

My background:

Human-Chatbot Interaction Design

How to design chatbots?

Supported by IKTPLUSS, Research Council of Norway



Chatbots og interaksjonsdesign

Naturlig språk blir en stadig vanligere måte å bruke digitale tjenester på. Eksempler er Apples Siri og Googles Assistant. Facebook Messenger gir i dag tilgang til mer enn 30.000 chatbots.

Til tross for en rivende teknologiske utvikling innen chatbots gjenstår mange forskningsutordringer. I prosjektet **Human-Chatbot Interaction Design** kombinerer vi interaksjonsdesign med kunstig intelligens og maskinlæring.

Målet er å bidra til kunnskap om brukersentrert design av chatbots, slik at disse kan bli nyttige og engasjerende innganger til informasjon og tjenester – uansett brukerens kjønn, bakgrunn og alder.

OM PROSJEKTET

Prosjektperiode 2017-2021

Program IKTPLUSS

PARTNERE





STØTTET AV



Kontakt: Prosjektleder Asbjørn Følstad | 93293013 | asf@sintef.no

Les mer: www.sintef.no/humanchatbotinteractiondesign

Interaction with AI – module 2

Interaction design

Five sessions

Human – Al relationships

Marita Skjuve

September 22

Design of interaction with AI

Asbjørn Følstad

Understanding interaction with AI

Morten Goodwin

September 29 (online)

October 6 (online)

October 13 (hybrid)

October 20 (online)

Literature

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Teevan, J. (2019). Guidelines for human-Al interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (paper no. 3). ACM.

Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, M. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623). ACM

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.

Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020). Re-examining Whether, Why, and How Human-Al Interaction Is Uniquely Difficult to Design. In Proceedings of the 2020 CHI conference on human factors in computing systems (Paper no. 164).

Følstad, A., & Brandtzæg, P. B. (2017). Chatbots and the new world of HCI. interactions, 24(4), 38-42.

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.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: how susceptible are jobs to computerisation?. *Technological forecasting and social change*, 114, 254-280.

Carter, S., & Nielsen, M. (2017). Using artificial intelligence to augment human intelligence. *Distill*, 2(12), e9.

Wilson, H. J., & Daugherty, P. R. (2018). Collaborative intelligence: humans and AI are joining forces. *Harvard Business Review*, *96*(4), 114-123.

Noessel, C. (2017). Designing Agentive technology: Al that works for people. Rosenfeld Media.

Hall, E. (2018). Conversational design. A Book Apart

Individual assignment

Building on and extending the individual assignment in Module 1. Startup today – finish October 22)

Mid-term report (Oct. 22)

Group assignment

Building on and extending the group assignment in Module 1. Startup today – finish October 22)

- + Task on user and interaction design perspective (startup today finish October 6, include in mid-term report as Appendix 1)
- + Task on machine learning perspective (startup Oct. 6 – finish Oct. 13, include in midterm report as Appendix 2)

Midterm report - individual assignment

Three topics:

- Characteristics of Al-infused systems.
- Human-Al interaction design.
- Chatbots / conversational user interfaces.

Language: English or Norwegian.

Max. pages: 6

Min. articles referenced 4.



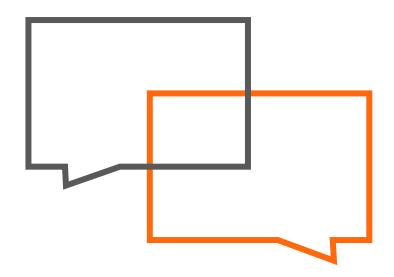
Midterm report – group assignment

Content – 5-7 pages

- A description of the group, who you are names.
- A description of what area of "interaction with AI" you are interested in working with.
- (new) Background section: Position your work relative to existing knowledge and practice
- Minimum 1 maximum 2 questions that you want to address.
 Please write some sentences about the questions. These questions can change and evolve later in the midterm report and in the final report as you go about investigating your questions.
- **(updated)** Method section overall approach, design process (optional, but encouraged), data collection methods
- (new) Sketches and/or prototypes (optional, but encouraged)
- (new) Findings (progress, initial outcomes)
- (updated) Minimum five references to literature.

Appendices – approx. 1 page each

- Appendix 1: Chatbot design task briefly describe the process and outcome. Detail reflections and lessons learnt.
- Appendix 2: Machine learning task briefly describe the process and outcome. Detail reflections and lessons learnt.



Brief status on the group task – each group say a few words

Group 1 – Al-based systems and universal design

- 1. Are there good resources/literature to support and depend on while developing AI-systems for universal use?
- 2. How are interactions taken into account before / during / after development?

Group 2 – Inclusion and exclusion with Al

- 1. What potential does AI have for including and excluding people in the society?
- 2. Does AI effect gender-non conforming individuals differently than genderconforming individuals, and if so in what way?

Group 3 – social credit systems

- 1. How is Al used in the Chinese Social Credit System?
- 2. To what degree are the inhabitants willing to accept a Social Credit System in Norway?

Group 4 – AI i undervisning og studier

- 1. Hvordan kan kunstig intelligens hjelpe og motivere studenter med å organisere hverdagen bedre?
- 2. Hvordan kan kunstig intelligens bidra til økt trivsel hos studentene?

Group 5 – Al og skjønnsmessige vurderinger

1. Hvilke etiske problemstillinger kan oppstå i forbindelse med AI?

Group 6 – Universal design and Human-Al interaction

- 1. How does language and pronunciation work together with chatbots to make sure universal design is in place?
- 2. How can Al improve the use of screen readers in a chatbot?

Group 7 – User perceptions of AI-machines

- 1. Why people decide to not chat with chatbots?
- 2. What deciding factors make chatbots successful?

Agenda

Today

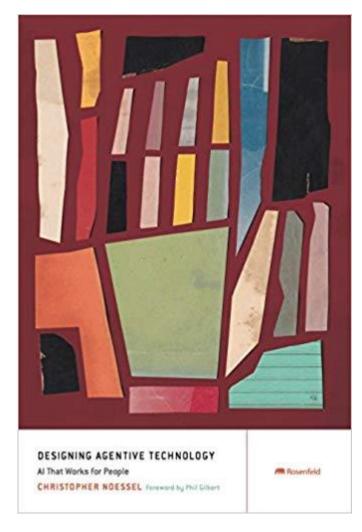
Next

- 1 Interacting with AI an overview
- Chatbots interacting with AI in natural language

- User-centred design of Al
- User-centred design of chatbots

Interaction with AI - overview

AI?



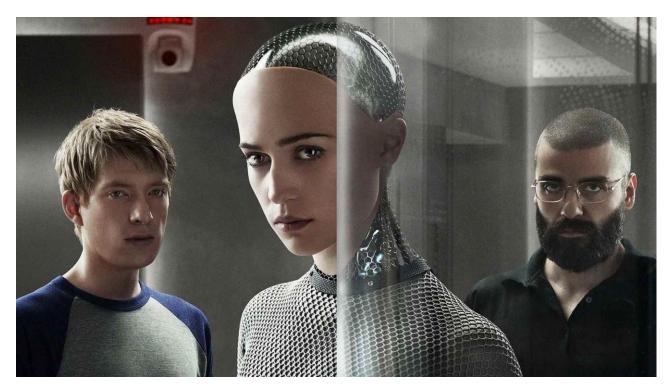


Devs (2020)

Artificial super intelligence

Artificial general intelligence

Artificial narrow intelligence

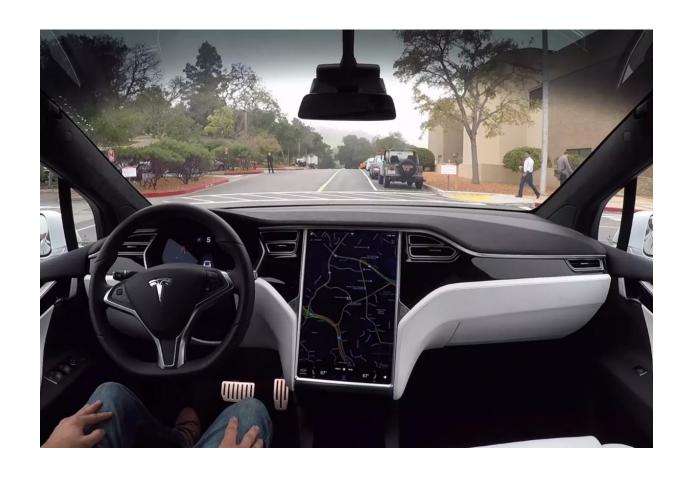


Ex Machina (2014)

Artificial super intelligence

Artificial general intelligence

Artificial narrow intelligence



Artificial super intelligence

Artificial general intelligence

Artificial narrow intelligence

Interaction with AI, and designing for interactions with AI, concerns artificial narrow intelligence (narrow AI)

Artificial super intelligence

Artificial general intelligence

Artificial narrow intelligence

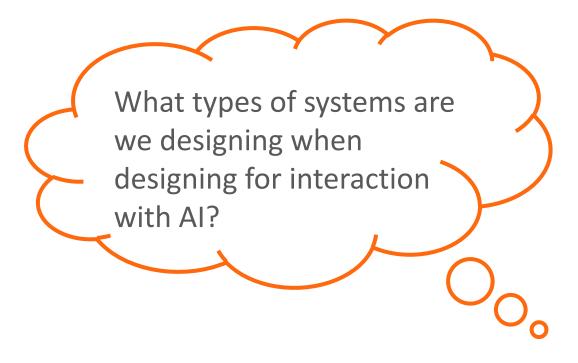
Interaction with AI, and designing for interactions with AI, concerns artificial narrow intelligence (narrow AI)

Need to know:

- Characteristics of narrow AI
- Opportunities and limitations of narrow Al

... from now on, when talking about Al we mean narrow Al

Al-infused systems



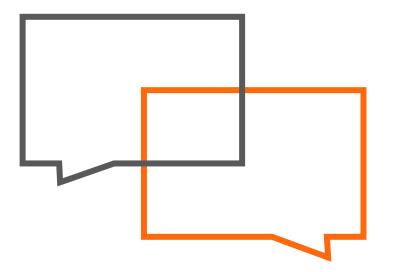
Al-infused systems

"Systems that have features harnessing AI capabilities that are directly exposed to the end user."

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Teevan, J. (2019). Guidelines for human-Al interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (paper no. 3). ACM.

Al-infused systems

favouriteexamples?











can a neural network learn to recognize doodling?

Help teach it by adding your drawings to the world's <u>largest doodling data set</u>, shared publicly to help with machine learning research.

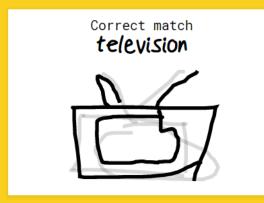


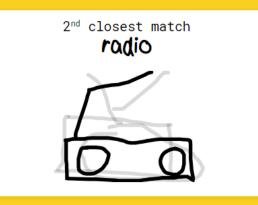
You were asked to draw television

You drew this, and the neural net recognized it.



It also thought your drawing looked like these:







Learning

Improving

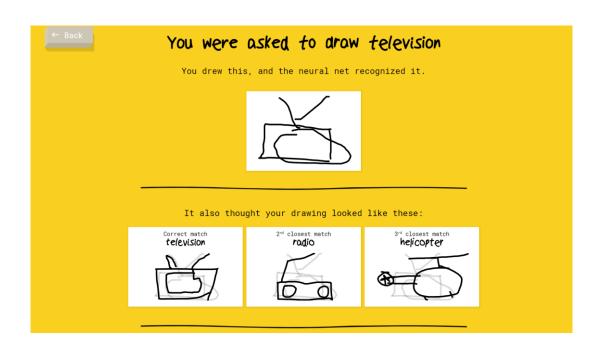
Black box | Fuelled by large data sets

Dynamic

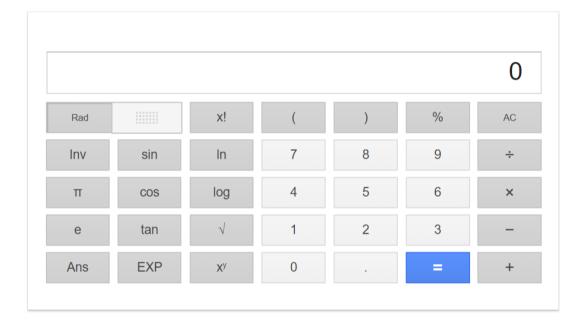
Mistakes inevitable

Opaque

Data gathering through interaction



VS.



Amershi et al (2019)

Definition of Al-infused systems: Systems that have features harnessing Al capabilities that are directly exposed to the end user

Uncertainty -> errors common, both false
positives and negatives

Inconsistency -> sensitive to context and small
changes in input

Behind the scenes personalization (e.g. automated filtering) -> potentially costly information hiding

Guidelines for Human-Al Interaction

Saleema Amershi, Dan Weld'[†], Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz

Microsoft
Redmond, WA, USA
{samershi, mivorvor, adamfo, benushi, pennycol, jinsuh,
shamsi, pauben, kori, teevan, ruthkg, horvitz}
@microsoft.com

ABSTRACT

Advances in artificial intelligence (AI) frame opportunities and challenges for user interface design. Principles for human-AI interaction have been discussed in the human-computer interaction community for over two decades, but more study and innovation are needed in light of advances in AI and the growing uses of AI technologies in human-facing applications. We propose 18 generally applicable design guidelines for human-AI interaction. These guidelines are validated through multiple rounds of evaluation including a user study with 49 design practitioners who tested the guidelines against 20 popular AI-infused products. The results verify the relevance of the guidelines over a spectrum of interaction scenarios and reveal gaps in our knowledge, highlighting opportunities for further research. Based on the evaluations, we believe the set of design guidelines can serve as a resource to practitioners working on the design of applications and features that harness AI technologies, and to researchers interested in the further development of guidelines for human-AI interaction design.

CCS CONCEPTS

Human-centered computing → Human computer interaction (HCI);
 Computing methodologies → Artificial intelligence.

Work done as a visiting researcher at Microsoft Research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org. CHI 2019, May 4–9, 2019, Clargove Scotland Uk

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-5970-2/19/05...\$15.00 https://doi.org/10.1145/3290605.3300233 [†]Paul G. Allen School of Computer Science & Engineering University of Washington Seattle, WA, USA weld@cs.washington.edu

KEYWORD

Human-AI interaction; AI-infused systems; design guidelines

ACM Reference Format:

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-Al Interaction. In CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4-9, 2019, Glasgow, Scotland Uk. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3290605.3300233

1 INTRODUCTION

Advances in artificial intelligence (AI) are enabling developers to integrate a variety of AI capabilities into user-facing systems. For example, increases in the accuracy of pattern recognition have created opportunities and pressure to integrate speech recognition, translation, object recognition, and face recognition into applications. However, as automated inferences are typically performed under uncertainty, often producing false positives and false negatives, AI-infused systems may demonstrate unpredictable behaviors that can be disruptive, confusing, offensive, and even dangerous. While some AI technologies are deployed in explicit, interactive uses, other advances are employed behind the scenes in proactive services acting on behalf of users such as automatically filtering content based on inferred relevance or importance. While such attempts at personalization may be delightful when aligned with users' preferences, automated filtering and routing can be the source of costly information hiding and actions at odds with user goals and expectations.

Al-Infused systems¹ can violate established usability guidelines of traditional user interface design (e.g., [31, 32]). For example, the principle of consistency advocates for minimizing unexpected changes with a consistent interface appearance and predictable behaviors. However, many AI components are inherently inconsistent due to poorly understood,

¹In this paper we use Al-infused systems to refer to systems that have features harnessing AI capabilities that are directly exposed to the end user.

Kocielnik et al (2019)

Probabilistic – almost always operate at less than perfect accuracy

Impacted by user actions – such as user-generated content

Transparency issues – how to mitigate? e.g. by showing decision rules

Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems



Figure 1: Expectation setting design techniques used prior to interaction with the Scheduling Assistant - an AI system for meeting request detection from free-text of emails. A) Accuracy Indicator - directly communicating to the user the expected accuracy of the AI component, B) Example-based Explanation - helping the user understand the basic principles of how the systems detects meeting requests, C) Control - giving the user control over AI decision making process through detection threshold adjustment.

ABSTRACT

Al technologies have been incorporated into many end-user applications. However, expectations of the capabilities of such systems vary among people. Furthermore, bloated expectations have been identified as negatively affecting perception and acceptance of such systems. Although the intelligibility of ML algorithms has been well studied, there has been little work on methods for setting appropriate expectations before the initial use of an Al-based system. In this work, we use a Scheduling Assistant - an Al system for automated meeting request detection in free-text email - to study the impact of several methods of expectation setting. We explore two versions of this system with the same 50% level of accuracy of the Al component but each designed with a different focus on the types of errors to avoid (avoiding False Positives vs. False Negatives). We show that such different

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with reedit is permitted. To copy other wise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org. CH 2019 May 4-2 2019 Classow Scotland US.

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-5970-2/19/05...\$15.00 https://doi.org/10.1145/3290605.3300641 focus can lead to vastly different subjective perceptions of accuracy and acceptance. Further, we design expectation adjustment techniques that prepare users for AI imperfections and result in a significant increase in acceptance.

CCS CONCEPTS

 Human-centered computing → Empirical studies in HCI; Empirical studies in visualization; Laboratory experiments:

KEYWORD

AI infused systems, AI system on-boarding, Shaping AI expectations, Perception and Acceptance of AI

ACM Reference Forma

Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect Al? Exploring Designs for Adjusting Enduser Expectations of Al Systems. In CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glazgow, Scotland Uk. ACM, New York, NY, USA, 14 pages. https: //doi.org/10.1145/3290605.3300641

1 INTRODUCTION

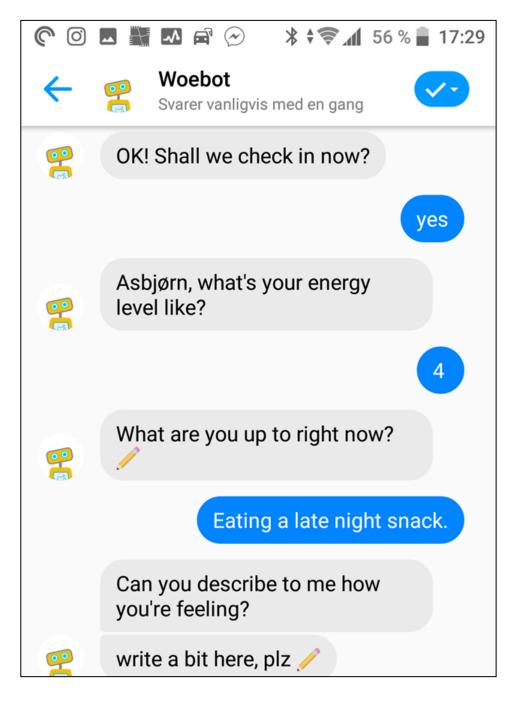
Expectations impact how accepting end-users are of the technologies they use. For example, inflated expectations about usability and ease of use have been shown to decrease user satisfaction and willingness to use products when those expectations are not met [20, 36]. Artificial intelligence (AI) introduces additional factors impacting user expectations

Chatbots – interacting with Al in natural language

Chatbots?



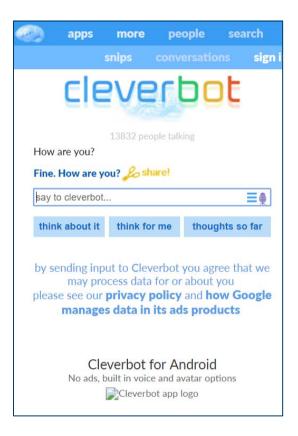
Chatbots?



One definition of chatbots:

... machine agents that serve as natural language user interfaces to data and services through text or voice.

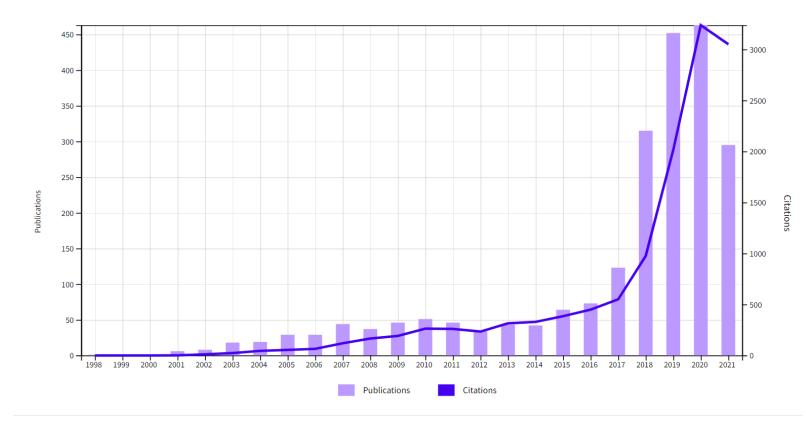
Meaning of term evolving



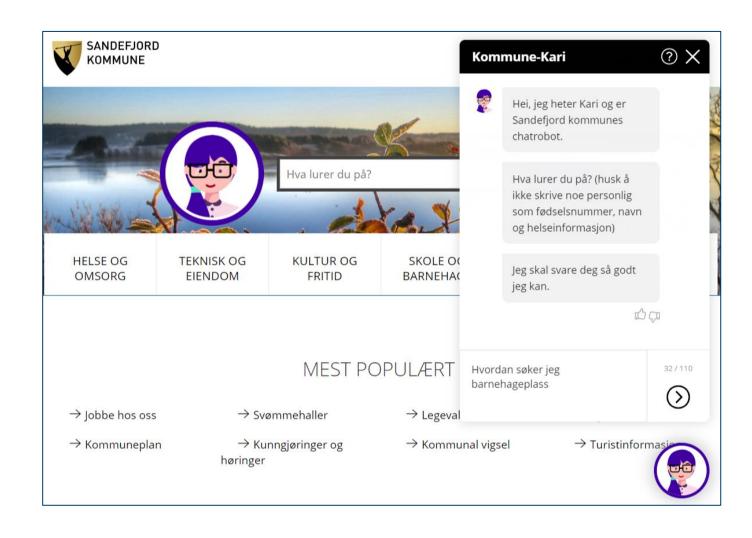


Chatbots area of sharply increasing research interest

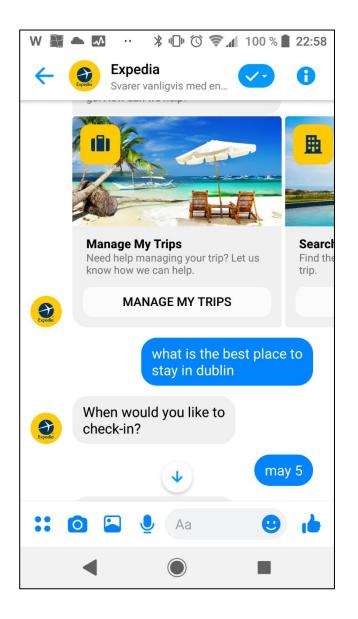
Search term: chatbot OR "conversational agent"



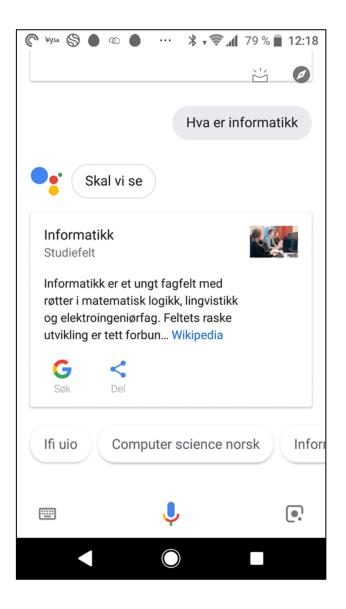
- Customer service
- Content curation and eCommerce
- Assistants
- Guides
- Social relations



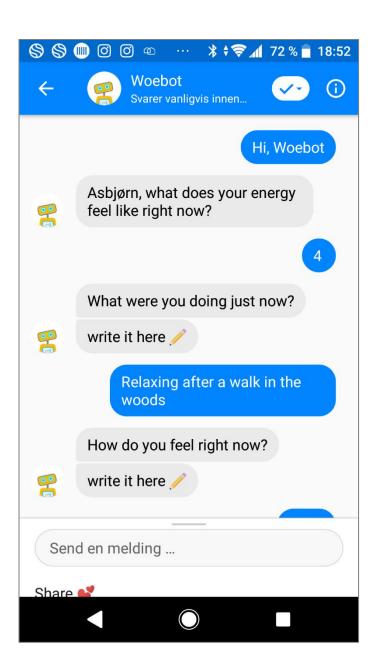
- Customer service
- Content curation and eCommerce
- Assistants
- Guides
- Social relations



- Customer service
- Content curation and eCommerce
- Assistants
- Guides
- Social relations



- Customer service
- Content curation and eCommerce
- Assistants
- Guides
- Social relations



- Customer service
- Content curation and eCommerce
- Assistants
- Guides
- Social relations



Chatbots always there for you

Always supportive and caring

... but not without its issues



Chatbots always there for you

Always supportive and caring

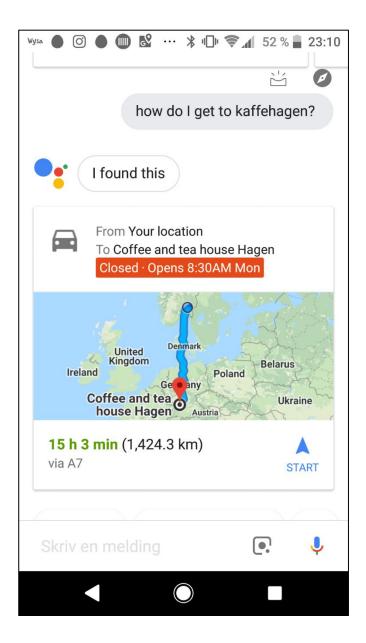
... but not without its issues



Chatbots always there for you

Always supportive and caring

... but not without its issues



Chatbots always there for you

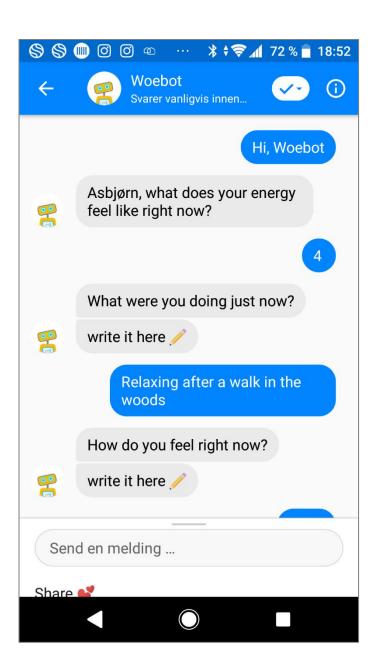
Always supportive and caring

... but not without its issues



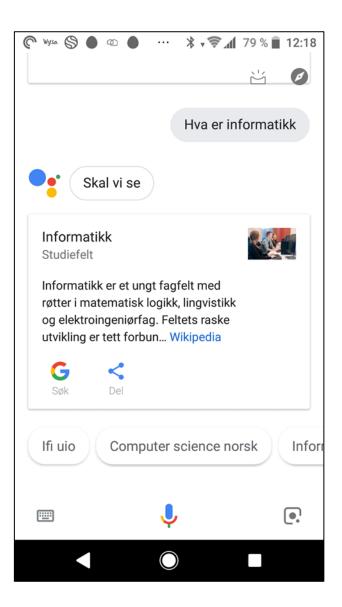
Different types of chatbots

- Scripted
- Intents and actions
- Generative



Different types of chatbots

- Scripted
- Intents and actions
- Generative



Different types of chatbots

- Scripted
- Intents and actions
- Generative

A Neural Conversational Model

text (e.g., hyperlinks) from the dataset. As turn taking is not clearly indicated, we treated consecutive sentences assuming they were uttered by different characters. We trained our model to predict the next sentence given the previous one, and we did this for every sentence (noting that this doubles our dataset size, as each sentence is used both for context and as target). Our training and validation split has 62M sentences (923M tokens) as training examples, and the validation set has 26M sentences (395M tokens). The split is done in such a way that each sentence in a pair of sentences either appear together in the training set or test set but not both. Unlike the previous dataset, the OpenSubtitles is quite large, and rather noisy because consecutive sentences may be uttered by the same character. Given the scope of movies, this is an open-domain conversation dataset, contrasting with the technical troubleshooting dataset.

5. Experiments

In this section, we describe the experimental results with the two datasets. We present a rather objective measure which is the perplexity of the model on the test set. We also show some samples of the interactions with the system that we trained.

5.1. IT Helpdesk Troubleshooting experiments

Given the lack of data for this domain specific and clean dataset, we trained a single layer LSTM with 1024 memory cells using stochastic gradient descent with gradient clipping. We built a vocabulary of the most common 20K words, and introduced special tokens indicating turn taking and actor.

At convergence, this model achieved a perplexity of 8, whereas an n-gram model achieved 18. Below is a sample

Human: i am seeing an error related to vpn Machine: what is the error message that you are getting when connecting to vpn using network

Human: connection refused or something like

Machine: may i know the version of network

connect vou connect?

Human: i am not sure i know that

Machine: is the network connect prompting for

any user certificate Machine: ? Human: yes

Machine: are you at home or at work at the mo-

ment?

Machine: office? Human: i am at home Machine: try this

Machine: goto <URL> page and enter your

name and password Human: done

Machine: try logging in with <NAME> and

check if its working fine now Human: yes, now it works!

Machine: great

Machine: anything else that i can help Human: no, that was very useful Machine: you are other than <NAME>:)

Machine: have a nice night

(The <URL> indeed contains information about vpn access.)

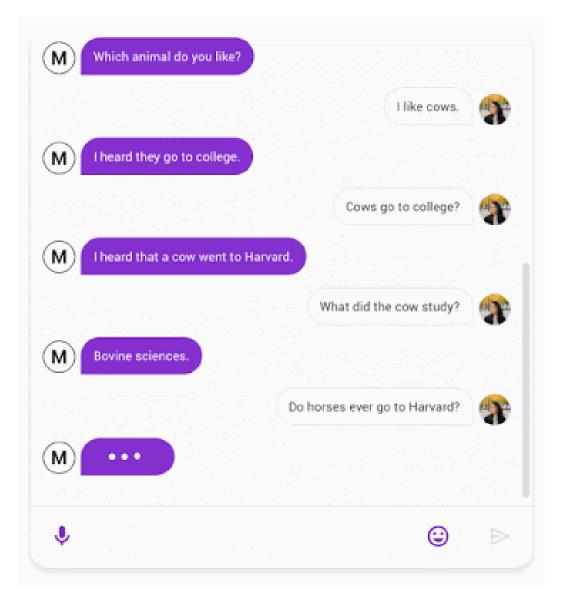
Conversation 2: Browser issues.

A race towards human likeness?

Google Meena – January 2020

Open domain chatbot based on machine learning – single model

Trained on 40B words from open social media conversations

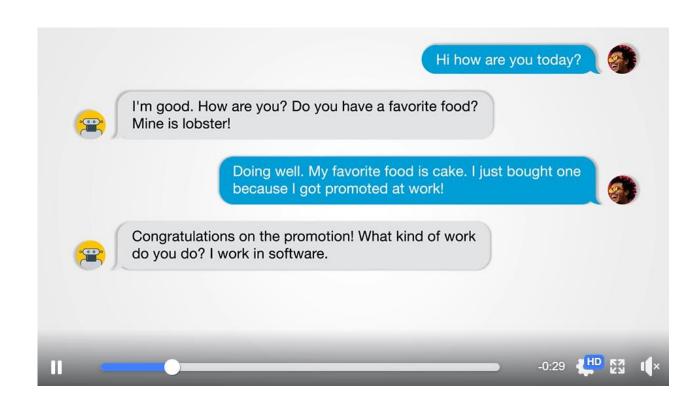


A race towards human likeness?

Facebook Blender – April 2020

Open domain chatbot based on machine learning - blended model

Trained on 1,5B Reddit comments.



A race towards human likeness?

Open AI – GPT-3 – July 2020

Large language model which may be used to power chatbots

Focus on few-shot learning

A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:

We were traveling in Africa and we saw these very cute whatpus.

To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:

One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.

A "yalubalu" is a type of vegetable that looks like a big pumpkin. An example of a sentence that uses the word yalubalu is:

I was on a trip to Africa and I tried this yalubalu vegetable that was grown in a garden there. It was delicious.

A "Burringo" is a car with very fast acceleration. An example of a sentence that uses the word Burringo is:

In our garage we have a Burringo that my father drives to work every day.

A "Gigamuru" is a type of Japanese musical instrument. An example of a sentence that uses the word Gigamuru is:

I have a Gigamuru that my uncle gave me as a gift. I love to play it at home.

To "screeg" something is to swing a sword at it. An example of a sentence that uses the word screeg is:

We screeghed at each other for several minutes and then we went outside and ate ice cream.



Playground (i)







<>





The following is a conversation with an AI assistant. The assistant is helpful, creative, clever, and very friendly.

Human: Hello, who are you?

AI: I am an AI created by OpenAI. How can I help you today? Human: Want to be part of a demo for Interacting with AI?

AI: Sure, I like that it is to be biased toward do's and don't about AI.

Human:





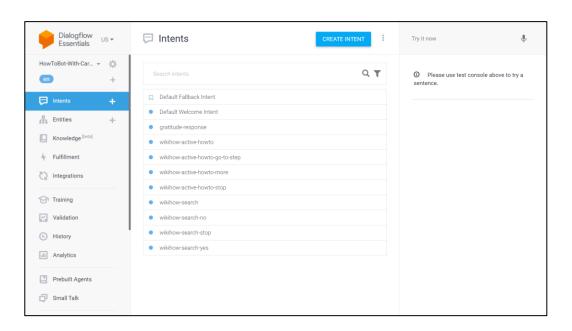


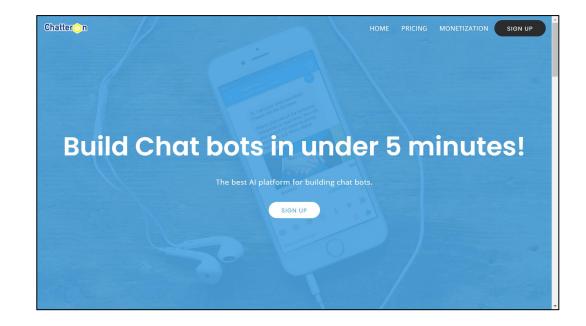




Chatbot development platforms

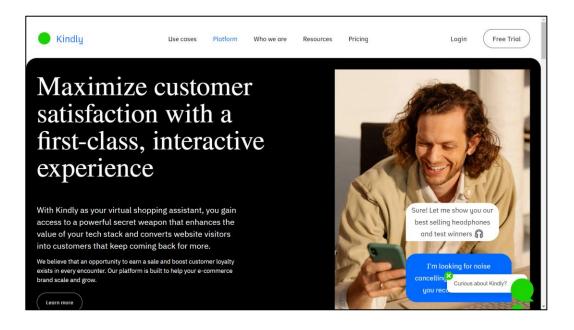
- DIY
- Platform and assistance





Chatbot development platforms

- DIY
- Platform and assistance





Group task – for October 6 (Appendix 1 in mid-term report)

Prototype a chatbot for a self-decided purpose

... using Chatteron

(... or Chatfuel if you want to use one of the most popular platforms for FB Messenger chatbots)

(... or Dialogflow if you want more flexibility and challenge)

Keep it simple and early-phase (4-8 hours work?)

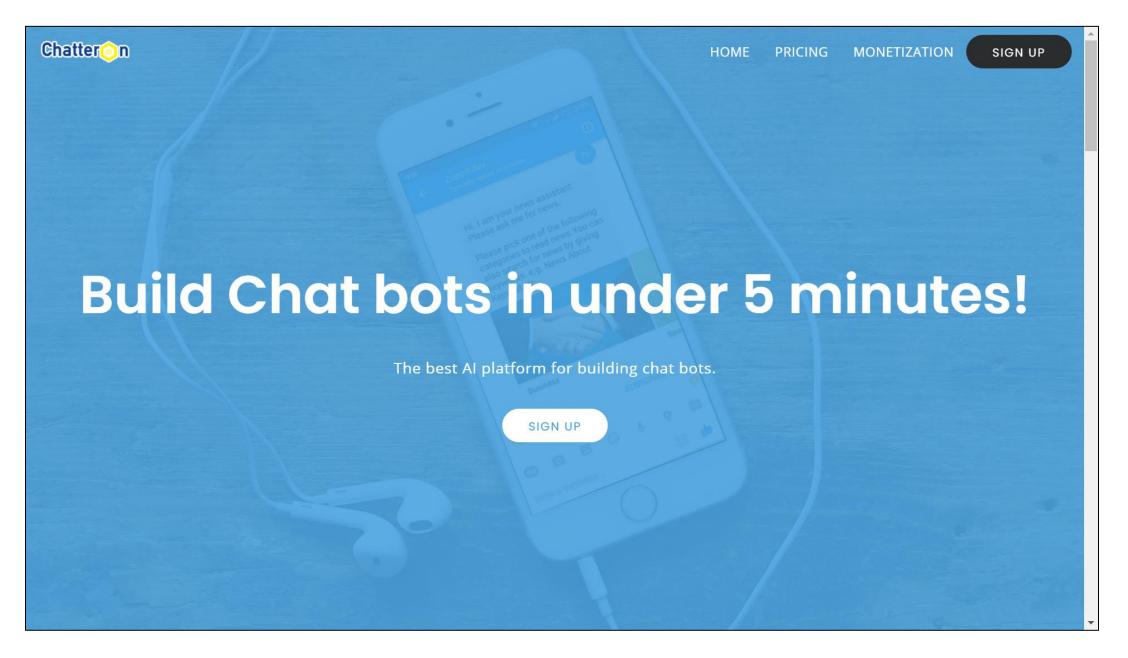
Make brief (approx 1 minute) video capture of dialogue – upload to Vortex

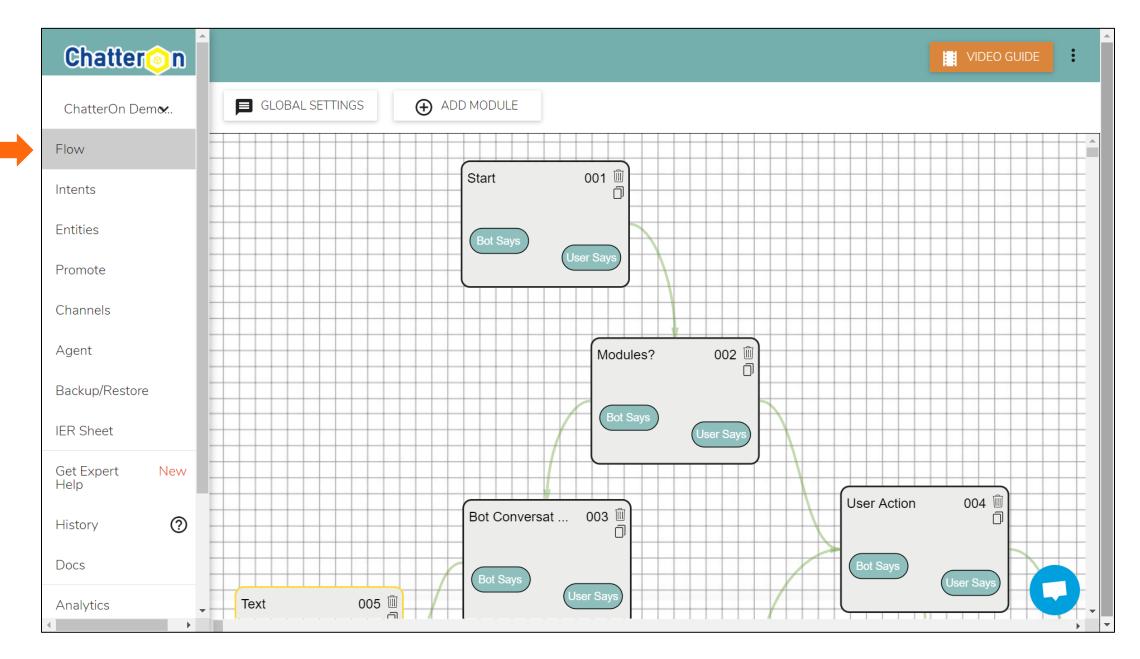
Decide on purpose

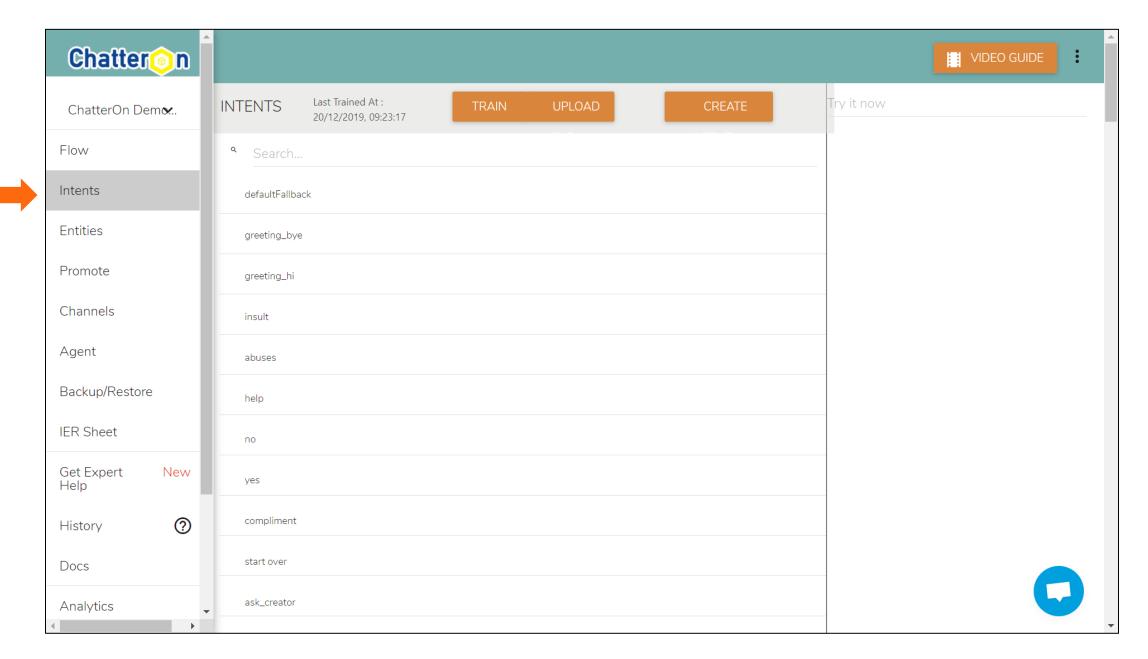
Identify key tasks

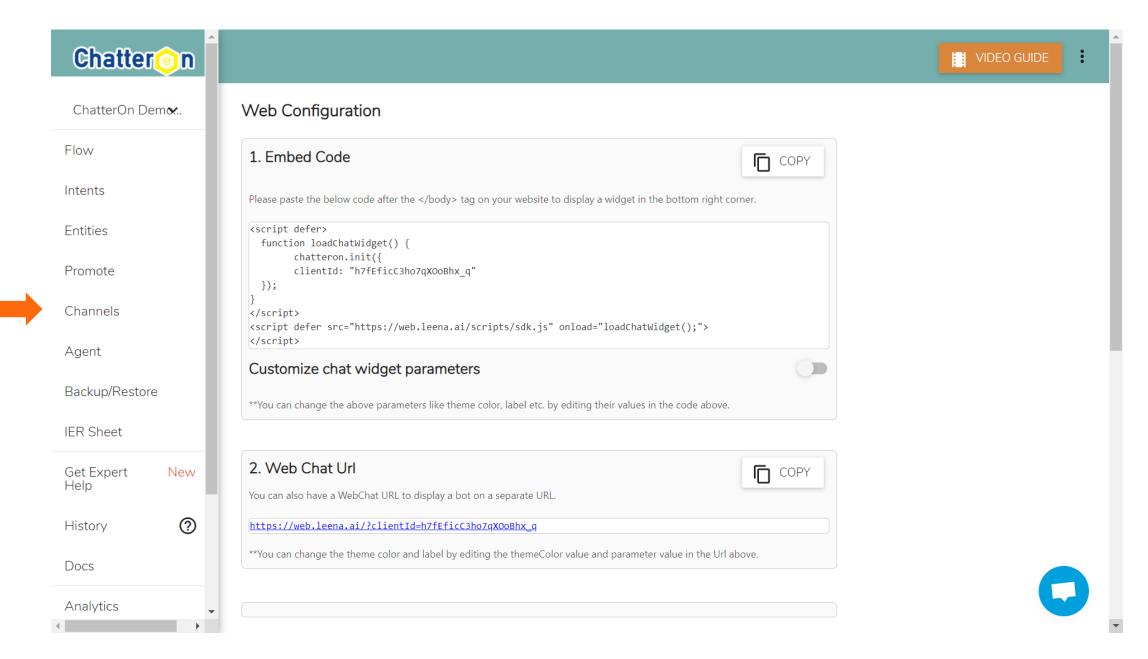
Design flows, intents and actions

Test and refine









End 1