springen.

Digimens: Sophie

Je kan zelf in gesprek gaan met Sophie (in het engels), via https://sophie.digitalhumans.com/, mits je computer een microfoon heeft. Wat ik opmerk is dat Sophie het initiatief neemt in het gesprek: zij stelt de vragen. Probeer het eens om te draaien, en stel haar in de plaats wat vragen. In zulke gevallen kom je al snel in situaties terecht dat ze niet inhoudelijk reageert op wat je zegt.

In Tegenlicht wordt als voorbeeld gegeven dat Sophie ingezet kan worden als assistent bij DigiContact, bijvoorbeeld als voorselectie voor wie er een menselijke gesprekspartner nodig heeft. Zie hier het commentaar van Alix Rübsaam bij, verder hieronder. Interessant is wel dat ze erop hameren dat het dus niet vanuit een vervangperspectief wordt ingezet, maar als extra gereedschap voor mensen.

Digimens: Betsy

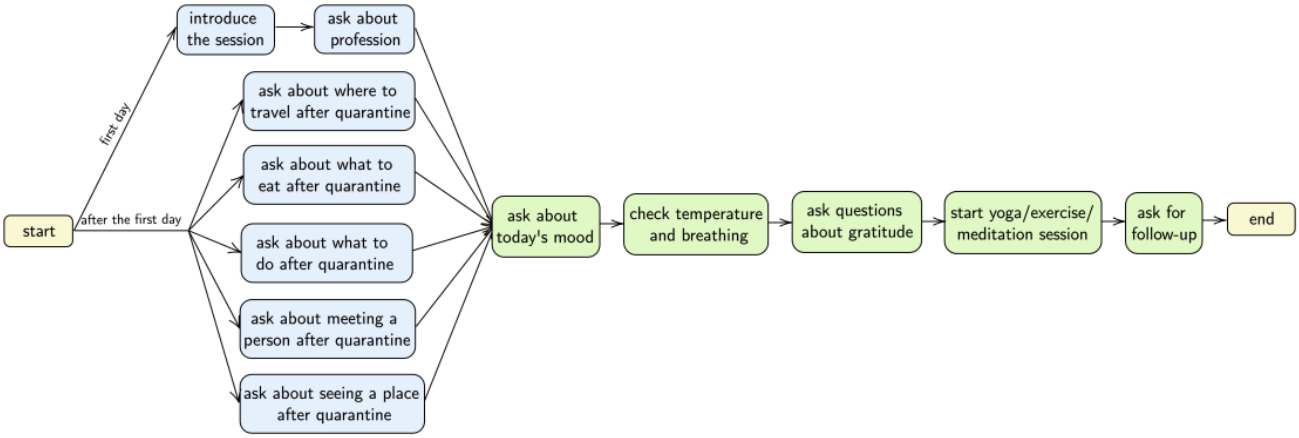
Betsy wordt momenteel onderzocht door Almira Osmanovic om in te zetten als therapeut. De technische opzet is waarschijnlijk vergelijkbaar met die van Sophie, maar dus met een andere invulling van het dialoogstructuur. De opzet is ditmaal een therapeutisch gesprek, waar de AI de standaardvragen stelt om de gemoedstoestand van een persoon te achterhalen. Hierbij wordt de digitale avatar gebruikt zodat mensen makkelijker praten. De onderliggende technologie die zorgt voor inhoudelijk begrip is echter nog steeds gelijk aan die van chatbots. Ook lijkt de avatar voornamelijk te reageren op de inhoud van de tekst, waarschijnlijk begrijpt ze sarcasme bijvoorbeeld niet (ik heb dit zelf getest op de bovenstaande link van Sophie, dat doet het daar inderdaad niet). Het is wel mogelijk om de eigenschappen van stemgeluid te gebruiken om dit te herkennen met behulp van AI, dit is onder andere gebruikt in een weerbaarheidstraining tegen babbeltrucs.

Hans van den Brink (ICT'er) en Filip Bergquist (neuroloog) praten vervolgens over een ander scenario: oudere patienten met eventueel een aangetast denkvermogen denken dat ze met een echt persoon te maken hebben. Bijvoorbeeld bij een digitale kopie van Filip (waarvan nog maar te bezien is of dat mogelijk is op de korte termijn) wordt ook de Turing test aangehaald. Hier wordt dus weer gesproken over een ouderwetse definitie van intelligentie en vanuit het vervangperspectief: als een digitale kopie niet te onderscheiden is van een mens, dan is het gelijk aan een mens. Hierbij verwijs ik graag door naar het commentaar van Vincent Müller.

Aan het eind van de documentaire noemt Almira nog een belangrijke mogelijke toepassing van AI in therapeutische context: het bieden van mentale steun aan vluchtelingen in opvangcentra, waarbij een AI het overneemt van mensen om verergering (in het geval dat ze compleet onbehandeld blijven) te voorkomen. Ik zou hier echter tegenover willen stellen dat het toepassen van AI op deze schaal wel degelijk voor verergering kan zorgen, bijvoorbeeld doordat het toegepast wordt op een nieuwe doelgroep waar het niet op is getest. Meer onderzoek is op dit gebied nodig voordat dergelijke systemen ingezet kunnen worden. Gelukkig is Almira hiermee bezig!

Digimens: Nora

Nora is ontwikkeld door (studenten van) Pascale Fung. Omdat het hier, in tegensteling tot de vorige voorbeelden, om wetenschappelijk gaat is er meer informatie te vinden over de achterliggende techniek van Nora. Zo kunnen we ook de dialoogstructuur zien waar Nora gebruik van maakt.



De dialoogstructuur voor Nora.

Hier zien we dat er een standaard volgorde in het gesprek zit, en dieper begrip eigenlijk mist. Bij het interpreteren van de antwoorden van een mens, "voelt" Nora dan ook geen emoties, maar classificeert aan hand van geluids- en textpatronen de onderliggende emoties. Deze classificatie beschikt over 6 categorieën voor geluid: cynisch, angstig, boos, eenzaam, blij en verdrietig. Bij tekst gaat het om drie categorieën: angstig, blij en verdrietig. Dit is natuurlijk erg gesimplificeerd in vergelijking met complexe emoties waarover mensen beschikken. De patroonherkenning is in beide gevallen 60-65% accuraat, en bij Nora wordt geen beeldherkenning of gezichtsherkenning gebruikt. Omdat het werk dusdanig recent is (juni 2021), is het moeilijk te zeggen wat voor effect de gesprekken hebben op mensen. Waarschijnlijk zijn de experimenten nog bezig.

Pascale zegt aan het eind ook nog dat machines getraind kunnen worden om "unbiased" te zijn. Echter, zoals ik al verteld heb en ook blijkt uit talloze onderzoeken, is het compleet weghalen van bias vrijwel onmogelijk. In plaats daarvan is het nodig om de systemen in context te plaatsen, te onderzoeken hoe ze gebruikt worden, wat de gevolgen zijn en er voornamelijk geen valse beloftes over maken. Fung benoemt wel het bestuderen van de mogelijke ethische en maatschappelijke impact. Dergelijke standaarden zijn voor AI onderzoekers steeds relevanter geworden, en zijn nu

bijvoorbeeld een belangrijk onderdeel van EMNLP, een van de beste conferenties voor AI-onderzoekers op het gebied van taalbegrip.

Fysieke robot: Phi

Phi is de Pepper robot van Philadephia, en tevens de enige fysieke robot in deze Tegenlicht documentaire. In het eerste stuk van de documentaire wordt Phi gebruikt om een interactie tot stand te brengen met de client. Voor deze interactie, in de vorm van een gesprek, wordt dan ook dialoogstructuur toegepast. Phi stelt vragen, waarop de client antwoord kan geven. Afhankelijk van het antwoord zegt Phi dan weer wat. Wat mij hieraan opviel is dat clienten hun antwoord geven via de tablet op de buik van Phi: waarschijnlijk is het verstaan van de client (technische term: speech to text, STT) te lastig in de omgeving waar Phi zich bevindt. Ook weten we niet of Phi in staat is om rond te rijden in huis, ze blijft immers op één plek staan. Desondanks lijken de clienten baat te hebben bij de interactie. Hierbij wil ik vooruitgrijpen naar iets wat Alix Rübsaam later in de documentaire noemt, namelijk dat het de vraag is of wij als mensen daadwerkelijk begrepen willen worden, of dat we alleen het gevoel willen hebben dat we gehoord worden. In het laatste geval kan Phi een prima apparaat zijn.

Mensen en AI-systemen

De twee filosofen die tegen het einde van de documentaire aan het woord komen hebben een misschien wat pessimistischer, maar ook wel realtischer beeld over hoe deze systemen moeten worden ingezet.

Alix Rübsaam slaat namelijk wat mij betreft de spijker op z'n kop. Haar eerdergenoemde opmerking over het projecteren van empathie is ook hoe ik er tegenaan kijk. Het "meevoelen" gebeurt niet in een computer, maar mensen voelen zich wel gehoord. In zulke gevallen kan een systeem wel degelijk toegevoegde waarde hebben. Het is inder daad wel cruciaal om hier transparant over te zijn, als onderzoekers.

In het geval dat systemen klakkeloos worden ingezet, schetst Alix het volgende:

> *Een chatbot doet een voorselectie om mensen wel of niet door te laten om een dokter te spreken. Mensen die dus niet overweg kunnen met zo'n chatbot, vallen buiten de boot, en worden niet door een dokter gehoord. Sterker nog, deze mensen zullen dus in het vervolg ook de dokter niet meer bellen. In het geval dat er ook een leersysteem is, waarbij de chatbot leert van feedback van gebruikers, wordt deze feedback gegeven door mensen die wel met een chatbot overweg kunnen (de anderen zouden namelijk simpelweg ophangen). De chatbot leert dus alleen van zij die wel door het proces heenkomen.*

Hier komt een gevaarlijke selectie bias om de hoek kijken, die ervoor kan zorgen dat mensen die nu al moeite hebben met het benaderen van een dokter nog minder aan bod zullen kunnen komen. Ook binnen de AI is dit een bekend probleem. Steeds vaker wordt er tegenwoordig aandacht

gegeven aan groepen mensen die buiten de boot zouden vallen, en het oplossen van deze biases is een steeds belangrijkere onderneming.

Vincent Müller heeft ook goeie punten, specifiek over empathie. De techneuten die eerder aan het woord komen (Pascale, Hans) stellen dat het herkennen van emotie en het naboosten van gedrag gelijkt aan het hebben van empathie. Echter, digitale systemen kunnen mogelijk goed nabootsen alsof ze empathisch zijn, maar hiermee houdt het ook op. Een computer heeft geen gevoel, dus kan zich ook niet voorstellen hoe jij je voelt. Hierbij is het dus gevaalijk als je een mens op een verkeerde manier inlicht over de werking van je systeem. Houd er dan ook rekening mee dat de manier waarop omgegaan wordt met herkende emoties voorgeprogrameeerd is door een mens: het systeem loopt intern een vast protocol af. Kort noemt Vincent ook nog hybride intelligentie: hoe kunnen we computers gebruiken om mensen beter hun werk te laten doen, niet om werk te vervangen. Vincent is dan ook betrokken bij mijn zuster-consortium, ESDiT.

Als laatste: de onderzoekers praten vaak gepersonificeerd over hun systemen: als er gezegd worden dat een systeem "nog moet leren" of "iets nog niet begrijpt," betekent het niet dat er een actief leermechanisme in zit. In het geval van dialogen, zou het kunnen betekenen dat de onderzoeker de dialoogstructuur moet aanpassen om een mogelijke vraag te kunnen beantwoorden. Hoewel ik moet erkennen dat onderzoekers hier voorzichtiger mee om zouden kunnen gaan, denk ik dat in veel gevallen termen als "leren" en "begrijpen" in de context van AI-systemen met een korreltje zout genomen moet worden.