

En analyse av NAV sin chatbot 'Frida'

Gruppe 1

Anna, Hedda, Nadia, Sara



Semesterrapport

Spesialisering innen forskning i design av IT

10 studiepoeng

Institutt for informatikk

Det matematisk naturvitenskapelige fakultet

UNIVERSITETET I OSLO

Oktober/2021

1. Introduksjon	3
1.1 Innledning	3
1.2 Definisjon	4
2. Bakgrunn	5
2.1 Hvorfor bruker mennesker chatboter?	5
2.2 NAV	5
3. Metode	6
3.1 Eksisterende rammeverk og guidelines for design av netttinnhold	7
3.1.1 WCAG 2.1	7
3.1.2 Amershi et al. (2019)'s 18 guidelines	7
4. Foreløpige funn	9
4.1 Forventninger fra datainnsamlingen	9
4.2 Presentasjon av funn	9
5. Tanker til fremtidig diskusjonsdel	10
6. Konkluderende refleksjon	10
Vedlegg 1 - Design av tidlig Chatbot-prototype	11
Introduksjon	11
Prosess	11
Refleksjon og Resultat	11
Vedlegg 2 - Refleksjoner til maskinlæringsoppgaven	14
Forberedelser	14
Kjøring av kode og endringer av verdier	14
Refleksjon	16
Vedlegg 3 - UU-analyse av NAV sin chatbot i Excel	18
Vedlegg 4 - Skjermbilder av initiell interaksjon med chatboten	20
Vedlegg 5 - Manuskript	22
Vedlegg 6 - Feedback-håndtering	24
Referanser	25

1. Introduksjon

Vår gruppe består av:

Anna Wasaznik Johannessen (annawas@uio.no)

Sara Hashemi Karoliussen (sarakaro@uio.no)

Nadia Hammar Skovgaard Møller (nhmoller@uio.no)

Hedda Kjetun Dyngeland (heddad@uio.no)

Vi tok alle vår bachelorgrad i Informatikk: design, bruk og interaksjon ved Universitetet i Oslo, og er nå førsteårsstudenter på masterprogrammet i samme studielinje.

1.1 Innledning

Denne oppgaven har vært gjennom mange stadier. Fra start har temaet vært å jobbe med universell utforming - især i forbindelse med interaksjon med det offentlige Norge. For å lande hvor oppgaven er nå, falt fokuset på NAV sin chatbot 'Frida'. Dette skjedde etter interaksjon med ulike offentlige chatboter som 'Kommune-Kari' og Frida. På disse utforsket vi ulike responser, slik som hvordan Frida responderer på vanlige skrivefeil som kan komme av dysleksi og hvordan denne chatboten tilpasser seg brukere som trenger alternative forklaringer på formelle ord (f.eks de med norsk som fremmedspråk).

Basert på dette ønsket vi å gjøre videre undersøkelser, og satte derfor opp en analyse i Excel med punktene fra prinsipp 3 'forståelse' i WCAG 2.1 standarden, samt Amershi et al. (2019) sine 18 retningslinjer for design av menneske-AI interaksjon (UUTilsynet 2021; Amershi et al. 2019, s. 3). Resultatet av disse initielle undersøkelsene var at chatboten hadde forbedringspotensiale både i forhold til prinsippene om forståelse, men også flere av retningslinjene til Amershi et al. (2019). I vår første iterasjon med bruk av chatboten 'Frida' hadde vi særlig fokus på hvorvidt brukerne forstod Frida (eks. gjennom forståelse av begreper).

Fra disse forundersøkelser i starten av denne iterasjonen til hvor vi er nå, har vi gjort oss flere erfaringer rundt hvordan ulike rammeverk kan brukes til ulike analyser.

I lys av vår initielle interaksjon med chatboten Frida og interesser for universell utforming, gikk vi først fra en generell tilgang til litteratur på området, til å lande på følgende

forskningsspørsmål som tar utgangspunkt i et konkret rammeverk for analyse av menneske-AI-interaksjon:

- Hvordan kan retningslinjene G4, G10, G12 fra Amershi et al. (2019) brukes til å forbedre forståelsen mellom chatboten Frida og brukerne?

Funnene fra vår første forundersøkelse førte til en innsnevring mot tre retningslinjer fra Amershi et al. (2019) som vi har valgt å føre videre i vår undersøkelser (G4 ‘show contextually relevant information’, G10 ‘scope services when in doubt’, G12 ‘remember recent interactions’). Disse retningslinjer planlegger vi å ta utgangspunkt i, slik at vi kan diskutere hvorvidt og hvordan Frida eventuelt treffer disse samt komme med potensielle forbedringsforslag. Endringen og innsnevringen av selve forskningsspørsmålet førte til at vi videre ønsket å foreta en systematisk hermeneutisk evaluering av NAV sin chatbot, og skiftet fokus mot Fridas forståelse av henvendelser fra brukeren og om chatboten også klarer å forstå mer sammensatte utfordringer brukeren kan ha.. Vi utformet derfor et felles manus (se vedlegg 5) med temaet arbeidsledighet som vi vil teste på chatboten i neste iterasjon og føre en mer systematisk samtale med Frida.

I nåværende iterasjon har vi beveget oss fra å fokusere på brukerens forståelse, til å undersøke chatboten sin forståelse av brukeren og dennes ofte komplekse situasjoner.

1.2 Definisjon

Det finnes mange definisjoner på hva en chatbot er, og gjør. Ifølge Luger og Sellen (2016), er chatboter et program som forsøker å etterligne menneskelige samtaler gjennom grafiske grensesnitt og/eller stemmeinteraksjon (2016, s. 2). Vi som gruppe bet oss merke i ordet *samtaler*, som er et komplekst og sammensatt fenomen. Foruten ord som settes sammen til setninger (sekvenser), spiller tidligere erfaring med temaer og situasjoner inn, og selve konteksten og bakgrunnen til personene som interagerer, påvirker samtalen i form av timing, kroppsspråk og valg av ord. Disse sosiale normene er svært viktige i det som utgjør kommunikasjon. I tillegg er selve formålet med å samtale og danne en *felles forståelse* (Moore, 2018, s. 186). Å danne en felles forståelse er allerede en utfordring mellom mennesker, og vi som gruppe anså det derfor som veldig spennende å forsøke oss på å etterstrebe en potensielt bedre *forståelse* mellom chatboten Frida og brukeren. Her forstår vi

forståelse i den forstand at chatboten kan prosessere input fra brukeren, og være i stand til å komme med passende tilbakemeldinger.

2. Bakgrunn

2.1 Hvorfor bruker mennesker chatboter?

Dette spørsmålet utforsker Petter Brandtzaeg og Asbjørn Følstad i sin artikkel om hva som motiverer mennesker til å bruke chatboter. De presenterer chatboter som erstatning for “special-purpose” applikasjoner og nettsider, og at dette kommer av at “*Chatbots are seen as a means for direct user or customer engagement through text messaging for customer service or marketing purposes*” (Brandtzaeg & Følstad, 2017, s. 377). Derimot har også mottakelsen av chatbotene vært kjøligere i noen settinger hvor innføringen av chatbot ikke levde opp til forventningene. Her tenker forfatterne at grunnen til dette er at chatboter ikke klarer å tilfredsstille brukernes behov. Chatboter kan være fylt med uklare mål, ulogiske svar og utilstrekkelig brukbarhet som gjør at opplevelsen etter å ha brukt en chatbot vil være underveldende eller skuffende (Brandtzaeg & Følstad, 2017; Kocielnik et al., 2019)

Et annet poeng gjort av forfatterne er at siden vi bruker mange ulike samtaleapplikasjoner som Messenger, Slack, Teams etc. er vi blitt vant til naturlige samtaler med andre mennesker via en applikasjon/nettsted (Brandtzaeg & Følstad, 2017). Dette tenker vi at kan skape høyere forventninger til chatboter, siden de følger et slikt format.. Videre tenker vi at dette innebærer at brukeren forventer forståelse fra chatboter på grunn av denne likheten til vanlige kommunikasjonskanaler. De ser da bort ifra at chatboter er program med AI-egenskaper og ikke mennesker.

2.2 NAV

NAV kan ses på som grunnsteinen i offentlig sektor og er en landsdekkende, statlig virksomhet som er underlagt Arbeids- og sosialdepartementet. Etaten har ansvaret for organisering og finansiering av arbeidsmarkedstiltak, trygdeytelser og sosialhjelp, og er derfor en viktig del av sikkerhetsnett i velferdsstaten. Organisatorisk er NAV en del av Arbeids- og velferdsdirektoratet og består av en ytre etat, regionskontorer og lokalkontorer som håndterer rundt 1/3 del av Norges statsbudsjett (Store Norske Leksikon). NAV sine tjenester treffer derfor en bred gruppe med brukere og har en særlig viktig rolle innen inkludering. Per januar 2021 var det 193 800 helt eller delvis arbeidsledige personer som

registrerte seg hos NAV (nav.no, a). Spesielt under korona-pandemien har mye av den første kontakten mellom brukere og NAV måttet gjøres på alternative måter enn fysisk oppmøte. Det er og lagt opp til at nav.no (inkludert veiledning via chatbot) skal være førstevalget til alle brukere som har tilstrekkelig digital kompetanse. Det er derfor svært viktig at tjeneste på nett oppleves av brukerne som enkle og intuitive i bruk (NAVs Personbrukerundersøkelse og Arbeidsgiverundersøkelse 2020).

Ifølge NAVs rapport om brukertilfredshet og tillit i koronasituasjonen fra 2020 har kun 9 prosent av brukerne erfaring med veiledning fra chatbot, og under halvparten av disse er fornøyde med denne løsningen (NAVs Personbrukerundersøkelse og Arbeidsgiverundersøkelse 2020). Mye kan derfor tyde på at chatbot-tjenesten ikke helt treffer, eller forstår brukernes behov.

3. Metode

Vår initielle datainnsamling er foregått gjennom en innledende interaksjon med ulike offentlige chatboter (Kommune-Kari og NAVs Frida). Dette ble utført med inspirasjon fra deltakende observasjon hvor vi selv inngikk som brukere for å observere den grenseflate brukere møtes med (Lazar et al., 2017). Her ble inntrykk notert og interaksjon dokumentert. Denne observasjonen ble brukt som utgangspunkt for videre undersøkelser.

Rammen for analysen i denne fasen tok utgangspunkt i WCAG 2.1 og Amershi et al. (2019)s 18 retningslinjer for design av interaksjon. Skriptet for denne første analyse er vedlagt i vedlegg 3.

Fra denne initielle analysen har fokus blitt innsnevret og videre i iterasjon 3 ønsker vi å gjennomføre en systematisk analyse av chatboten Frida. Dette skal vi gjøre ved å gjennomføre en innholdsanalyse av data vi har samlet inn gjennom interaksjon med chatboten (Lazar et al. 2017). Til dette har vi utformet et skript for å undersøke hvordan chatboten responderer til sammensatte brukersituasjoner og analysere de svar vi får tilbake fra systemet (se vedlegg 5). Vi vil videre lene oss på relevant akademisk litteratur for å diskutere og besvare vårt forskningsspørsmål, og komme med forslag til hvordan NAV sin interaksjon eventuelt kan forbedres for disse brukerne. Litteraturen vi kommer til å lene oss på er som følger:

- Artikkel: Guidelines for Human-AI Interaction av Amershi et al. (2019).

- Artikkel: Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems av Kocielnik et al. (2019).
- Masteroppgave: Bruk av chatbot i praksis av Steinstrøm, T. (2020).
- Egne analyser:
 - Analyse i Excel med WCAG 2.1 prinsipp 3 ('forståelse') hentet fra UUTilsynet samt (se vedlegg 3) og designretningslinjene fra Amershi et al. (2019).
 - Et felles skript for systematisk analyse av chatboten Frida med hovedtema sammensatte situasjoner

3.1 Eksisterende rammeverk og guidelines for design av nettinhold

I det følgende utdypes vi først hva WCAG 2.1 og Amershi et al. (2019)s retningslinjer er. Heretter presiserer vi hvilke retningslinjer fra Amershi et al. vi har valgt å ta med videre i analysen i iterasjon 3.

3.1.1 WCAG 2.1

WCAG 2.1 er et rammeverk fra 2018 (siden oppdatert) som har som formål å gjøre nettinhold tilgjengelig for alle slags brukere. Det vil si at rammeverket er ment for å sikre åpne og universelt utformet design av applikasjoner og nettsider. Rammeverket er bygget rundt 4 prinsipper (mulig å oppfatte, mulig å betjene, forståelig og robusthet) med underlagte retningslinjer som går mer spesifikt til verks for å veilede utviklere til å skape universelle design (UUTilsynet, 2021). Som det også ses nedenfor, så passer prinsippene for nettbasert innhold ikke nødvendigvis i forbindelse med design av AI. Vi har således benyttet oss av WCAG 2.1 i startfasen av prosjektet især til inspirasjon og guiden, men har valgt å ikke inkludere dette videre i neste analyse.

3.1.2 Amershi et al. (2019)s 18 guidelines

Amershi et al. (2019) argumenterer for at mangelen på forståelse om hva AI-systemene gjør, hvordan det fungerer og hvor ulikt AI kan tolke ulike situasjoner - kan føre med seg uheldige konsekvenser for brukerne. Ved design av AI-systemer lener man seg mye på prinsipper innen HCI-feltet, men forfatterne mener man bør utvikle egne retningslinjer for AI-baserte systemer (Amershi et al, 2019). Amershi et al. presenterer til slutt 18 retningslinjer for hvordan det kan være mulig å designe mer brukervennlige AI-systemer (se tabell 1). Etter

første analyse valgte vi å velge ut tre konkrete retningslinjer å jobbe videre med, som er markert i gult i tabellen under.

Initially	During interaction	When wrong	Over time
G1. Make clear what the system can do.	G3. Time services based on context.	G7. Support efficient invocation.	G12. Remember recent interactions.
G2. Make clear how well the system can do what it can do.	G4. Show contextually relevant information.	G8. Support efficient dismissal.	G13. Learn from user behavior.
	G5. Match relevant social norms.	G9. Support efficient correction.	G14. Update and adapt cautiously.
	G6. Mitigate social biases.	G10. Scope services when in doubt.	G15. Encourage granular feedback.
		G11. Make clear why the system did what it did.	G16. Convey the consequences of user actions.
			G17. Provide global controls.
			G18. Notify users about changes.

Tabell 1. Presentasjon av Amershi et al. 2019 sine evaluerte guidelines

Siden vi er interessert i å se næyere på G4, G10 og G12 (gulet ut i tabell 1), vil vi gå litt mer i detalj i hva de innebærer under.

G4: Show contextual relevant information

- Display information relevant to the user's current task and environment.

I.e., 'searching a movie title return show times in near my location for today's date'.

G10: Scope in services when in doubt

- Engage in disambiguation or gracefully degrade the AI system's services when uncertain about a user's goals

I.e., "It usually provides 3-4 suggestions instead of directly auto completing it for you"

G12: Remember recent interactions

- Maintain short term memory and allow the user to make efficient references to that moment.

I.e., “[The search engine] remembers the context of certain queries, with certain phrasing, so that it can continue the thread of the search (e.g., ‘who is he married to’ after a search that surfaces Benjamin Bratt)”

4. Foreløpige funn

4.1 Forventninger fra datainnsamlingen

Vi gjennomførte en forundersøkelse i Excel av NAV sin chatbot ved å liste opp WCAG 2.1 prinsipp 3 ‘forståelse’ (UUtilsynet, 2021) samt Amershi et al. (2019) sine 18 retningslinjer for design av menneske-ai interaksjon (s. 3). Dette hjalp oss blant annet med å skaffe en oversikt over hva hver av prinsippene betyr, samt snevre inn skopet for oppgaven. Vi oppdaget raskt at WCAG-prinsippene ikke nødvendigvis egner seg like godt til chatbot-interaksjon som til nettsider eller applikasjoner og at enkelte av Amershi et al. (2019) sine retningslinjer stakk seg mer ut i forhold til vårt forskningsspørsmål. Derfor endte vi opp med å fokusere spesifikt på tre av retningslinjene fra Amershi et al.: (G4, G10, G12). Denne forundersøkelsen ble også brukt som et grunnlag til å lage skriptet vi skal bruke til å gjennomføre en systematisk undersøkelse av chatboten.

Vår initielle funn fra forundersøkelsen viste at chatboten ikke nødvendigvis forstår brukeren, om denne ga uttrykk for å ikke forstå ulike begreper (dette var også vårt fokus for chatbot-oppgaven i Appendix 1, se vedlegg 1).

4.2 Presentasjon av funn

Siden vi fortsatt jobber med en til iterasjon av analyse, vil vi enda ikke presentere konkrete ‘funn’. I vår innledende pilotering av vårt nye skript (se vedlegg 3 + 4) ser vi dog antydninger til at vi kan endre fokus fra brukerens forståelse av chatboten, til chatbotens forståelse av brukeren og dennes ofte komplekse situasjon.

5. Tanker til fremtidig diskusjonsdel

Jobben med denne iterasjonen har medført overveielser som vi ønsker å ta opp i neste iterasjon. Disse overveielsene vil vi presentere her.

Menneskelige interaksjoner og samtaler skiller seg fra interaksjoner med maskiner på flere ulike vis. En av de mest åpenbare forskjellene er mangelen på fysisk tilstedeværelse i form av kroppsspråk, reaksjoner, mimikk og kontekst. I tillegg til dette er ikke chatboten i stand til undring og refleksjon - og det er heller ikke dens mål å bli kjent med deg. Dette henger sammen med chatboten sitt formål. Målet er 'problemløsning': å forstå et problem og gi deg en løsning på problemet. Den har derfor ikke et like stort behov for tilpasning til menneskelige relasjoner og interaksjoner. Ettersom chatboter blir stadig mer brukt som førstelinjekontakt, ser vi at det fortsatt er et stort forbedringspotensiale og behov for bedre tilpasninger til menneskelige samtaler og relasjoner, om denne type interaksjon skal fylde mer. F.eks vil et menneske kunne 'lese'/oppfatte når motparten sitter fast/sliter med noe og kanskje stille et oppfølgingsspørsmål for å undersøke om vedkommende har gjort seg forstått. Dette er utfordrende å ivareta for en chatbot, men kunne muligens imøtekommes med å legge inn oppfølgingsspørsmål av typen 'hvordan går det? Forstod du det jeg sa?' etter x antall minutter med stillhet fra brukeren.

Det er også i forbindelse med disse refleksjonene vi finner det relevant å trekke inn Kocielnik et al. (2019) og deres tanker om å justere brukeres forventninger til AI-baserte systemer. Dette vil undersøkes mer i dybden i iterasjon 3 etter at all datainnsamling har kommet på plass.

6. Konkluderende refleksjon

Gitt vårt sted i prosessen reserverer vi vår konkluderende refleksjoner til iterasjon 3.

Vedlegg 1 - Design av tidlig Chatbot-prototype

Introduksjon

Etter mye frem og tilbake i prosessen for å finne vårt fokus for oppgaven, landet vi på å jobbe konkret med å adressere prinsipp 3 i WCAG 2.1: 'Understanding'. Dette var et prinsipp vi så som overordnet oppfylt hos de chat-bottene vi så på i vår undersøkende fase (NAV, Kommune-Norges chatbot; Kommune-Kari og Skatteetaten). Vi startet dette prosjektet med å være interessert i inkludering og ekskludering som ledd i universiell utforming av AI. Med dette i bakhodet fant vi at forståelse av systemer også avhenger i høy grad av om hvorvidt man skjønner de begrepene og termer som brukes av eksempelvis offentlige etater og institusjoner. Det er i dette skjæringspunktet vi har lagt oss med denne chat-botten og denne annen iterasjonen av vårt prosjekt.

Prosess

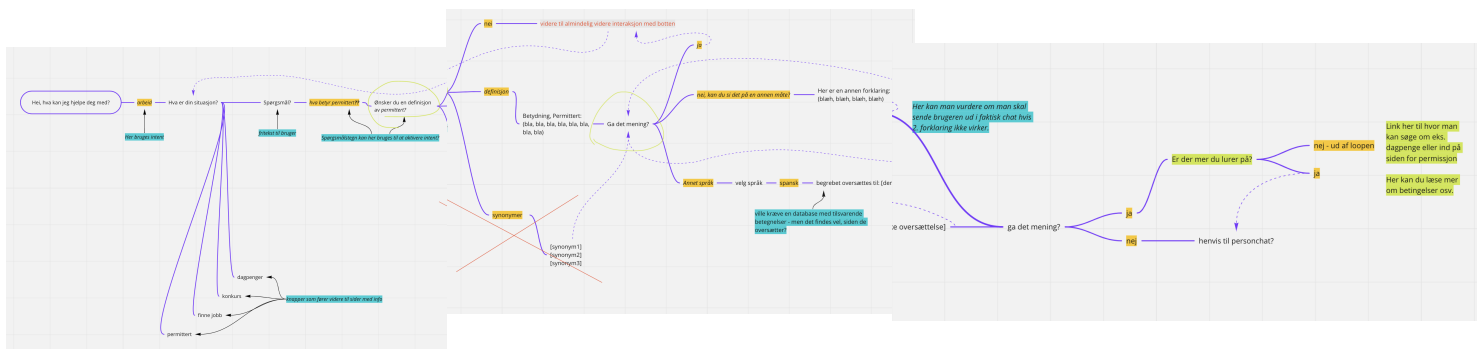
Vår chatbot er basert på en tidlig sondering foretatt av en av vår gruppe-medlemmer. I denne sonderingen undersøkte og testet hun ut hvordan interaksjonen var med de ulike chat-bottene. Denne tidlige sonderingen (som blir supplert av mer systematisk undersøkelse etter dette), ga oss inntrykk av hvor der kunne være et problem vi kunne adressere.

Sideløpende med dette utforsket vi Chatteron (og litt Dialogflow) for å se hvordan vi kunne jobbe med denne. For å få overblikk over flowet for interaksjonen for vår chat-bot, skisserte vi dette ved hjelp av Miro. Hovedfunksjonen til vår chatbot ble å gi brukerne muligheten for å få definisjoner på begreper som brukes i det offentlige (eks. 'Permittert' og 'Dagpenger') og muligheten for å få disse forklaringene oversatt for økt forståelse.

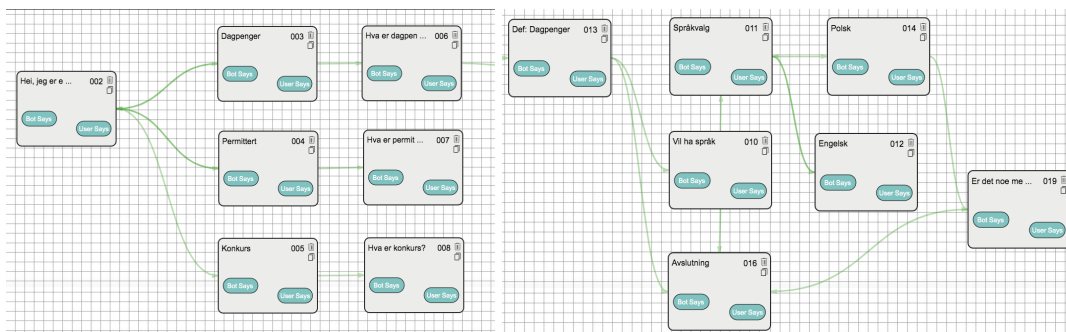
Refleksjon og Resultat

Chat-botten for denne iterasjonen er en tidlig prototype på hvordan man kan håndtere dette forklaringsaspekt for å fremme forståelse av det offentlige. Der gjenstår fortsatt arbeid for oss med å se dette feltet og hvordan dette prinsippet som ledd i omformingen av den offentlige sektoren mot universell utforming, jobbes med og blir adressert.

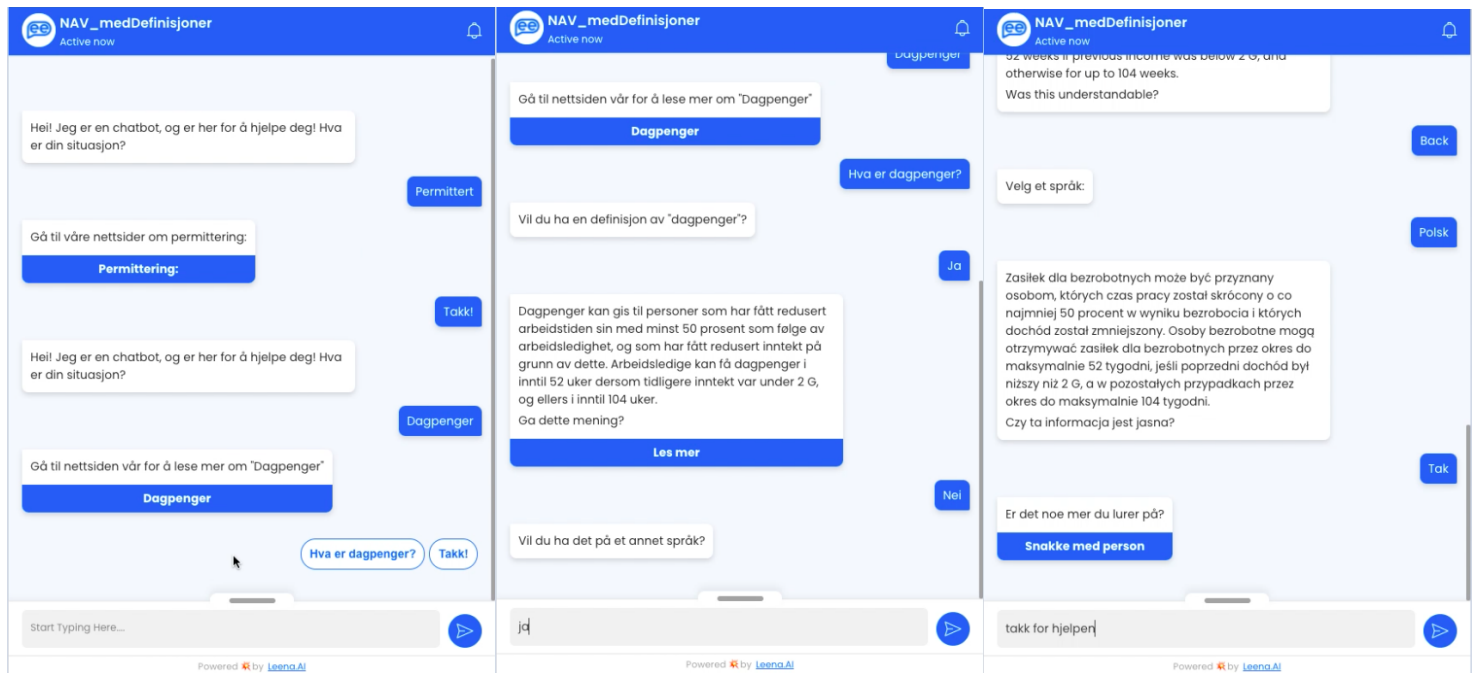
For hva er forståelse - og hvordan kan man kommunisere begreper så 'vanlige' borgere uten videre innsikt i offentlig sektor skjønner hva de innebærer og betyr? Hva skjer da om man legger på en språkbarriere - både for folk som kommer utenfra eller for folk som har enten bokmål eller nynorsk som sidemål? Jobben med å skissere vårt flyt var en øvelse i både å velge hva som minimum må med og hva som er for avanserte flyt som kan være gjenstand for videre jobbing. Dermed gjenstår jobben med å utvide og avgrense assistanse-skopet til chat-botten, innen vår valgte områder (språk og offentlig formidling), til en sådan grad at vårt flyt-nett kan fange mange potensielle misforståelser og feiltolkninger.



Figur 1: Skisse i Miro av et flyt i chat-botten



Figur 2: Skjermbilde av chat-bottens flyt i Chatteron



Figur 3: Skjerm bilde av interaksjonen med vår chatbot

Vedlegg 2 - Refleksjoner til maskinlæringsoppgaven

Forberedelser

Vi brukte en del tid bare på å forstå og tolke selve oppgaven, og hva som var krevd av oss for å løse den. Vi gjorde det også litt tungvint for oss selv, siden vi valgte å bruke andre tekstredigeringsprogrammer, slik som Atom, som vi selv hadde brukt tidligere, kontra å bruke Google Colab. Siden det ble mye knot og vi hadde diverse problemer med å kjøre koden, fikk vi ikke så mye tid til å teste og endre verdiene i koden.

Kjøring av kode og endringer av verdier

Vi begynte først med å teste ut chatboten etter de allerede satte verdiene i MovieChatbot.py-koden. Dette tenkte vi var et fint utgangspunkt for å vurdere effekten av ulike endringer av verdier. Vi valgte videre å kjøre en form moving, seeing, moving tilnærming for å se hvilke endringer i koden som ga mest endring i samtaleflyten og/eller intelligensen til chatboten.

1. Original setting

```
batch_size = 32@
max_words = 1000
epochs = 3

tempdata = open("movie_lines.txt", errors='replace').readlines()
x_train_temp = np.array([i.split("+++$+++")[-1].strip() for i in tempdata[:1000]])
y_train_temp = [i.split("+++$+++")[-2].strip() for i in tempdata[:1000]]

#Model
model = Sequential()
model.add(Dense(512, input_shape=(max_words,)))
model.add(Activation('relu'))
model.add(Dense(num_classes))
model.add(Activation('softmax'))

model.compile(loss='categorical_crossentropy',
              optimizer='adam',
              metrics=['accuracy'])

history = model.fit(x_train, y_train,
                  batch_size=batch_size,
                  epochs=epochs,
                  verbose=1,
                  validation_split=0.1)
print("Finished training")
```

Skjerm bilde 1 og 2: Originale verdier i chatbot-koden (inkludert epochs og dense).

```
Human:no
Chatbot:What's this?
Human:you are not cool
Chatbot:I know the porn career's a lie.
Human:what?
Chatbot:You decided.
Human:i did?
Chatbot:You're welcome.
Human:thank you?
Chatbot:Can we go now?
Human:no
Chatbot:The duck?
Human:dont curse
Chatbot:Just once. Afterwards, I told him I didn't want to anymore. I wasn't ready. He got pissed. Then he broke up wi
Human:he broke up with you?
Chatbot:Were you in jail?
Human:what are you talking about?
Chatbot:No, Daddy.
Human:i am not your daddy
Chatbot:That's what you want, isn't it?
Human:i am not sure what you mean
Chatbot:No, I won't go with you
Human:alright
Chatbot:Other than my upchuck reflex? Nothing.
Human:
```

Skjermbilde 3: Første samtale med chatboten med diverse tilbakemeldinger ("følger flowen")

For å utforske litt mer, hadde vi også et fokus på reaksjonen til chatboten, og/eller om den ville reagere annerledes dersom vi svarte likt basert på det den hadde skrevet. Ville chatboten havne i en 'loop', eller ville den svare annerledes selv med samme repeterte input? Vi fikk inntrykk av at output fra chatboten var utelukkende random og ikke var påvirket av input fra bruker.

```
Chatbot:My mission in life.
Human:life
Chatbot:Something like that
Human:life
Chatbot:I don't like to do what people expect. Then they expect it all the time and they get disappointed when you change
Human:life
Chatbot:You want me to climb up and show you how to get down?
Human:life
Chatbot:I know, just let me sleep
Human:life
Chatbot:I don't like to do what people expect. Then they expect it all the time and they get disappointed when you change
Human:life
Chatbot:So did you
Human:
```

Skjermbilde 4: Gjentakende respons fra menneske

Vi ville også teste om det skjedde noe annerledes dersom vi responderte med neste linje i tekstfilen (movie_lines.txt) for å se om dette ville skape en annen reaksjon fra chatboten. Dette forsterket inntrykket vårt om at responsen er helt tilfeldig.

```
L249 +++$+++ u5 +++$+++ m0 +++$+++ KAT +++$+++ William didn't even go to high school
L248 +++$+++ u6 +++$+++ m0 +++$+++ MANDELLA +++$+++ William would never have gone to a state school.
L247 +++$+++ u5 +++$+++ m0 +++$+++ KAT +++$+++ So he has this huge raging fit about Sarah Lawrence and insists that I go to his male-dominated, puking frat boy, number one golf team
school. I have no say at all.
L152 +++$+++ u6 +++$+++ m0 +++$+++ MANDELLA +++$+++ You could always go with me. I'm sure William has some friends.
L151 +++$+++ u5 +++$+++ m0 +++$+++ KAT +++$+++ The people at this school are so incredibly foul.

Chatbot:So he has this huge raging fit about Sarah Lawrence and insists that I go to his male-dominated, puking frat boy
Human:You could always go with me. I'm sure William has some friends.
Chatbot:Listen to you! You sound like Betty, all pissed off because Archie is taking Veronica.
Human:
```

Skjerm bilde 5 og 6: Neste linje som respons

2. Endring av Epoch

- Vår initielle tolkning var at dersom vi endret på “epochs” til et høyere tall, så ville det komme frem en tydeligere sammenheng mellom svarene til menneske og chatbot - men dette fant vi ikke særlig grunnlag for. Selv om vi da økte gjennomgangene av treningssettet, hadde vi ikke gjort i endringer i treningssettet i seg selv. En kombinasjon ville kanskje utløst en tydeligere endring i responsen til chatboten.

3. Endring av Dense

- Etersom “Dense” er det som bestemmer antall neuroner, tenkte vi at det ville være relevant å endre på dens verdier. Endring av verdien fra 512 til 256 eller 1024 endret ikke særlig på responsen, og vi tenker at en endring av dense-verdien alene ikke er nok til å skape endring - uten å inkludere andre faktorer.

Vår mini-undersøkelse ga altså ikke forventet output (større endringer), og vi vil derfor heller bruke litt mer plass og tid på å reflektere over undersøkelsen, fremfor å legge inn og beskrive mange skjermbilder. Tidsbegrensningen til oppgaven førte også til at vi ikke rakk å få mer kunnskap for å kunne gjøre større endringer i koden.

Refleksjon

Vårt første inntrykk av koden var at den var fin og ryddig. Et spørsmål vi satt med etter første titt var om det skulle være en sammenheng mellom det chatboten sier og det vi som bruker skriver inn. Vi merket en god del bruk av random - var det meningen at det ikke skulle være sammenheng mellom interaksjonen? Vårt videre inntrykk var at koden var skrevet statisk - og ikke dynamisk. Det vi mener her, er at chatboten ikke så ut til å lære av datasettet, men kun plukket ut tilfeldige linjer fra et ferdig manuskript. Vi antar at det er treningsdataen og/eller

funksjoner for å trene chatboten som vi burde endret på for å få en effekt i interaksjonen med chatboten. Vi burde derfor ikke bare hatt fokus på selve koden som eksisterte fra før.

Vi gjorde også mye research, og merket raskt at maskinlæring er et komplekst tema, med både en terminologi og logikk som er ukjent for oss. Dette gjorde at vi ikke alltid forstod sammenhengen mellom endringene vi forsøkte oss på, og forventet og faktisk output. Samtidig synes vi dette er et spennende fagfelt, og skulle ønske vi hadde hatt mer teknisk introduksjon til maskinlæring. Som designere merker vi nå at det er viktig å forstå det tekniske aspektet, så vel som det etiske.

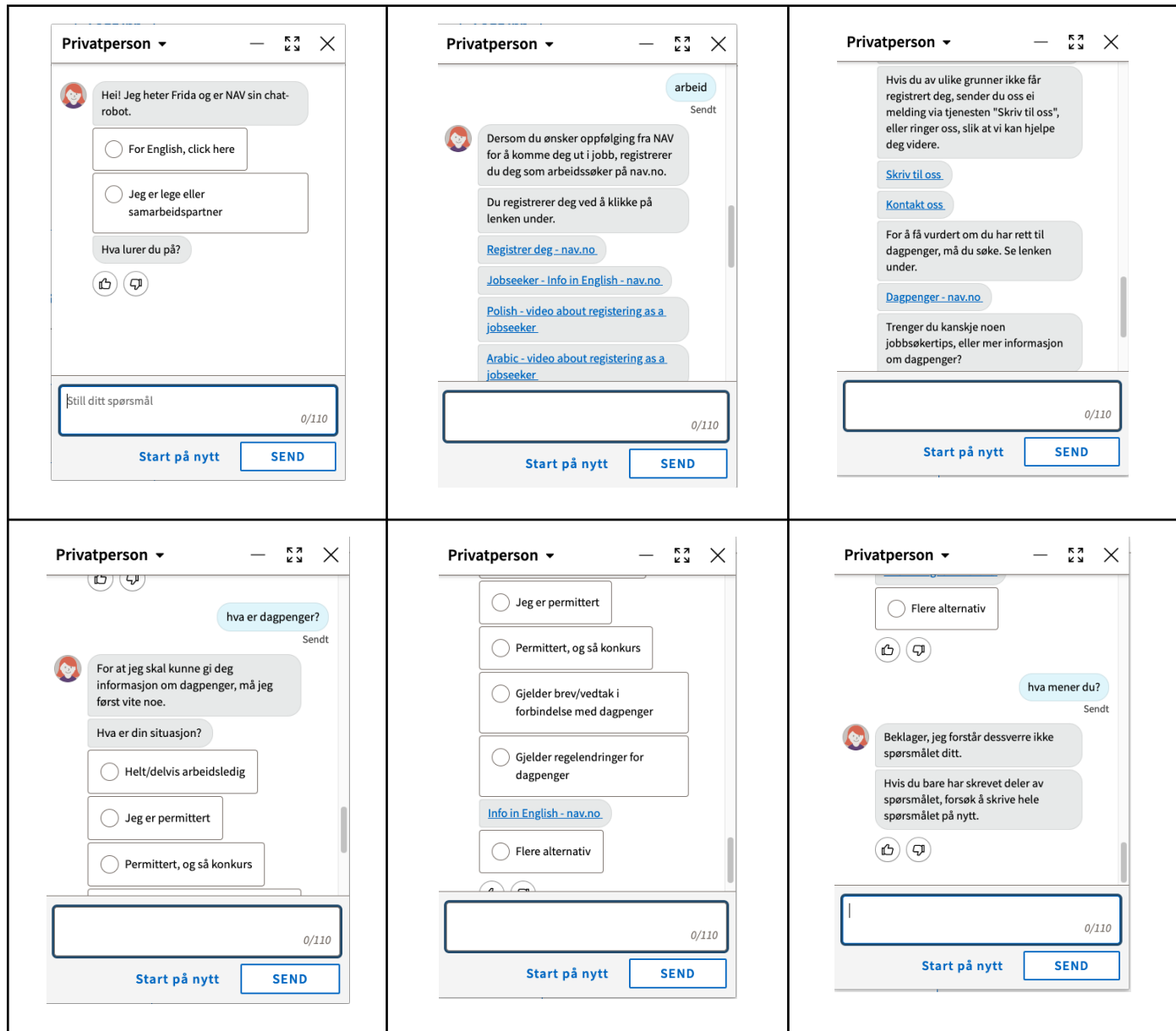
Vedlegg 3 - UU-analyse av NAV sin chatbot i Excel

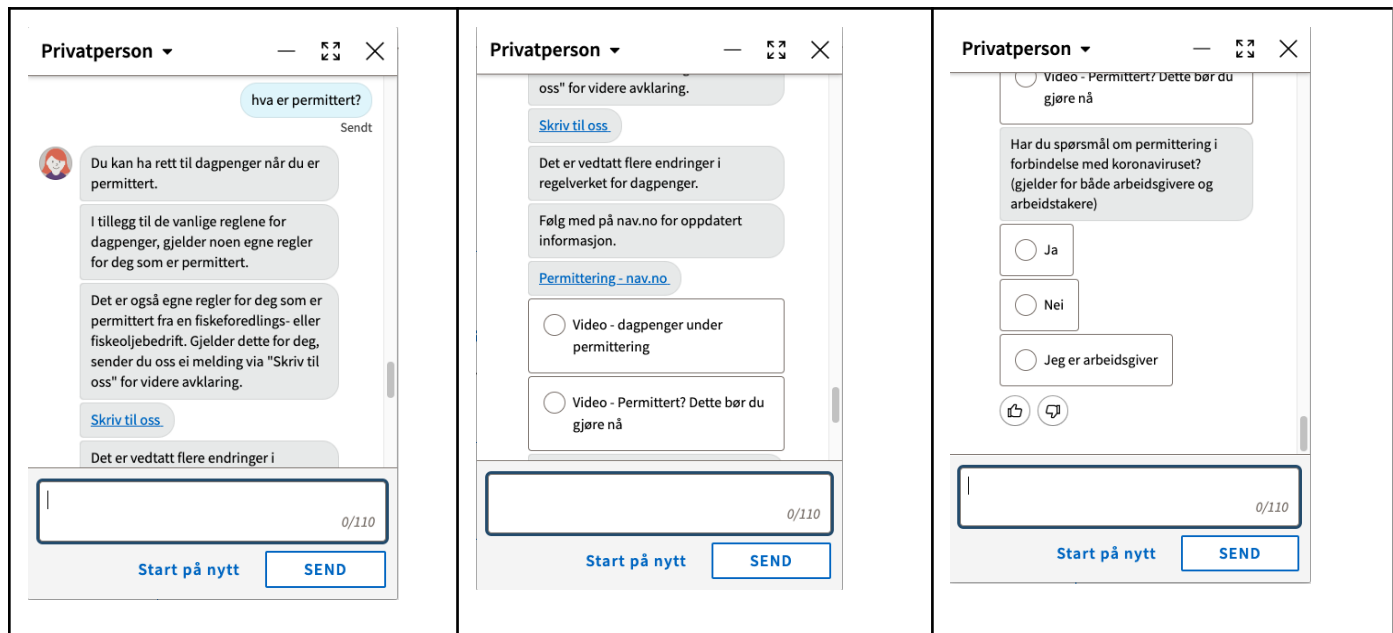
Initiell analyse av en interaksjon med chatboten Frida i forhold til WCAG 2.1 prinsipp 3 ('forståelse') samt Amershi et al. (2019) sine guidelines (2019, s. 3).

Kriterie	Forklaring	Plattform	Godkjent	Må utbredres	Kommentar	Verktøy	Skjermbilder (ligger i mappen)
Sjekk av NAVs chatbot ift WCAG prinsipp 3: forståelig + Amershi et al. (2019) sier		18 Human-AI design guidelines					
Prinsipp 3: Forståelig	Det må være mulig å forstå og betjene grensesnitt. Det vil si, både forstå hvordan chatboten brukes, og informasjonen som finnes.	Dataskrin (MacBook/Microsoft)				Manuelt	
3.1 Leselig	Gjør innholdet leselig og lett forståelig.		✓	✗	Både og, avhenger av flere faktorer	Manuelt	
3.1.1 Språk på siden - Nivå A	Sørge for at språket til innholdet på alle nettsider er angitt i koden.	Dataskrin (MacBook/Microsoft)			*Kan ikke vurderes av oss	Manuelt	
3.1.2 Språk på deler av innhold - Nivå AA	Sørge for at alle deler av innholdet som er på et annet språk enn resten av siden er markert i koden.	Dataskrin (MacBook/Microsoft)			*Kan ikke vurderes av oss	Manuelt	
3.2 Forutsigbar	Sørge for at nettsider presenteres og fungerer på forutsigbare måter.		✓			Manuelt	
3.2.1 Fokus - Nivå A	Når en komponent kommer i fokus medfører dette ikke automatisk betydelige endringer i siden.	Dataskrin (MacBook/Microsoft)	✓			Manuelt	
3.2.2 Inndata - Nivå A	Endring av verdien til et skjemafelt medfører ikke automatisk betydelige endringer i siden.	Dataskrin (MacBook/Microsoft)				Manuelt	
3.2.3 Konsekvens navigering - Nivå AA	Navigasjonslinker som gjentas på flere sider skal ha en konsekvent rekkefølge.	Dataskrin (MacBook/Microsoft)				Manuelt	
3.2.4 Konsekvens identifikasjon - Nivå AA	Elementer som har samme funksjonalitet på tvers av flere sider er utformet likt.	Dataskrin (MacBook/Microsoft)	✓		Knapper, tekstformat etc er likt.	Manuelt	
3.3 Inndatahjelp							
3.3.1 Identifikasjon av feil - Nivå A	For feil som oppdages automatisk må du vise hvor feilen har oppstått og gi en tekstbeskrivelse av feilen.	Dataskrin (MacBook/Microsoft)				Manuelt	
3.3.2 Ledetekster eller instruksjoner - Nivå A	Det vises ledetekster eller instruksjoner når du har skjemaelementer som må fylles ut.	Dataskrin (MacBook/Microsoft)	✓	✗	Både og, chatboten er god på å referere til gitte sider som kan løse utfordringen, men ikke så god til å forklare hva det egentlig betyr	Manuelt	
3.3.3 Forslag ved feil - Nivå AA	Dersom feil blir oppdaget automatisk, gi brukeren et forslag til hvordan feilen kan rettes.	Dataskrin (MacBook/Microsoft)		✗	Kan bli bedre på å foreslå alternativer ved feil. F.eks ved forsøk på å skrive "sjønner" var responsen: "Beklager men jeg forstår ikke spørsmålet ditt." Mao den vil ikke fungere godt for dyslektikere	Manuelt	
3.3.4 Forhindring av feil (juridiske feil, økonomiske feil, datafeil) - Nivå AA	For sider som medfører juridiske forpliktelser må det være mulig å kunne angre, kontrollere eller bekrefte dataene som sendes inn.	Dataskrin (MacBook/Microsoft)	✓		Forslagene som gis er i tråd med lovverket, som er det NAV bruker ift. stønader m.m. Brukeme blir bedt om å ikke oppgi personidentifiserende info i botten.	Manuelt	
3.4 Robust	Innholdet må være robust nok til å kunne tolkes på en pålitelig måte av brukere, inkludert kompensierende teknologi (omhandler koding)				*Ikke relevant for oss		

Amershi et al (2018) Guidelines for Human-AI Interaction (p. 3).						
G1 Make clear what the system can do. Help the user understand what the AI system is capable of doing.	[Activity Trackers, Product #1] "Displays all the metrics that it tracks and explains how. Metrics include movement metrics such as steps, distance traveled, length of time exercised, and all-day calorie burn, for a day."	Dataskin (MacBook/Microsoft)	✓			Manuelt
G2 Make clear how well the system can do what it can do. Help the user understand how often the AI system may make mistakes.	[Music Recommenders, Product #1] "A little bit of hedging language: 'we think you'll like'."	Dataskin (MacBook/Microsoft)	✓			Manuelt
G3 Time services based on context. Time when to act or interrupt based on the user's current task and environment.	[Navigation, Product #1] "In my experience using the app, it seems to provide timely route guidance. Because the map updates regularly with your actual location, the guidance is timely."	Dataskin (MacBook/Microsoft)		✗	Kan bli bedre på å spørre 'er du der?' når brukeren bruker lang tid på å svare evt foreslå "forstod du det jeg sa?" og en ja/nei knapp + forklaring	Manuelt
G4 Show contextually relevant information. Display information relevant to the user's current task and environment.	[Web Search, Product #2] "Searching a movie title returns show times in near my location for today's date"	Dataskin (MacBook/Microsoft)	✓	✗	Viser ofte relevant info om det man spør om hvis man vet hva man skal lete etter	Manuelt
G5 Match relevant social norms. Ensure the experience is delivered in a way that users would expect, given their social and cultural context.	[Voice Assistants, Product #1] "[The assistant] uses a semi-formal voice to talk to you - spells out 'okay' and asks further questions."	Dataskin (MacBook/Microsoft)	✓	✗	Vanskelig å svare utdelt ja/nei på. Er imidlertid sikker på at den kan bli mye bedre på dette.	Manuelt
G6 Mitigate social biases. Ensure the AI system's language and behaviors do not reinforce undesirable and unfair stereotypes and biases.	[Autocomplete, Product #2] "The autocomplete feature clearly suggests both genders [him, her] without any bias while suggesting the text to complete."	Dataskin (MacBook/Microsoft)	✓		Den virker ganske "nøytral" ved kortidsbruk	Manuelt
G7 Support efficient invocation. Make it easy to invoke or request the AI system's services when needed.	[Voice Assistants, Product #1] "I can say [wake command] to initiate."	Dataskin (MacBook/Microsoft)	✓		Let tilgjengelig på siden, og aktiveres raskt.	Manuelt
G8 Support efficient dismissal. Make it easy to dismiss or ignore undesired AI system services.	[E-commerce, Product #2] "Feature is unobtrusive, below the fold, and easy to scroll past... Easy to ignore."	Dataskin (MacBook/Microsoft)	✓			Manuelt
G9 Support efficient correction. Make it easy to edit, refine, or recover when the AI system is wrong.	[Voice Assistants, Product #2] "Once my request for a reminder was processed I saw the ability to edit my reminder in the UI that was displayed. Small text underneath stated 'Tap to Edit' with a chevron indicating something would happen if I selected this text."	Dataskin (MacBook/Microsoft)		✗	Du kan ikke redigere tidligere input, men du kan laste ned samtalen.	Manuelt
G10 Scope services when in doubt. Engage in disambiguation or gracefully degrade the AI system's services when uncertain about a user's goals.	[Autocomplete, Product #1] "It usually provides 3-4 suggestions instead of directly auto-completing it for you"	Dataskin (MacBook/Microsoft)		✗	Hvis de hadde arbeidet mer med å tilby forklaring av begreper hadde denne kanskje blitt godkjent.	Manuelt
G11 Make clear why the system did what it did. Enable the user to access an explanation of why the AI system behaved as it did.	[Navigation, Product #2] "The route chosen by the app was made based on the Fastest Route, which is shown in the subtext."	Dataskin (MacBook/Microsoft)	✓			Manuelt
G12 Remember recent interactions. Maintain short term memory and allow the user to make efficient references to that memory.	[Web Search, Product #1] "[The search engine] remembers the context of certain queries, with certain phrasing, so that it can continue the thread of the search (e.g., 'who is he married to' after a search that surfaces Benjamin Bratt)"	Dataskin (MacBook/Microsoft)	✓			Manuelt
G13 Learn from user behavior. Personalize the user's experience by learning from their actions over time.	[Music Recommenders, Product #2] "I think this is applied because every action to add a song to the list triggers new recommendations."	Dataskin (MacBook/Microsoft)		✗	Målet til denne chatboten er ikke personalisering.	Manuelt
G14 Update and adapt cautiously. Limit disruptive changes when updating and adapting the AI system's behaviors.	[Music Recommenders, Product #2] "Once we select a song they update the immediate song list below but keeps the above one constant."	Dataskin (MacBook/Microsoft)			*Ikke relevant for oss	Manuelt
G15 Encourage granular feedback. Enable the user to provide feedback indicating their preferences during regular interaction with the AI system.	[Email, Product #1] "The user can directly mark something as important, when the AI hadn't marked it as that previously."	Dataskin (MacBook/Microsoft)		✗		Manuelt
G16 Convey the consequences of user actions. Immediately update or convey how user actions will impact future behaviors of the AI system.	[Social Networks, Product #2] "[The product] communicates that hiding an Ad will adjust the relevance of future ads."	Dataskin (MacBook/Microsoft)		✗	*Ikke en feature i denne chatboten.	Manuelt
G17 Provide global controls. Allow the user to globally customize what the AI system monitors and how it behaves.	[Photo Organizers, Product #1] "[The product] allows users to turn on your location history so the AI can group photos by where you have been."	Dataskin (MacBook/Microsoft)		✗	*Ikke en feature i denne chatboten.	Manuelt
G18 Notify users about changes. Inform the user when the AI system adds or updates its capabilities.	[Navigation, Product #2] "[The product] does provide small in-app teaching callouts for important new features. New features that require my explicit attention are oob-us."	Dataskin (MacBook/Microsoft)			*Fikk ikke testet	Manuelt

Vedlegg 4 - Skjermbilder av initiell interaksjon med chatboten





Vedlegg 5 - Manuskript

1. **BOT: Hei jeg heter Frida og er NAV sin chat-robot! Hva lurere du på? Knapper**

<for English, click here>

<Jeg er lege eller samarbeidspartner>

<Uføretrygd etteroppgjør>

Bruker: Hei!

2. **BOT: Hei! Hva kan jeg hjelpe deg med?**

Bruker: Jeg er blitt arbeidsledig, hva gjør jeg nå?

3. **BOT: Dersom du ønsker oppfølging fra NAV for å komme deg ut i jobb, registrerer du deg som arbeidssøker på nav.no. Du registrerer deg ved å klikke på lenken under Registrer deg - nav.no.**

Jobseeker - Info in English - nav.no

Polish - video about registering as a jobseeker

Arabic - video about registering as jobseeker

Hvis du av ulike grunner ikke får registrert deg, sender du oss ei melding via tjenesten "Skriv til oss", eller ringer oss, slik at vi kan hjelpe deg videre.

Se lenken under.

Skriv til oss

Kontakt oss

For å få vurdert om du har rett til dagpenger, må du søke.

Se lenken under.

Dagpenger - nav.no

Trenger du kanskje noen jobbsøkertips, eller mer informasjon om dagpenger?

Knapp med

<ja>

<nei>

Bruker: <bruker velger dagpenger>

4. BOT: Du må velge din situasjon. knapp med

<helt/delvis arbeidsledig>

<permittert så konkurs>

Bruker: Velger <ja>

5. BOT: Hva ønsker du informasjon om? Knapp med

<jobbsøkertips/dagpenger>

Bruker: Velger <dagpenger>

6. BOT: For at jeg skal kunne gi deg informasjon om dagpenger, må jeg først vite noe. Hva er din situasjon? Knapp med:

<helt/delvis arbeidsledig>

<jeg er permittert>

<permittert, så konkurs>

<gjelder brev/vedtak i forbindelse med dagpenger>

Info in english.

Knapp med <Flere alternativ>

Bruker: Velger <Flere alternativ>

7. BOT: Kan noen av alternativene under kan være til hjelp? Knapper med:

<Arbeidsplassen er konkurs>

<Jeg er selvstendig næringsdrivende>

<Jeg er fisker>

<Dagpenger ved streik>

<Dagpenger etablering av virksomhet>

<Dagpenger ved utdanning>

<Dagpenger dimittert>

<Dagpenger offshorearbeider>

Bruker: hvilke rettigheter har jeg som arbeidsledig?

BOT: responderer helt likt som i punkt 3 (så ingen ny info, kun det samme)

Bruker: jeg er både student, arbeidsledig og alenemor?

BOT: Jeg er ikke sikker på om jeg forsto alt du spurte om, men jeg tror jeg kan hjelpe deg med: (Knapper med)

<Enslig forsørger>

<Arbeidssøker>

<Student>

Vedlegg 6 - Feedback-håndtering

På vårt Wonderdokument fikk vi tilbakemelding om at vi kunne spesifisert litt bedre vår metodologisk tilnærming til oppgaven.

Vi har jobbet med å spesifisere dette til den grad det har vært mulig for oss hittil. Vi har dog brukt tilbakemeldingen som en form for motivator for å jobbe med skiptet og systematisk sette opp retningslinjene på en måte så det var mulig å undersøke alle på en god måte.

Referanser

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Teevan, J. (2019). Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (paper no. 3). ACM.
<https://www.microsoft.com/en-us/research/uploads/prod/2019/01/Guidelines-for-Human-AI-Interaction-camera-ready.pdf>

Brandtzaeg, P. & Følstad, A. (2017) Why people use chatbots. *International Conference on Internet Science*. pp. 377-392.

Kocielnik, R., Amershi, S., & Bennett, P. N. (2019). Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (paper no. 411). ACM.
https://www.microsoft.com/en-us/research/uploads/prod/2019/01/chi19_kocielnik_et_al.pdf

Lazar, J. Feng, J. & Hochheiser, H. (2017). *Research Methods in Human-Computer Interaction*. Morgan Kaufmann, 2. ed.

Luger, E., & Sellen, A. (2016). Like having a really bad PA: the gulf between user expectation and experience of conversational agents. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 5286-5297). ACM.

Moore, J. R. (2018). Chapter 9 - A Natural Conversation Framework for Conversational UX Design. IBM Research. *Studies in Conversational UX Design*, Human-Computer Interaction Series. Springfield International Publishing. DOI:[10.1145/3304087](https://doi.org/10.1145/3304087)

Nav.no, a. (2021, 12. januar). 193 800 arbeidssøkere registrert hos NAV.
<https://www.nav.no/no/nav-og-samfunn/statistikk/arbeidssokere-og-stillinger-statistik/nyheter/193-800-arbeidssokere-registrert-hos-nav>

NAV's Personbrugerundersøkelse og Arbeidsgiverundersøkelse 2020: *Brukertilfredshet og tillit i koronasituasjonen*

https://www.nav.no/_/attachment/download/913acd81-00fb-4908-9fcb-6da9c8338551:240f0f23b8528ef977d3659c48ca42c31b9b7516/NAV%20brugerunders%C3%B8kelse%202020.pdf

Steinstø, T (2020) *Bruk av chatbot i praksis. En kvalitativ studie av utfordringer ved bruk av chatbot i offentlige tjenester* [Masteroppgave, Universitetet i Oslo]. DUO Vitenarkiv.

<https://www.duo.uio.no/bitstream/handle/10852/79645/Master-Tina-Steinst--NY.pdf?sequence=1&isAllowed=y>

Store Norske Leksikon, "Nav", hentet 18.10.21 fra

<https://snl.no/NAV>

UUTilsynet (2021). Oppbygging av WCAG 2.1. Hentet 05. oktober 2021 fra

<https://www.uutilsynet.no/wcag-standard/oppbygging-av-wcag-20/139>