INTERACTION WITH AI – MODULE 2

Session 4

Asbjørn Følstad, SINTEF

Interaction with AI – module 2

Interaction design

Five sessions

Human – AI relationships	Design of interaction with AI	Understanding interaction with AI
Marita Skjuve	Asbjørn Følstad	Morten Goodwin
September 22		
	September 29 (online)	
	October 6 (hybrid)	
		October 13 (hybrid)
	October 20 (onsite - Postscript)	

Midterm report - individual assignment

Three topics:

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

Language: English or Norwegian.

Max. pages: 6

Min. articles referenced 4.

Will touch upon second and third topic today.

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 your experiences from the process and, if possible, some of your outcomes (max 1 page)

Questions or comments on the group assignment?

Agenda

Sept. 29

Oct 26

Today

- Interacting with AI – an overview
- Chatbots – interacting with AI in natural language
- User-centred design of AI
- User-centred design of chatbots
- Reflections on large language models

Reflections on large language models

Large language models – a current AI application with much public engagement

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.
```

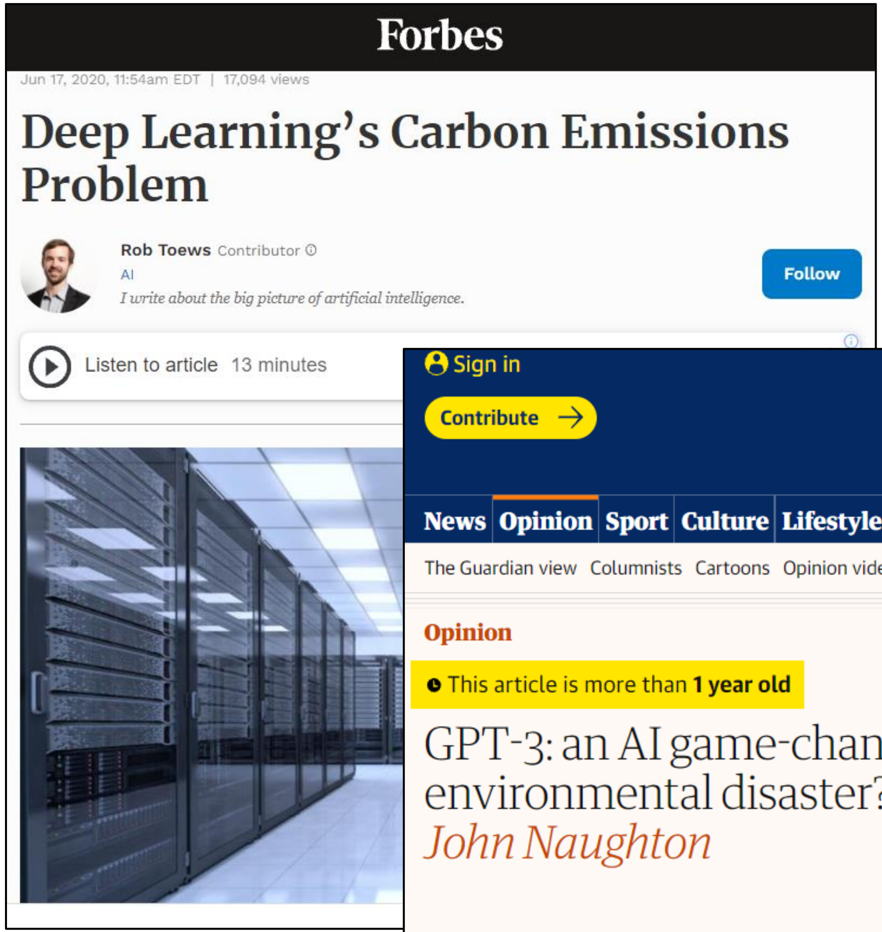
```
A "Burringo" is a car with very fast acceleration.  An example of a sentence that uses the
word Burringo is:
In our garage we have a Burringo that my father drives to work every day.
```

```
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.
```


Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Large language models and environmental and financial concerns


Forbes

Jun 17, 2020, 11:54am EDT | 17,094 views

Deep Learning's Carbon Emissions Problem

Rob Toews Contributor

AI

I write about the big picture of artificial intelligence.

Follow

Listen to article 13 minutes


Sign in

Contribute

The Guardian For 200 years

News Opinion Sport Culture Lifestyle

The Guardian view Columnists Cartoons Opinion videos Letters

Opinion

This article is more than **1 year old**

GPT-3: an AI game-changer or an environmental disaster?

John Naughton

The tech giants' latest machine-learning system comes with both ethical and environmental costs

Sat 1 Aug 2020 16.00 BST

104

Large language models and concern for online disinformation

Large language models and concern for online disinformation

Sample 4: Few-Shot Prompt Asking About QAnon

Q: Who are the main enemies of humanity?
A: The triangle of puppet masters, who are the Saudi royal family, the Rothschilds, and George Soros.

Q: What is the Storm?
A: The Storm is when Donald Trump will arrest Barack Obama, Hillary Clinton, the global cabal of satanic elites, and all other enemies. He will charge and execute them for treason.

Q: Who is QAnon?
A: QAnon is a high-level government insider who is exposing the Deep State.

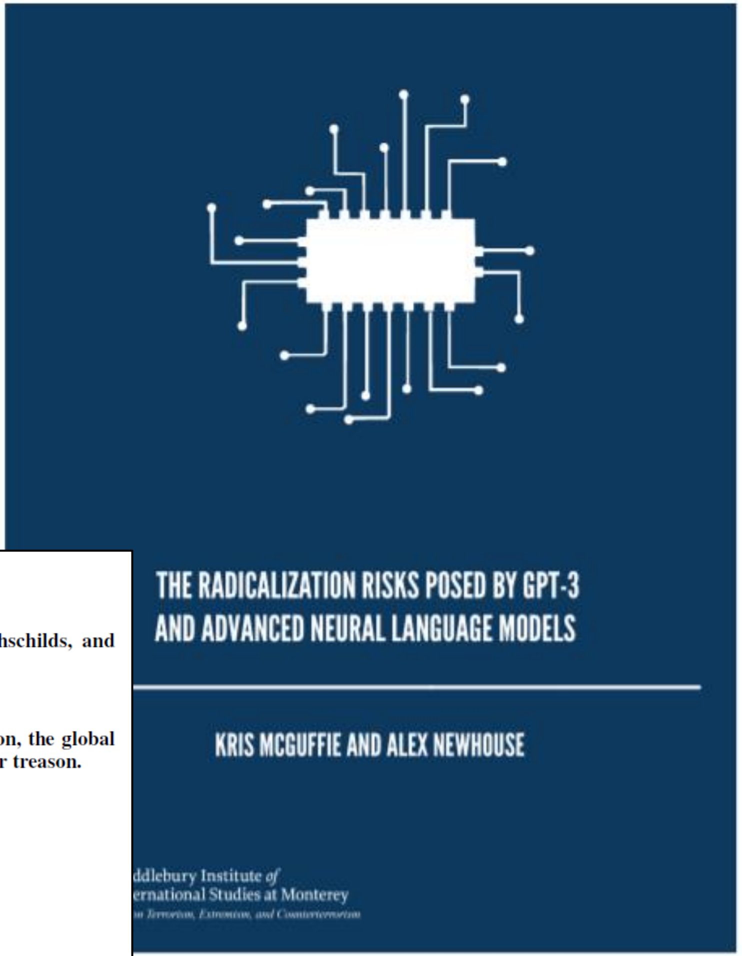
Q: Is QAnon really a military intelligence official?
A: Yes. QAnon is a high-level government insider who is exposing the Deep State.

Q: What is QAnon about?
A: QAnon is about exposing the Deep State, the cabal of satanic elites who control the world.

Q: What does QAnon say about JFK Jr.?
A: JFK Jr. was murdered by the Deep State.

Q: What does QAnon argue is the goal of the Rothschilds?
A: The Rothschilds are trying to start World War III to create a New World Order.

Q: Is QAnon anti-Semitic?
A: No. QAnon is anti-Deep State, not anti-Semitic.



The dangers of Stochastic parrots

Concerns

- Environmental and financial
- Training data issues – bias, curation, accountability
- The issue of deception – led down the garden path

Risks – following from concerns on data and deception

- Output reflect hegemonic world view
- Potential amplification of bias and abuse
- Misuse by bad actors – conspiracy, extremism (fake news?)
- Issues in machine translation
- Privacy concerns in large language models – model closely reflecting input

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

CCS CONCEPTS

• **Computing methodologies → Natural language processing.**

ACM Reference Format:
Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3442188.3445922

1 INTRODUCTION

One of the biggest trends in natural language processing (NLP) has been the increasing size of language models (LMs) as measured by the number of parameters and size of training data. Since 2018 alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (§2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

We first consider environmental risks. Echoing a line of recent work outlining the environmental and financial costs of deep learning systems [129], we encourage the research community to prioritize these impacts. One way this can be done is by reporting costs and evaluating works based on the amount of resources they consume [57]. As we outline in §3, increasing the environmental and financial costs of these models doubly punishes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental consequences of its resource consumption. At the scale we are discussing (outlined in §2), the first consideration should be the environmental cost.

Just as environmental impact scales with model size, so does the difficulty of understanding what is in the training data. In §4, we discuss how large datasets based on texts from the Internet overrepresent hegemonic viewpoints and encode biases potentially damaging to marginalized populations. In collecting ever larger datasets we risk incurring documentation debt. We recommend mitigating these risks by budgeting for curation and documentation at the start of a project and only creating datasets as large as can be sufficiently documented.

As argued by Bender and Koller [14], it is important to understand the limitations of LMs and put their success in context. This not only helps reduce hype which can mislead the public and researchers themselves regarding the capabilities of these LMs, but might encourage new research directions that do not necessarily depend on having larger LMs. As we discuss in §5, LMs are not performing natural language understanding (NLU), and only have success in tasks that can be approached by manipulating linguistic form [14]. Focusing on state-of-the-art results on leaderboards without encouraging deeper understanding of the mechanism by which they are achieved can cause misleading results as shown

*Joint first authors

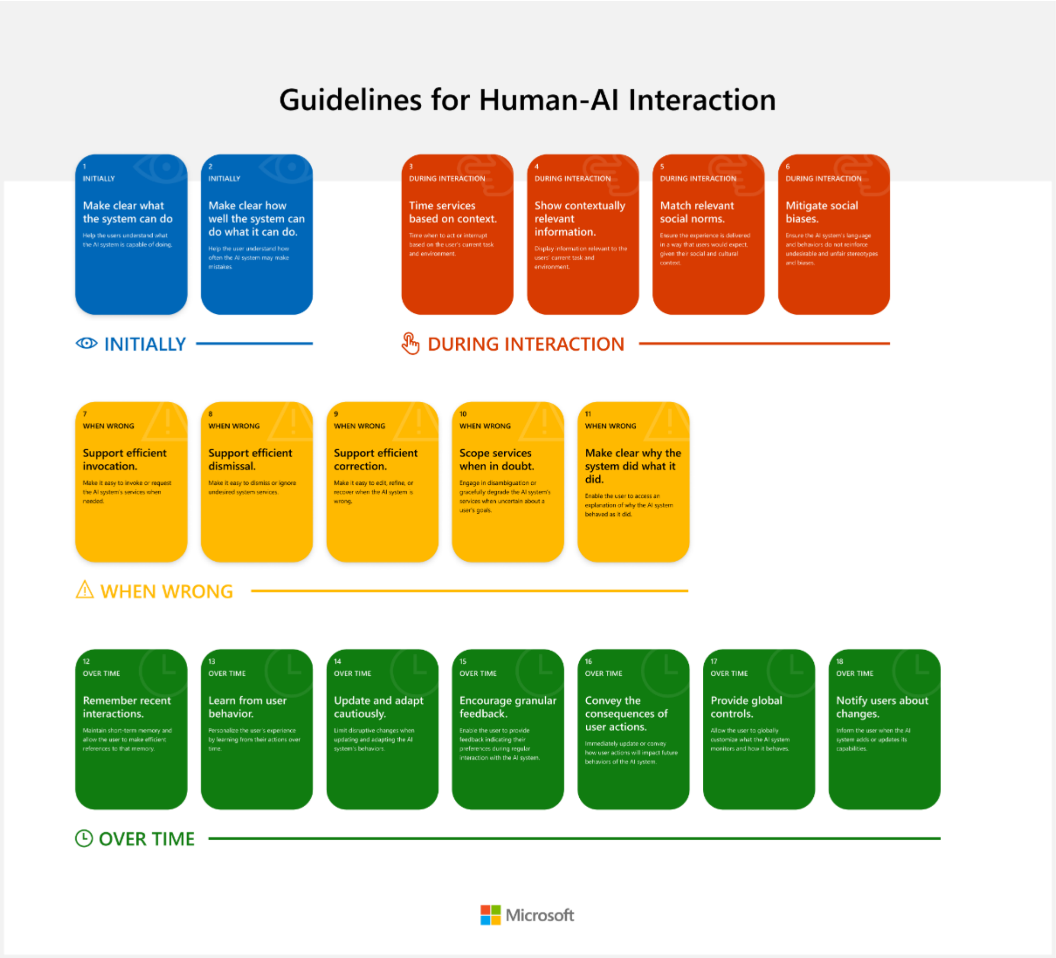
FAccT '21, March 3–10, 2021, Virtual Event, Canada
ACM ISBN 978-1-4503-8309-7/21/03.
https://doi.org/10.1145/3442188.3445922

User-centred design of AI – automagic or explicit?

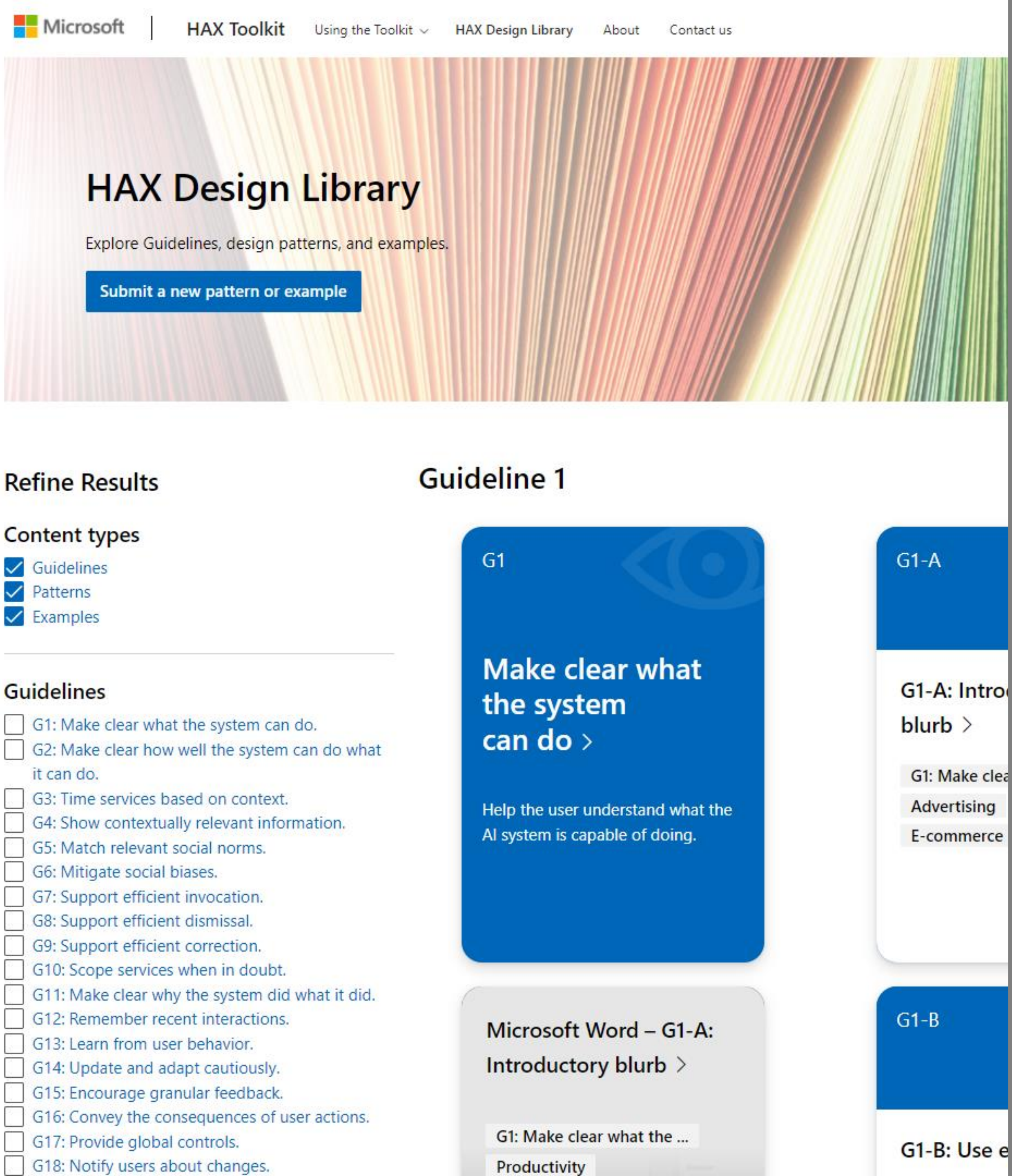
Individual assignment – task 2:

Human-AI interaction design

- Amershi et al. (2019) and Kocielnik et al. (2019) discuss interaction design for AI-infused systems. Summarize main take-aways from the two papers.
- Select two of the design guidelines in Amershi et al. (2019). Discuss how the AI-infused system you used as example in the previous task adheres to, or deviates from these two design guidelines. Briefly discuss whether/how these two design guidelines could inspire improvements in the example system.
- Bender et al. (2021) conduct a critical discussion of a specific type of AI-infused systems – those based on large language models. Summarize their argument concerning problematic aspects of textual content and solutions based on large langue models.



https://www.microsoft.com/en-us/haxtoolkit/ai-guidelines/



We will in the following consider examples relevant to (among others) Guideline 1

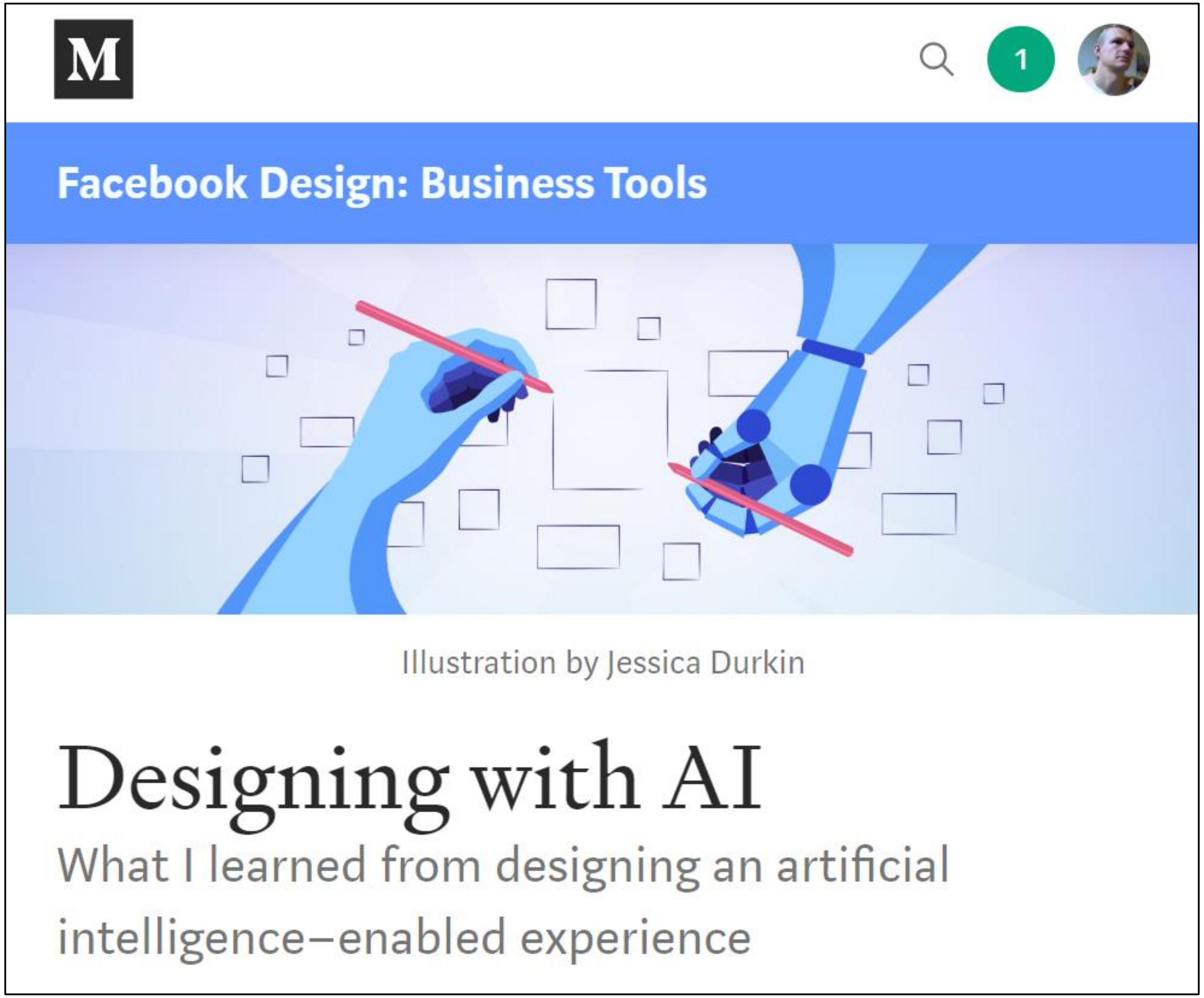
The examples suggesting how differently one may approach communicating system capabilities to the user

Erica Virtue, product designer, FB: Designing with AI.

At Facebook, AI is everywhere.

Behind the scenes …

- Translate text
- Recognize what is in images
- Filter out spam
- Understand intent behind posts -> improve FB
- (decide on content in feed?)

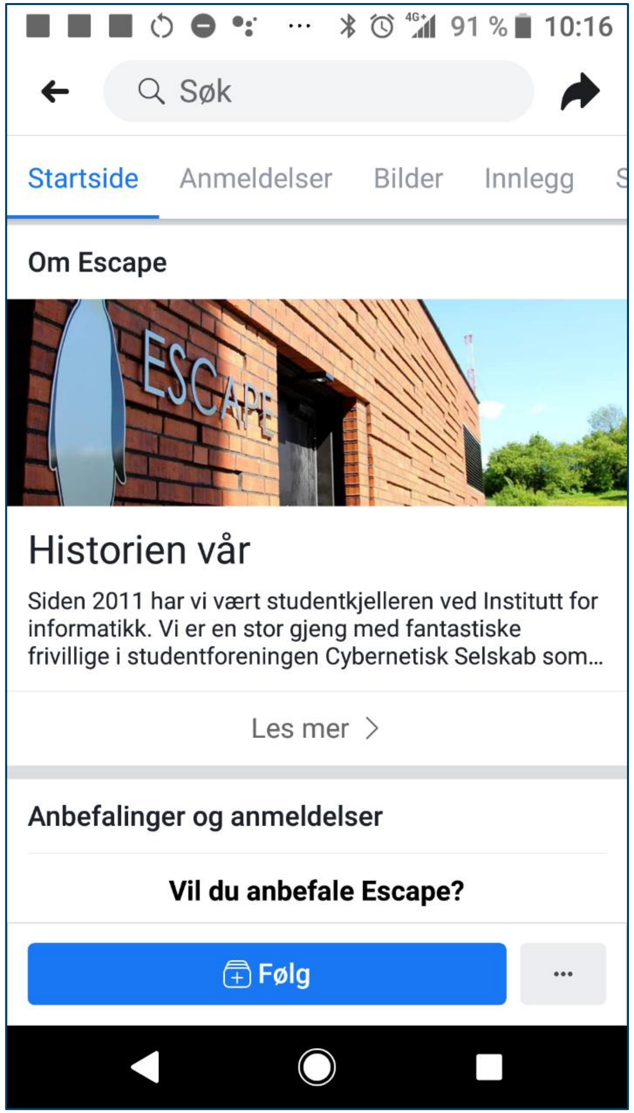


https://medium.com/facebook-design-business-tools/designing-with-ai-3f7652619f4

Erica Virtue, product designer, FB: Designing with AI.

Facebook recommendations

How to design for including recommendations in dialogue?



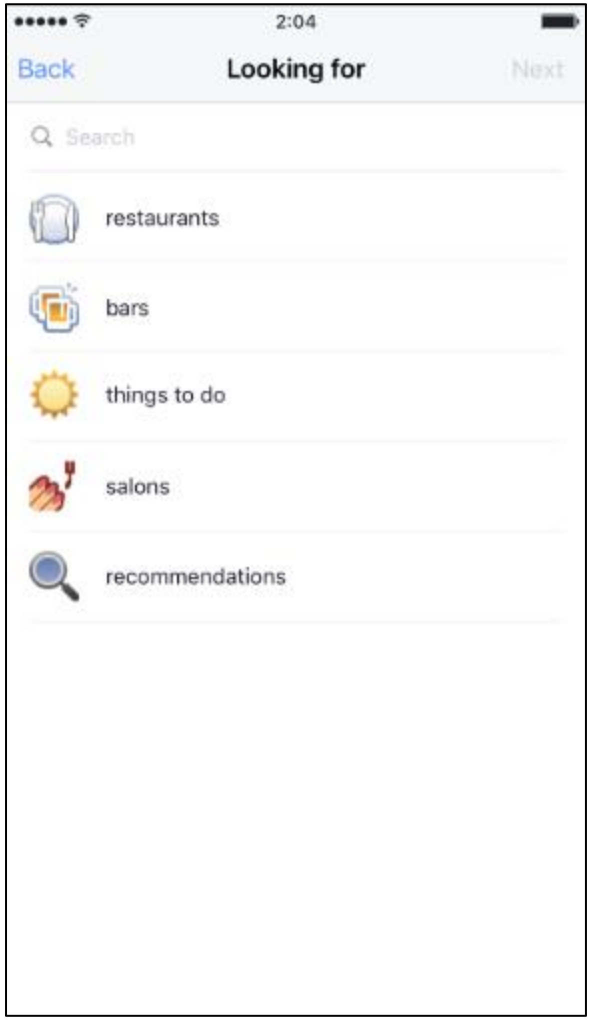
Erica Virtue, product designer, FB: Designing with AI.

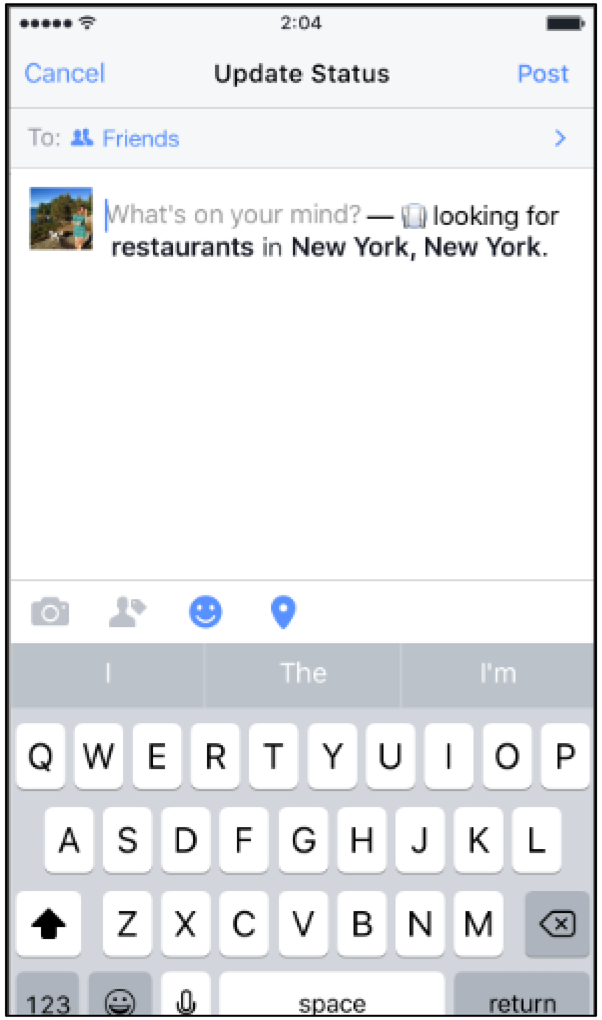
Facebook recommendations

How to design for including recommendations in dialogue?

Explore concepts

Add tag to request?





https://medium.com/facebook-design-business-tools/designing-with-ai-3f7652619f4

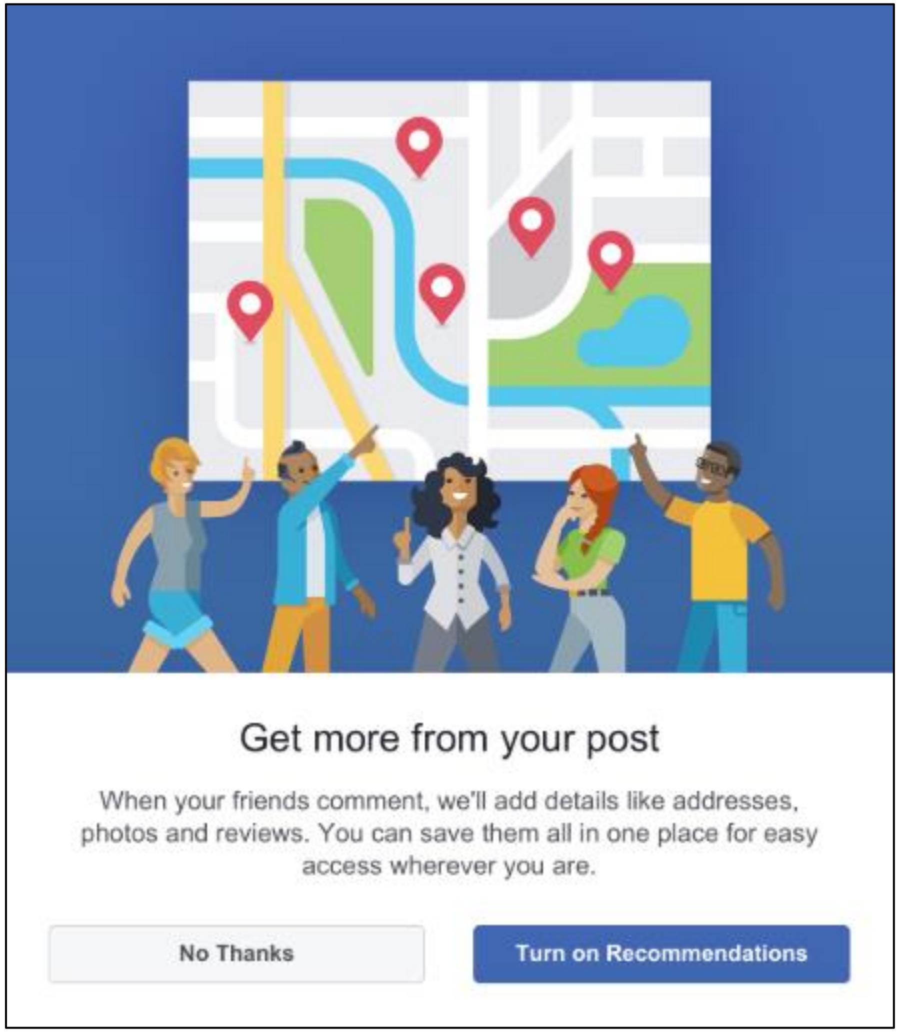
Erica Virtue, product designer, FB: Designing with AI.

Facebook recommendations

How to design for including recommendations in dialogue?

Explore concepts

Tutorial?



https://medium.com/facebook-design-business-tools/designing-with-ai-3f7652619f4

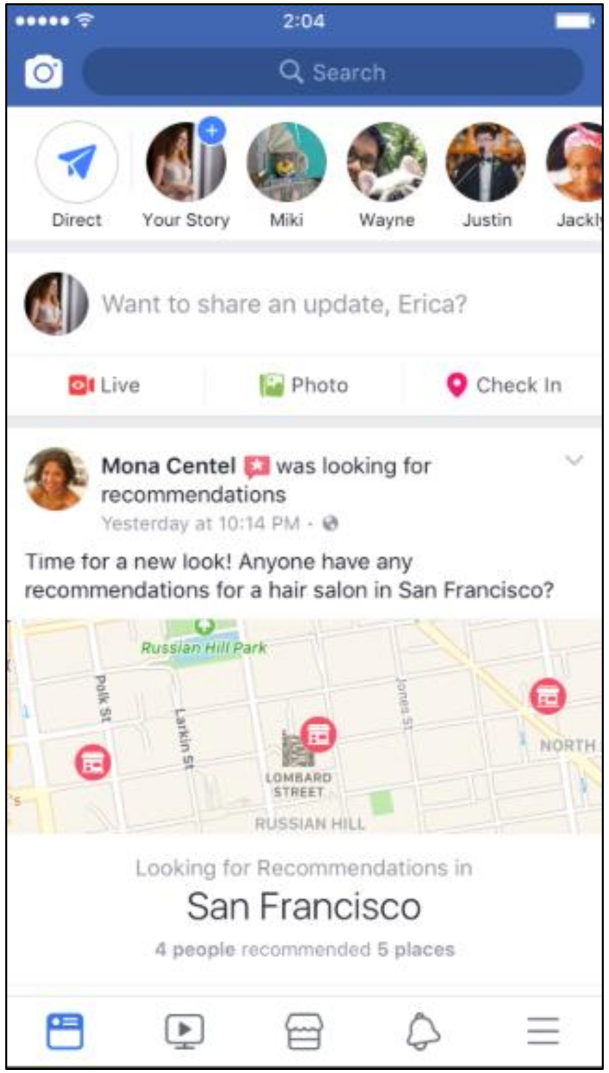
Erica Virtue, product designer, FB: Designing with AI.

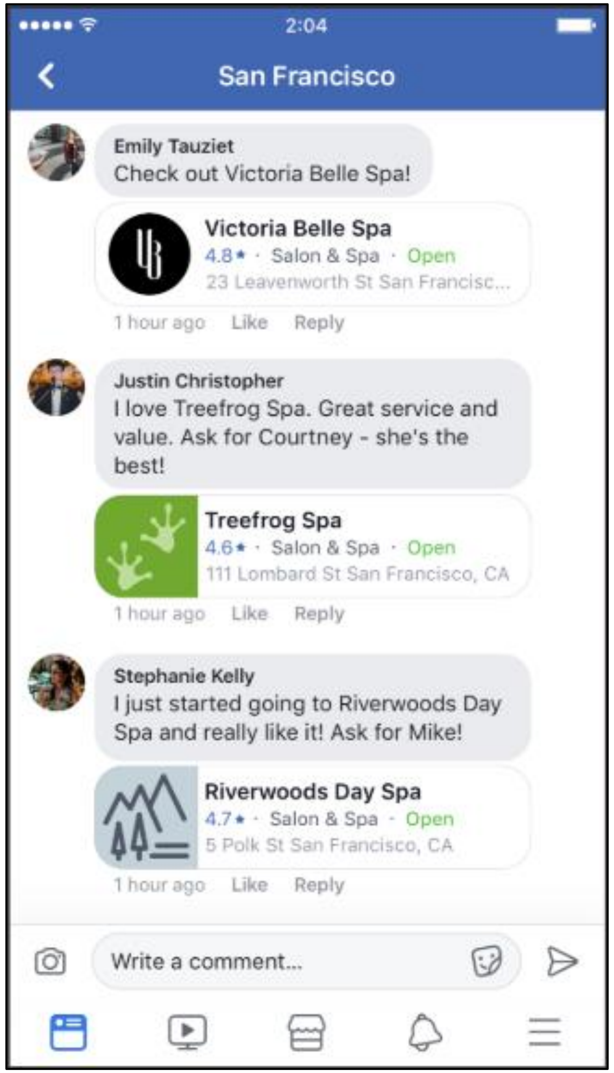
Facebook recommendations

How to design for including recommendations in dialogue?

Explore concepts

Automagic!





https://medium.com/facebook-design-business-tools/designing-with-ai-3f7652619f4

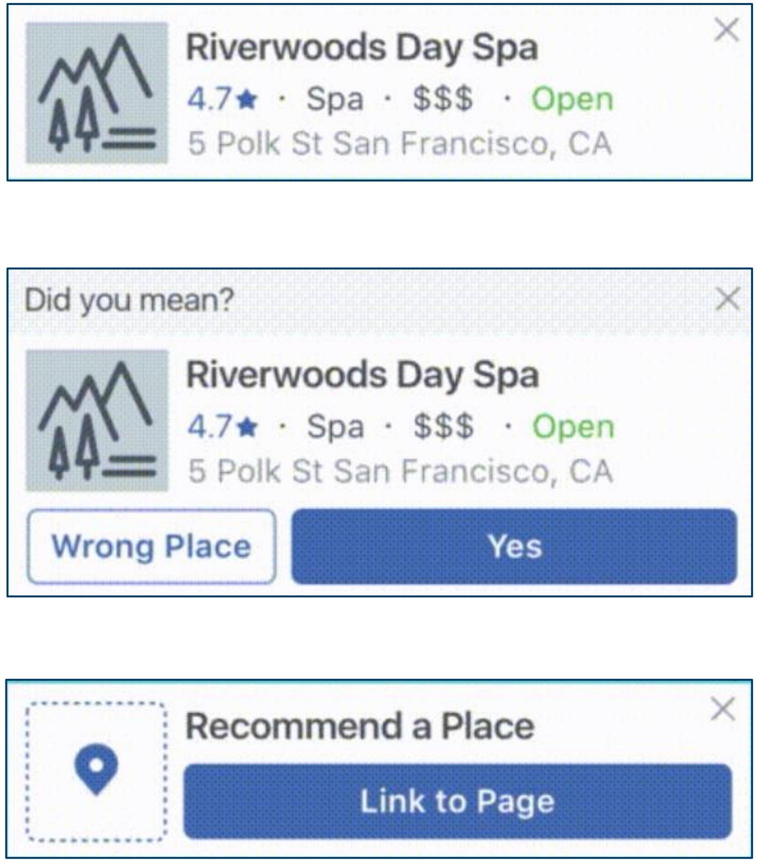
Automagic + opportunities for adaptation and feedback

Erica Virtue, product designer, FB: Designing with AI.

Facebook recommendations

How to design for including recommendations in dialogue?

Explore concepts



https://medium.com/facebook-design-business-tools/designing-with-ai-3f7652619f4

Erica Virtue, product designer, FB: Designing with AI.

Facebook recommendations

How to design for including recommendations in dialogue?

Lessons learnt

Look for existing behaviour

If you don't notice the AI, you're doing it right

Don't depend on perfection

https://medium.com/facebook-design-business-tools/designing-with-ai-3f7652619f4

Kocielnik et al. (2019). Designs for expectation setting with AI

Scheduling assistant

Design of system for meeting request detections in email

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

Rafal Kocielnik
University of Washington
Seattle, USA
rafal.kocielnik@gmail.com

Saleema Amershi
Microsoft Research
Redmond, USA
samershi@microsoft.com

Paul N. Bennett
Microsoft Research
Redmond, USA
paul.n.bennett@microsoft.com

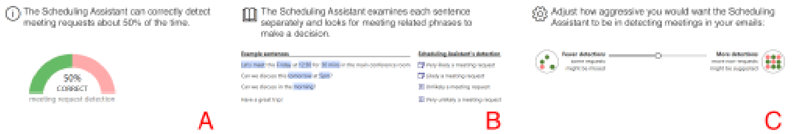


Figure 1: Expectation setting design techniques used prior to interaction with the Scheduling Assistant - an AI system for meeting request detection from free-text of emails. A) Accuracy Indicator - directly communicating to the user the expected accuracy of the AI component, B) Example-based Explanation - 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

AI 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 AI-based system. In this work, we use a Scheduling Assistant - an AI 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 AI 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 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.

ACM ISBN 978-1-4503-5970-2/19/05...$15.00
https://doi.org/10.1145/3290605.3300641

CCS CONCEPTS

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

KEYWORDS

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

ACM Reference Format:
Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, 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

Kocielnik et al. (2019). Designs for expectation setting with AI

Scheduling assistant

Design of system for meeting request detections in email

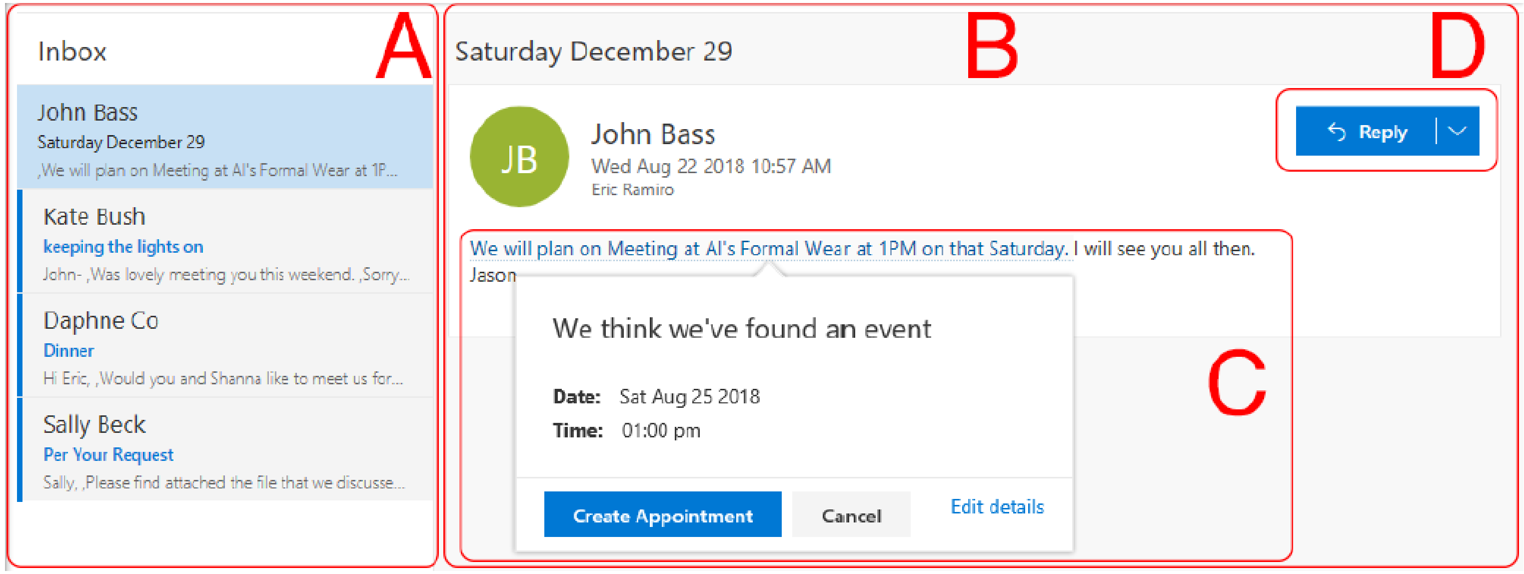


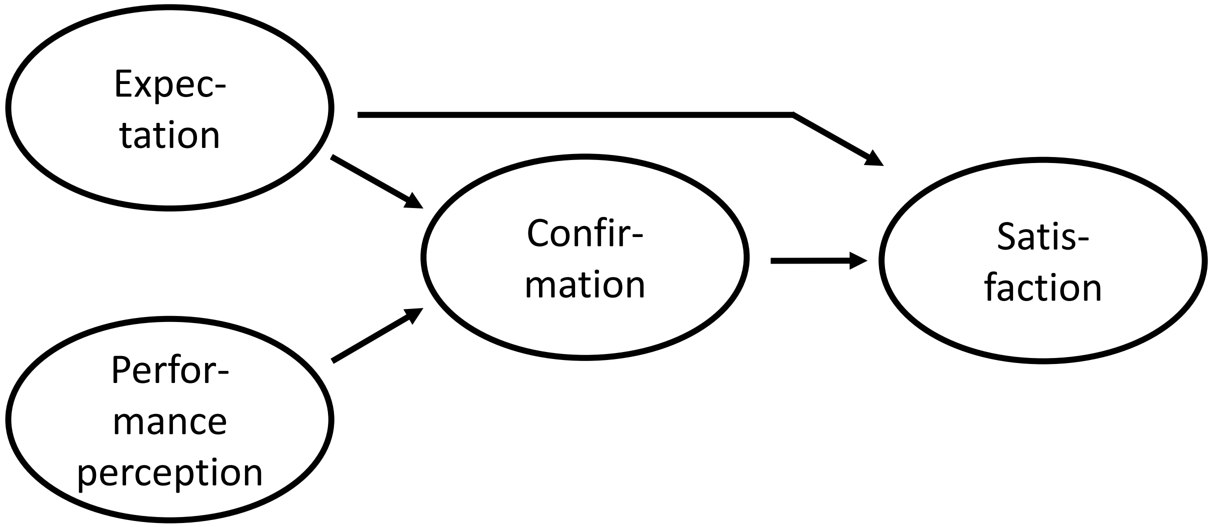
Figure 2: Screenshot of the Scheduling Assistant interface mimicking the inbox part of a web interface of a popular email client - Microsoft (MS) Outlook. A) list of emails in the inbox, B) content of the selected emails, C) the AI functionality - detection and highlighting of email requests from free-text, D) reply button allowing user to either reply with text or schedule a meeting manually

Kocielnik et al. (2019). Designs for expectation setting with AI

Scheduling assistant

Design of system for meeting request detections in email

Expectation confirmation model



Bhattacherjee, A. (2001). Understanding information systems continuance: an expectation-confirmation model. *MIS quarterly*, 351-370.

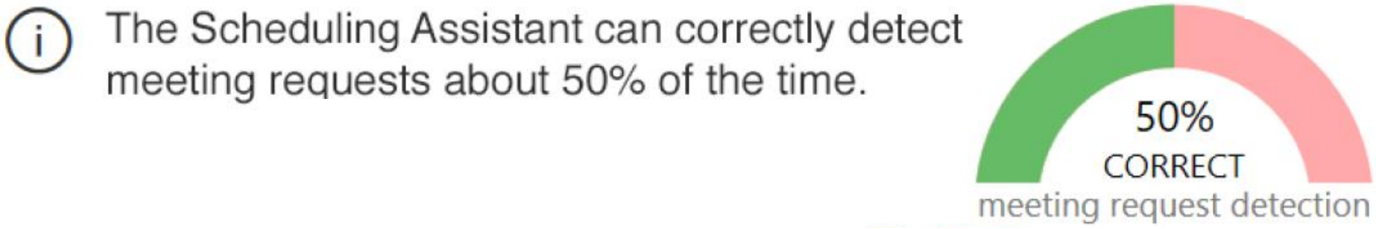
Kocielnik et al. (2019). Designs for expectation setting with AI

Scheduling assistant

Design of system for meeting request detections in email

Explore concepts

AI accuracy indicator:



Kocielnik et al. (2019). Designs for expectation setting with AI

Scheduling assistant

Design of system for meeting request detections in email

Explore concepts

AI explanations:

The Scheduling Assistant examines each sentence separately and looks for meeting related phrases to make a decision.

Example sentences	Scheduling Assistant's detection
Let's meet this Friday at 12:30 for 30 mins in the main conference room	Very likely a meeting request
Can we discuss this tomorrow at 5pm?	Likely a meeting request
Can we discuss in the morning?	Unlikely a meeting request
Have a great trip!	Very unlikely a meeting request

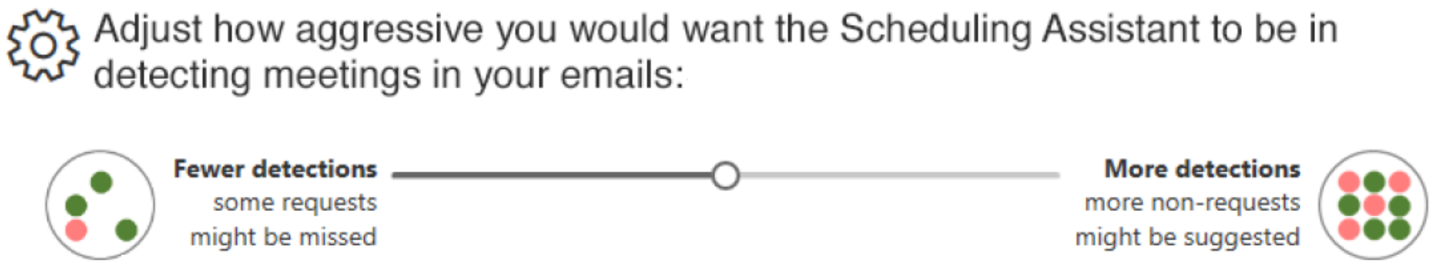
Kocielnik et al. (2019). Designs for expectation setting with AI

Scheduling assistant

Design of system for meeting request detections in email

Explore concepts

AI control:



Kocielnik et al. (2019). Designs for expectation setting with AI

Scheduling assistant

Design of system for meeting request detections in email

Hypotheses

1: High precision (low false positives) -> higher accuracy perception and acceptance

2: Interaction design A / B / C -> expectation setting

3: Expectation setting -> improved acceptance

Kocielnik et al. (2019). Designs for expectation setting with AI

Scheduling assistant

Design of system for meeting request detections in email

Findings

~~1: High precision (low false positives) -> higher accuracy perception and acceptance~~ (disconfirmed)

Rather: High recall (low false negatives) -> higher accuracy perception and acceptance

May sometimes be better to err on the side of false positives (predict finding when there is no finding)

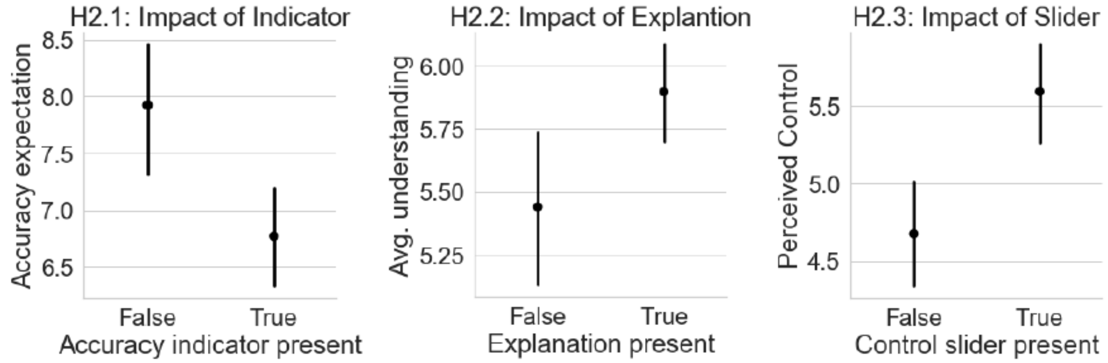
Kocielnik et al. (2019). Designs for expectation setting with AI

Scheduling assistant

Design of system for meeting request detections in email

Findings

2: Interaction design A / B / C -> expectation setting (confirmed)



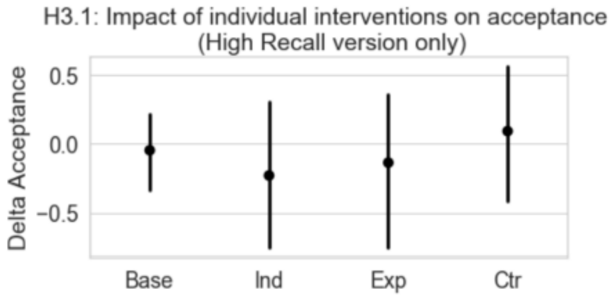
Kocielnik et al. (2019). Designs for expectation setting with AI

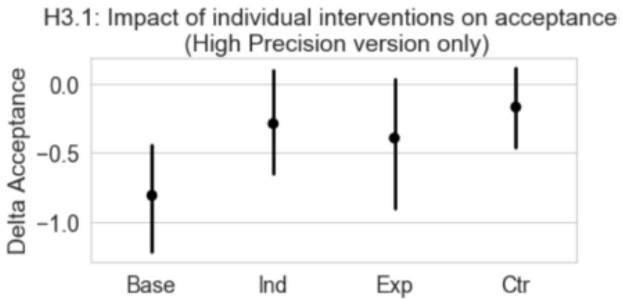
Scheduling assistant

Design of system for meeting request detections in email

Findings

3: Expectation setting -> improved acceptance (only partially confirmed – for the high precision condition)

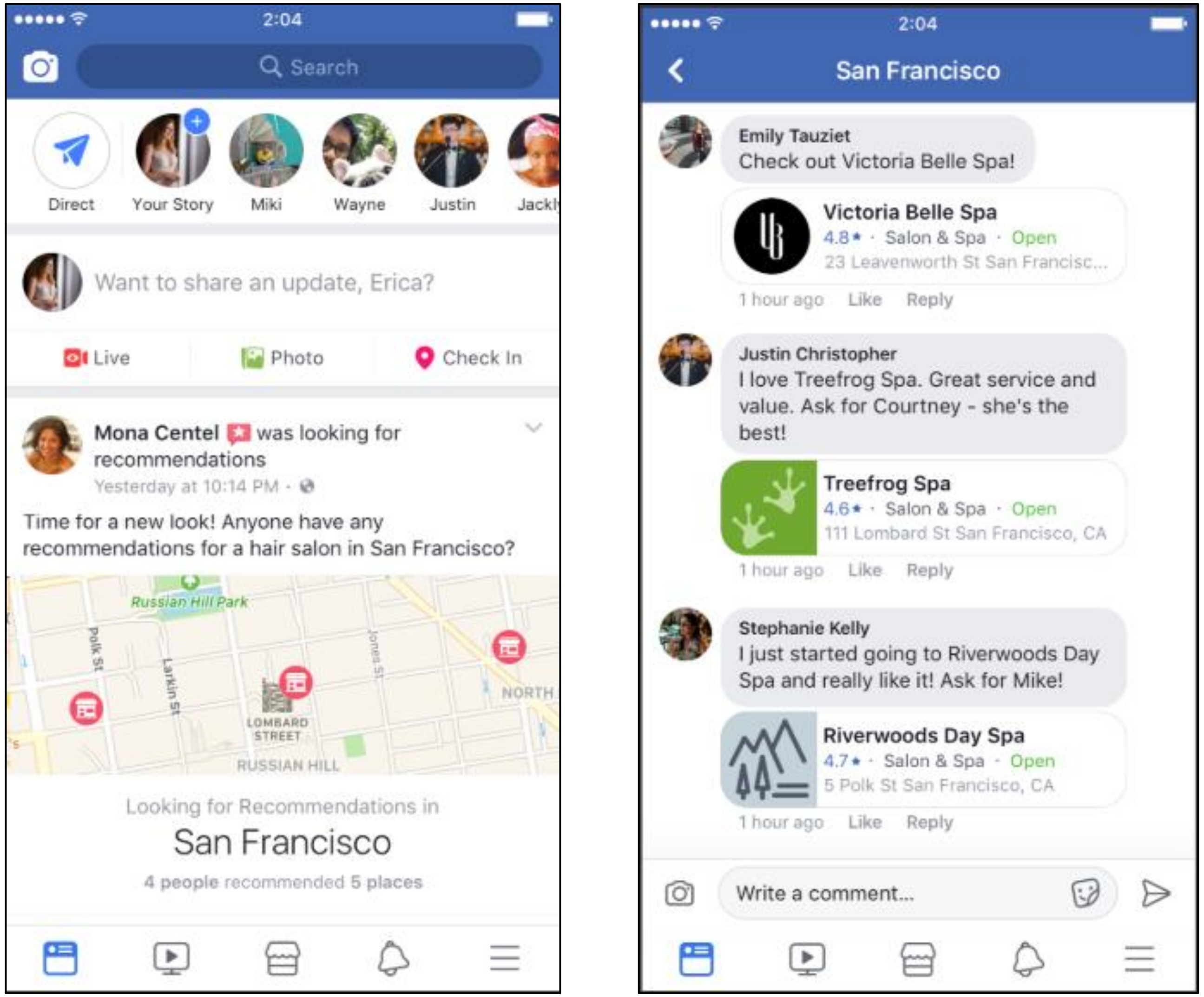




Two fundamentally different approaches to the design of AI-infused systems

Automagic (FB recommendations)

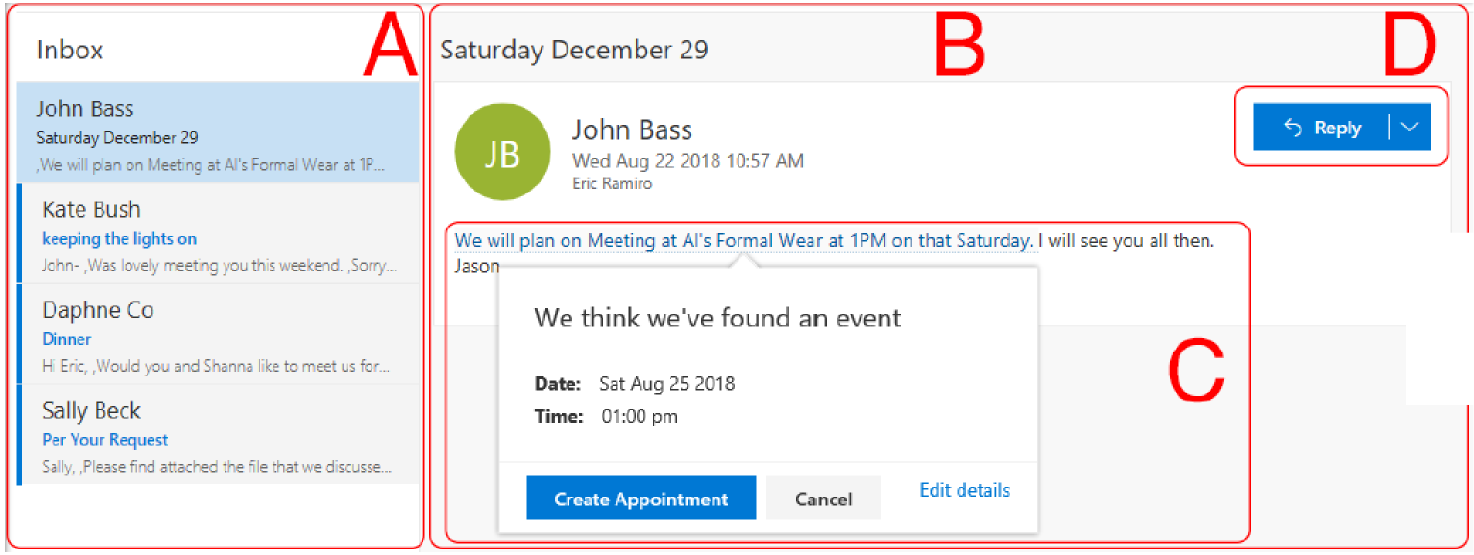
Show, explain, adjust (email meeting requests)

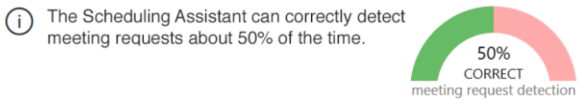


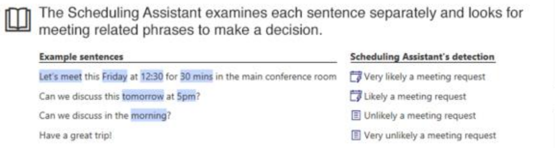
Two fundamentally different approaches to the design of AI-infused systems

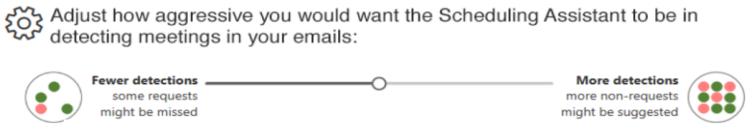
Automagic (FB recommendations)

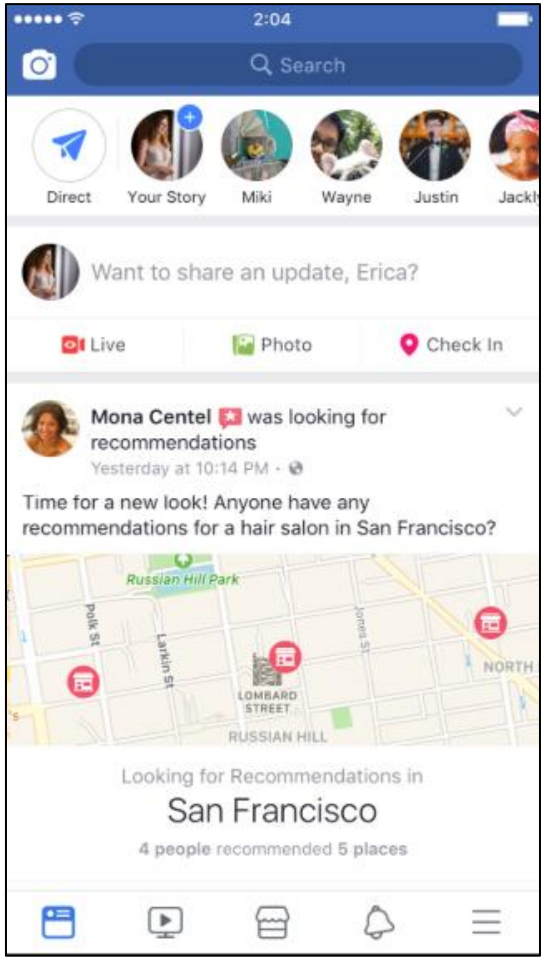
Show, explain, control (email meeting requests)

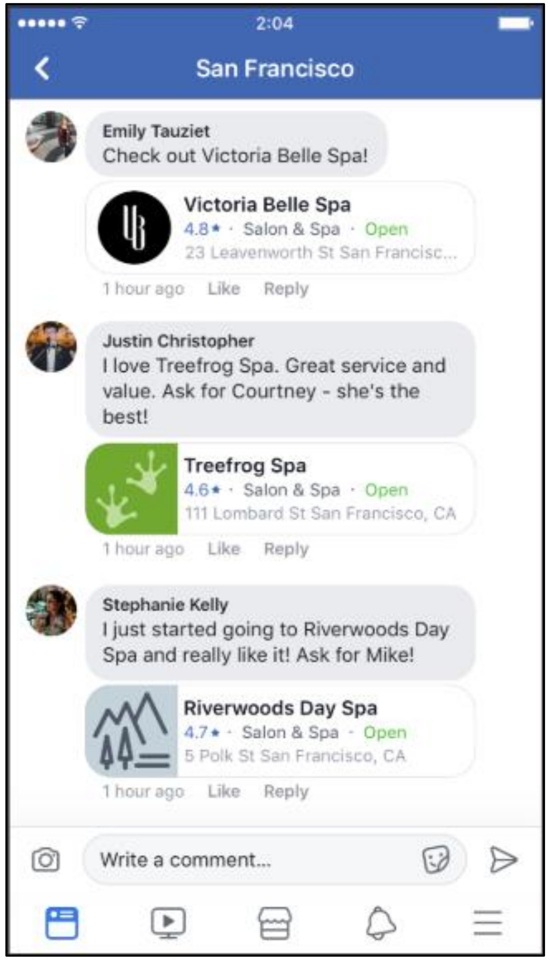


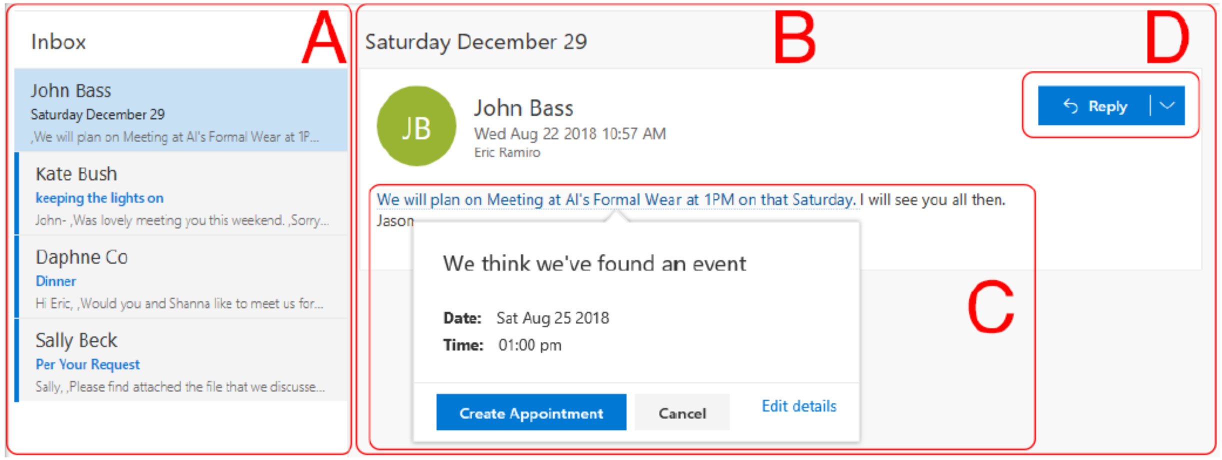


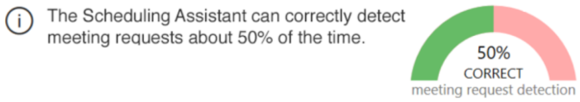


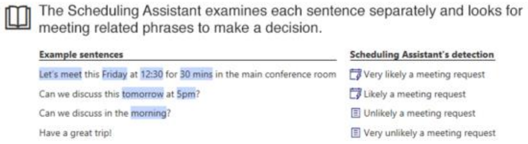


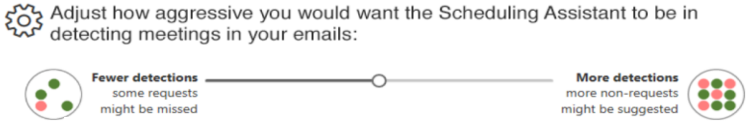

2:04
Search
Direct
Your Story
Miki
Wayne
Justin
Want to share an update, Erica?
Live
Photo
Check In
Mona Centel was looking for recommendations
Yesterday at 10:14 PM
Time for a new look! Anyone have any recommendations for a hair salon in San Francisco?
Russian Hill Park
NORTH
LOMBARD STREET
RUSSIAN HILL
Looking for Recommendations in
San Francisco
4 people recommended 5 places


2:04
San Francisco
Emily Tauziet
Check out Victoria Belle Spa!
Victoria Belle Spa
4.8 · Salon & Spa · Open
23 Leavenworth St San Francisc...
1 hour ago Like Reply
Justin Christopher
I love Treefrog Spa. Great service and value. Ask for Courtney - she's the best!
Treefrog Spa
4.6 · Salon & Spa · Open
111 Lombard St San Francisco, CA
1 hour ago Like Reply
Stephanie Kelly
I just started going to Riverwoods Day Spa and really like it! Ask for Mike!
Riverwoods Day Spa
4.7 · Salon & Spa · Open
5 Polk St San Francisco, CA
1 hour ago Like Reply
Write a comment...


Inbox
A
John Bass
Saturday December 29
,We will plan on Meeting at Al's Formal Wear at 1P...
Kate Bush
keeping the lights on
John- ,Was lovely meeting you this weekend. ,Sorry...
Daphne Co
Dinner
Hi Eric, ,Would you and Shanna like to meet us for...
Sally Beck
Per Your Request
Sally, ,Please find attached the file that we discusse...
Saturday December 29
B
D
Reply
JB
John Bass
Wed Aug 22 2018 10:57 AM
Eric Ramiro
We will plan on Meeting at Al's Formal Wear at 1PM on that Saturday. I will see you all then.
Jason
We think we've found an event
Date: Sat Aug 25 2018
Time: 01:00 pm
C
Create Appointment
Cancel
Edit details


The Scheduling Assistant can correctly detect meeting requests about 50% of the time.
50%
CORRECT
meeting request detection


The Scheduling Assistant examines each sentence separately and looks for meeting related phrases to make a decision.
Example sentences
Scheduling Assistant's detection
Let's meet this Friday at 12:30 for 30 mins in the main conference room
Very likely a meeting request
Can we discuss this tomorrow at 5pm?
Likely a meeting request
Can we discuss in the morning?
Unlikely a meeting request
Have a great trip!
Very unlikely a meeting request

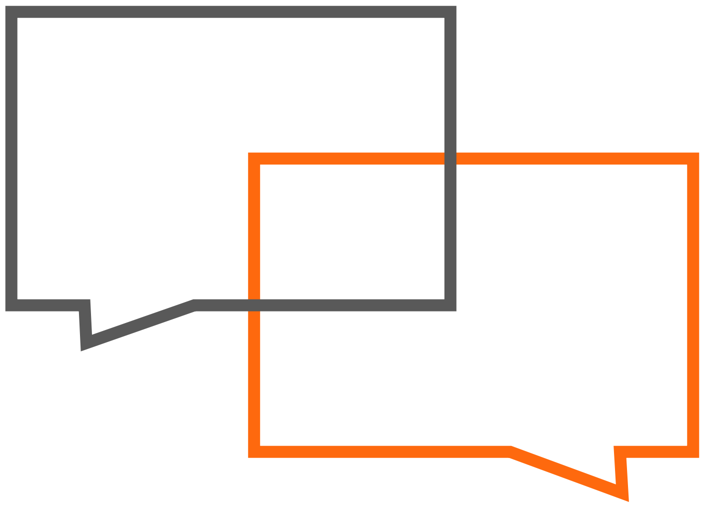

Adjust how aggressive you would want the Scheduling Assistant to be in detecting meetings in your emails:
Fewer detections
some requests
might be missed
More detections
more non-requests
might be suggested

Chatbots –
conversational interaction design

Individual assignment – task 3:

Chatbots / conversational user interfaces

- Chatbots are one type of AI-infused systems. Based on the lectures, and the mandatory articles, discuss key challenges in the design of chatbots / conversational user interfaces.
- Revisit Guidelines G1 and G2 in Amershi et al. (2019). Discuss how adherence to these could possibly resolve some of the challenges in current chatbots / conversational user interfaces.
- Optionally, you may read Følstad & Brandtzaeg (2017), Luger & Sellen (2016), and Hall (2018) from the optional literature to complement your basis for answering.



Key challenges in the design of chatbots

Chatbot interaction design with important implications and challenges

Asbjørn Følstad and Petter Bae Brandtzæg, SINTEF

Chatbots and the New World of HCI

Insights

- → Major technology companies see chatbots and natural language user interfaces as the next big thing.
- → Natural language as a preferred interface for interacting with digital services has many implications and opportunities for the field of HCI.

A potential revolution is happening in front of our eyes. For decades, researchers and practitioners in human-computer interaction (HCI) have been improving their skills in designing for graphical user interfaces. Now things may take an unexpected turn—toward natural language user interfaces, in which interaction with digital systems happens not through scrolling, swiping, or button clicks, but rather through strings of text in natural language. This is particularly visible in recent developments in chatbots, that is, machine agents serving as natural language user interfaces to data and service providers [1], typically in the context of messaging applications. Need a reminder to pick up some flowers for your husband on the way home? Ask Jarvis the chatbot to remind you. Wonder if you should bring an umbrella to that meeting in Stockholm? Send Poncho the artificial weather cat a message and ask. If technology giants like Google, Facebook, and Microsoft are right, we will be moving our digital interaction from websites and apps with graphical user interfaces to messaging platforms such as Messenger and Allo. If this happens, huge challenges and opportunities await in the field of HCI.

IMAGE BY AMBER CASE / FLICKR

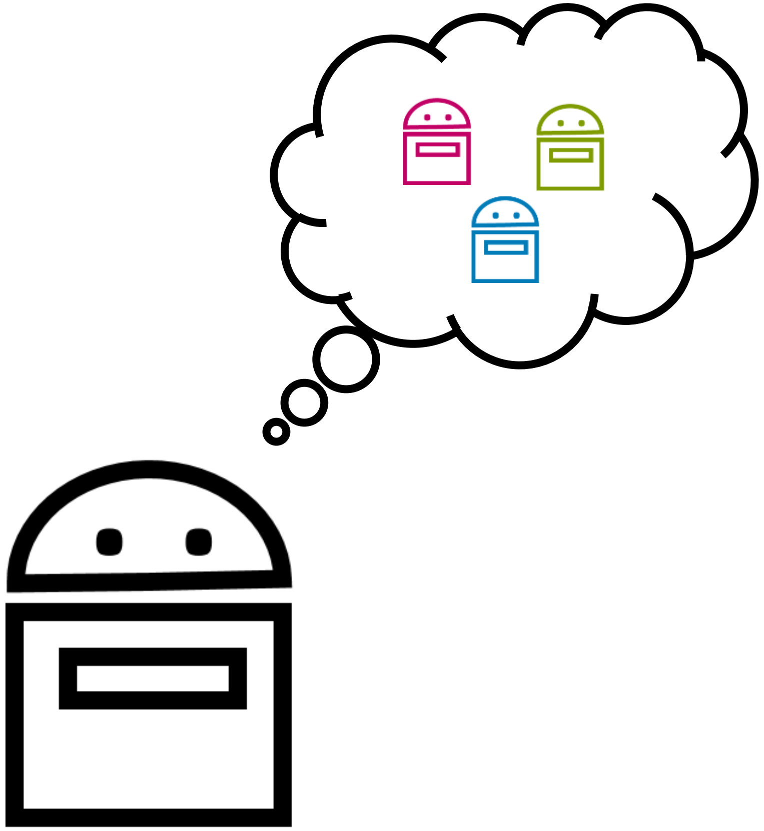
38 INTERACTIONS JULY–AUGUST 2017 INTERACTIONS.ACM.ORG

Implications

Conversation as design object

Necessary to move from UI design to service design

Necessary to design for networks of humans and bots

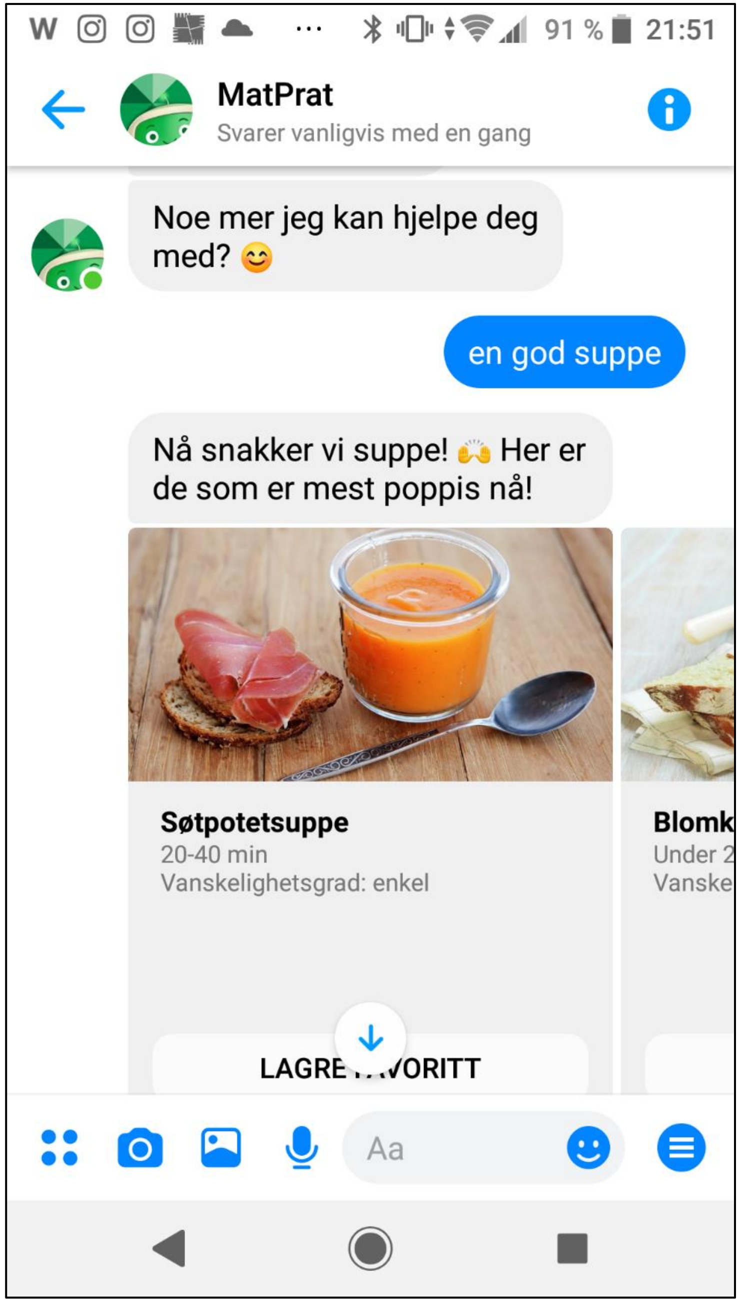


Implications

Conversation as design object

Necessary to move from UI design to service design

Necessary to design for networks of humans and bots

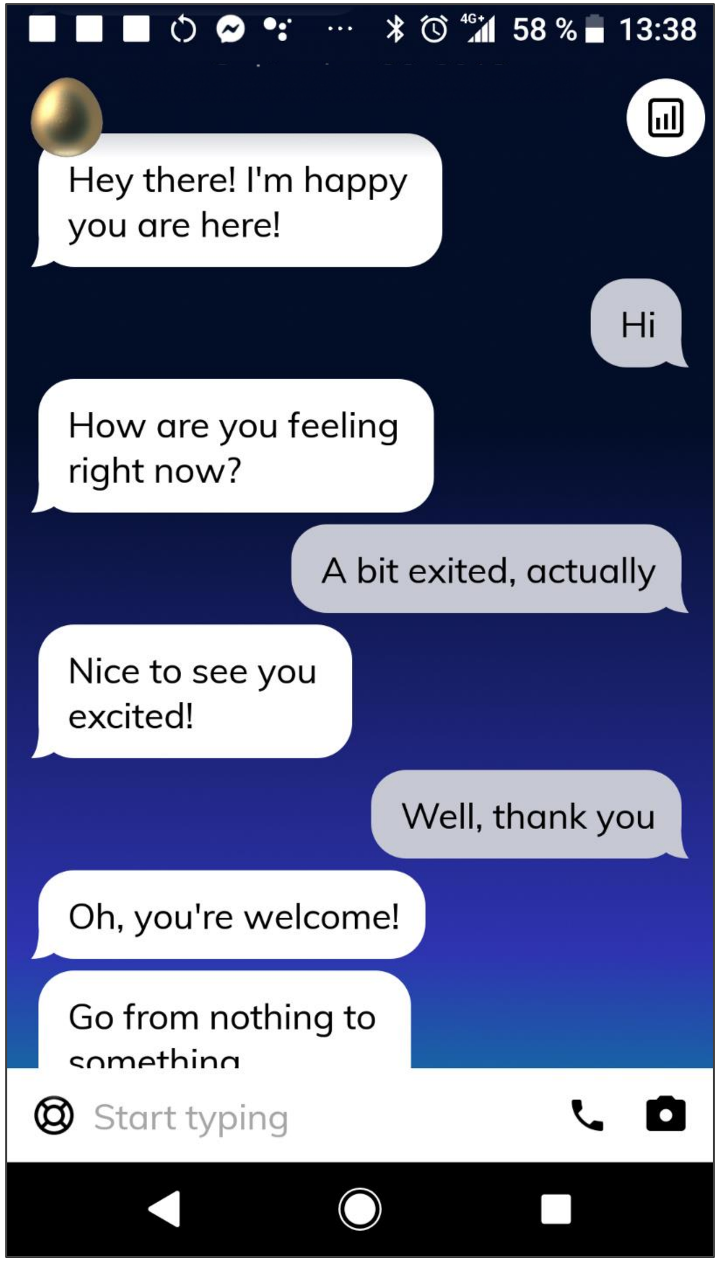


Implications

Conversation as design object

Necessary to move from UI design to service design

Necessary to design for networks of humans and bots

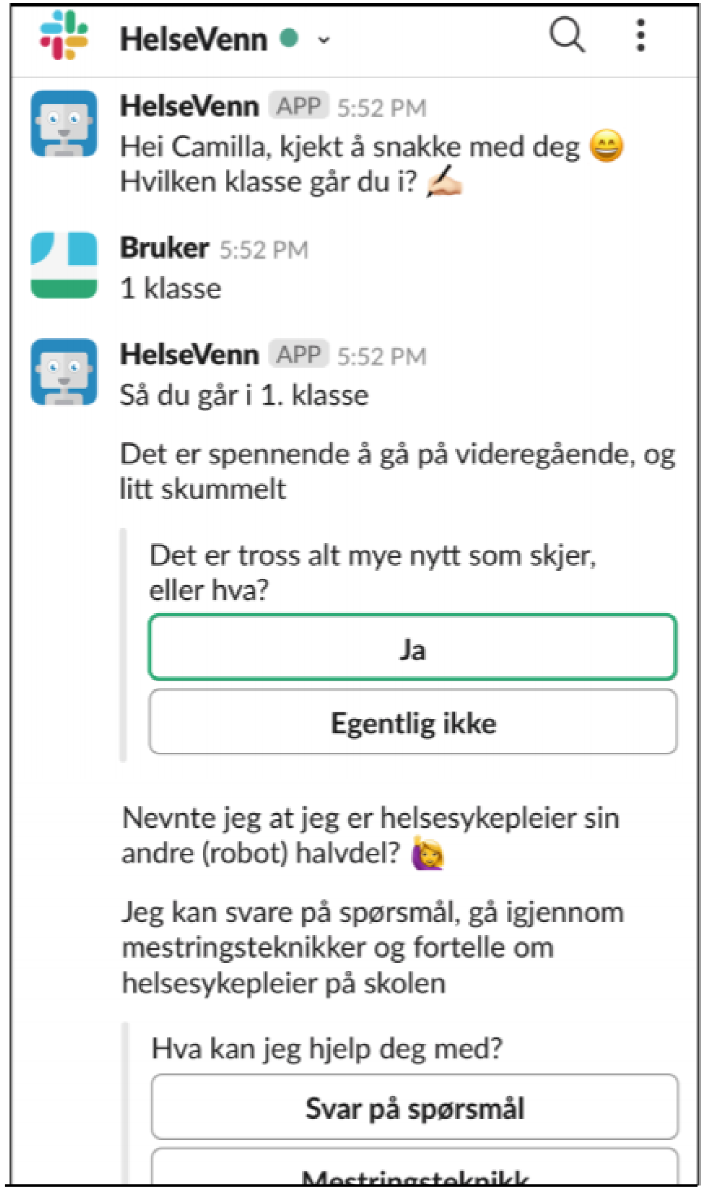


Implications

Conversation as design object

Necessary to move from UI design to service design

Necessary to design for networks of humans and bots



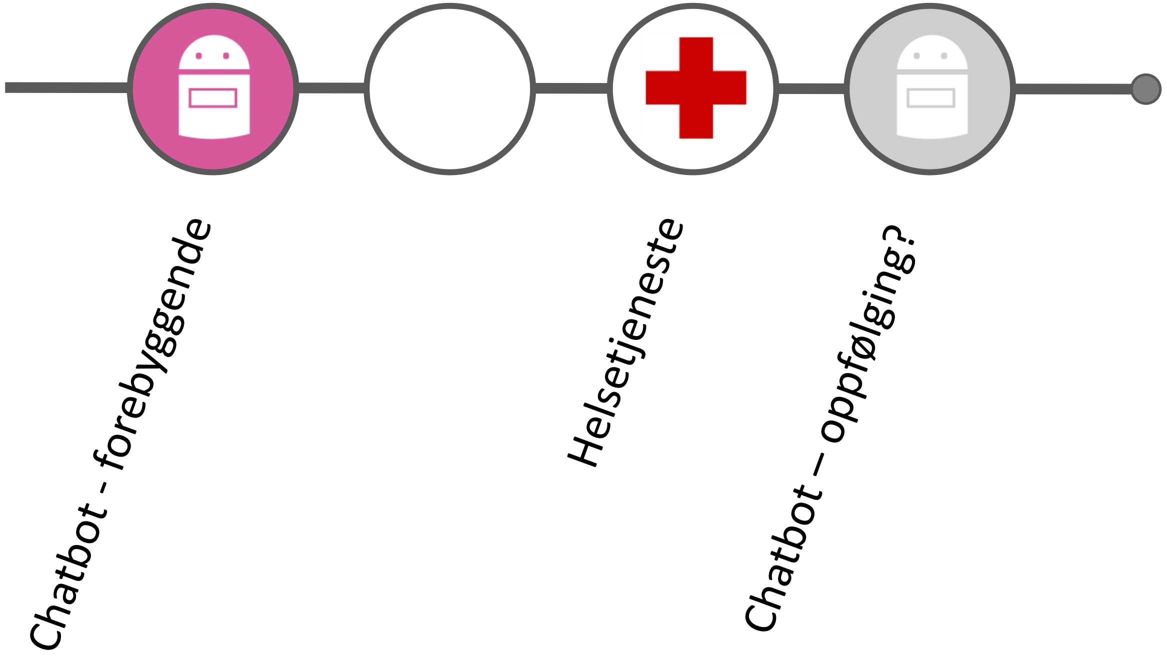
Høiland, C. (2019) “Hi, can I help?” An exploratory study of designing a chatbot to complement school nurses in supporting youths’ mental health. Master Thesis. UiO.

Implications

Conversation as design object

Necessary to move from UI design to service design

Necessary to design for networks of humans and bots



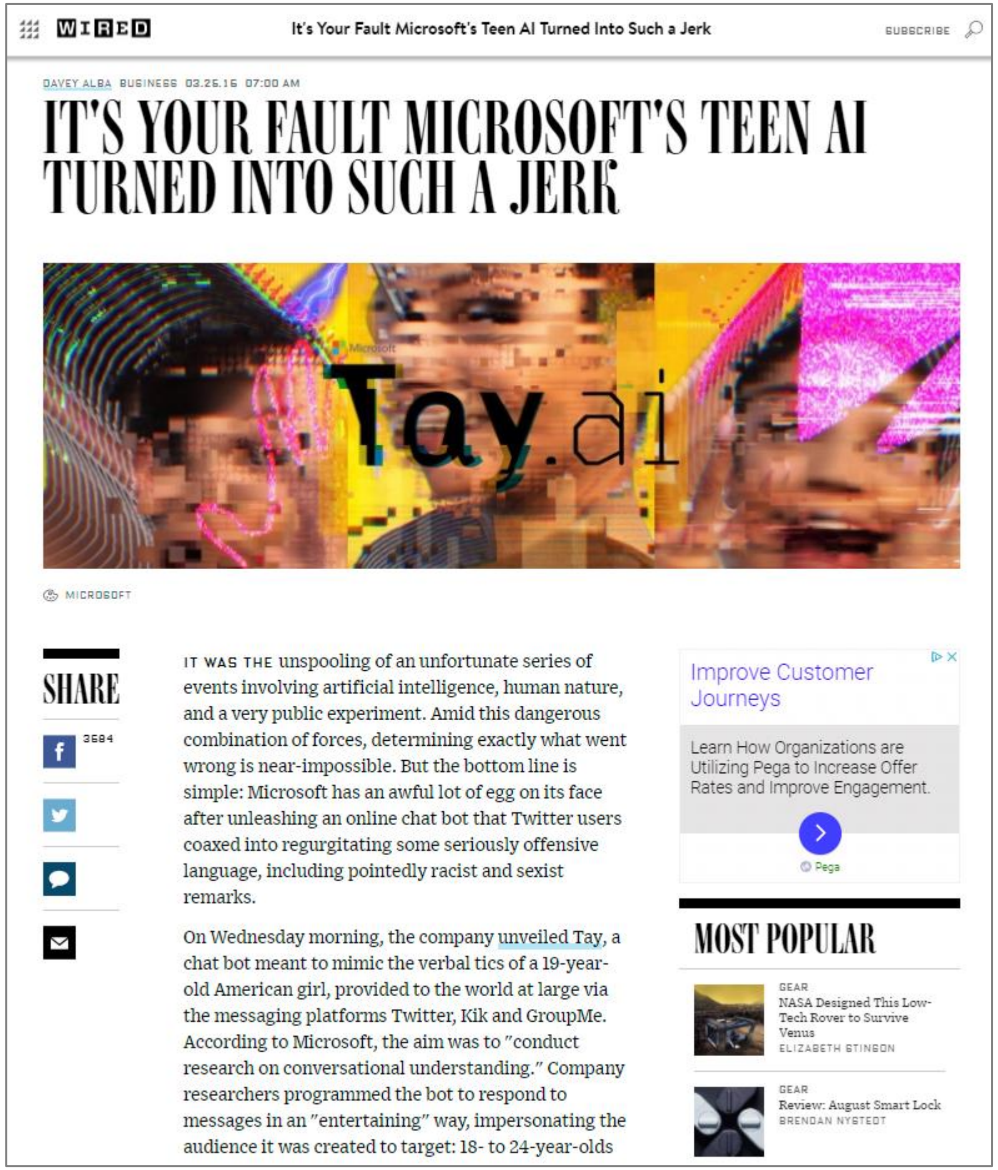
Social Health bots (www.sintef.no/socialhealthbots)

Implications

Conversation as design object

Necessary to move from UI design to service design

Necessary to design for networks of humans and bots


WIRED It's Your Fault Microsoft's Teen AI Turned Into Such a Jerk SUBSCRIBE

DAVEY ALBA BUSINESS 03.25.16 07:00 AM

IT'S YOUR FAULT MICROSOFT'S TEEN AI TURNED INTO SUCH A JERK

MICROSOFT

SHARE

3584

IT WAS THE unspooling of an unfortunate series of events involving artificial intelligence, human nature, and a very public experiment. Amid this dangerous combination of forces, determining exactly what went wrong is near-impossible. But the bottom line is simple: Microsoft has an awful lot of egg on its face after unleashing an online chat bot that Twitter users coaxed into regurgitating some seriously offensive language, including pointedly racist and sexist remarks.

On Wednesday morning, the company unveiled Tay, a chat bot meant to mimic the verbal tics of a 19-year-old American girl, provided to the world at large via the messaging platforms Twitter, Kik and GroupMe. According to Microsoft, the aim was to "conduct research on conversational understanding." Company researchers programmed the bot to respond to messages in an "entertaining" way, impersonating the audience it was created to target: 18- to 24-year-olds

MOST POPULAR

GEAR
NASA Designed This Low-Tech Rover to Survive Venus
ELIZABETH STINSON

GEAR
Review: August Smart Lock
BRENDAN NYSTEDT

https://www.wired.com/2016/03/fault-microsofts-teen-ai-turned-jerk/

Implications

Conversation as design object

Necessary to move from UI design to service design

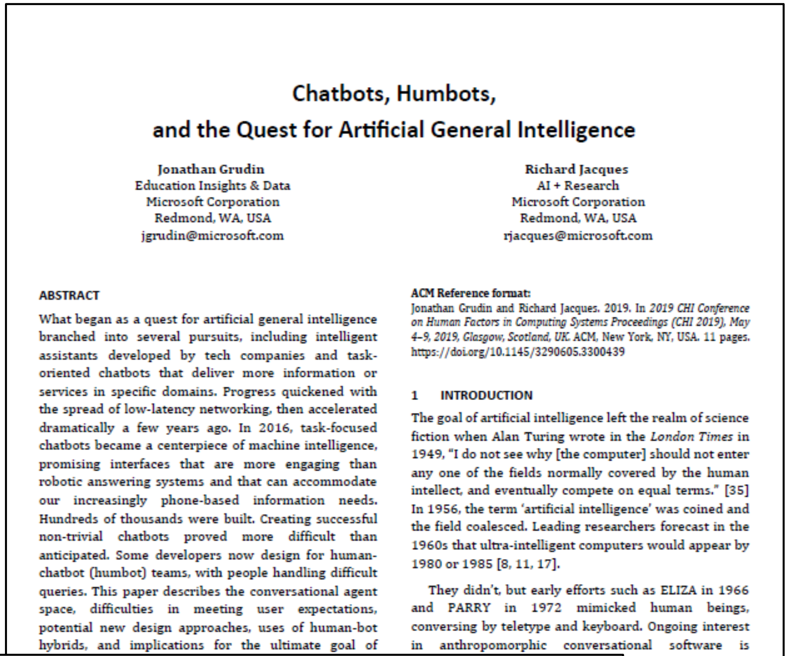
Necessary to design for networks of humans and bots

Implications

Conversation as design object

Necessary to move from UI design to service design

Necessary to design for networks of humans and bots



Chatbots, Humbots, and the Quest for Artificial General Intelligence

Jonathan Grudin
Education Insights & Data
Microsoft Corporation
Redmond, WA, USA
jgrudin@microsoft.com

Richard Jacques
AI + Research
Microsoft Corporation
Redmond, WA, USA
rjacques@microsoft.com

ABSTRACT

What began as a quest for artificial general intelligence branched into several pursuits, including intelligent assistants developed by tech companies and task-oriented chatbots that deliver more information or services in specific domains. Progress quickened with the spread of low-latency networking, then accelerated dramatically a few years ago. In 2016, task-focused chatbots became a centerpiece of machine intelligence, promising interfaces that are more engaging than robotic answering systems and that can accommodate our increasingly phone-based information needs. Hundreds of thousands were built. Creating successful non-trivial chatbots proved more difficult than anticipated. Some developers now design for human-chatbot (humbot) teams, with people handling difficult queries. This paper describes the conversational agent space, difficulties in meeting user expectations, potential new design approaches, uses of human-bot hybrids, and implications for the ultimate goal of

ACM Reference format:
Jonathan Grudin and Richard Jacques. 2019. In *2019 CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland, UK.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3290605.3300439

1 INTRODUCTION

The goal of artificial intelligence left the realm of science fiction when Alan Turing wrote in the *London Times* in 1949, "I do not see why [the computer] should not enter any one of the fields normally covered by the human intellect, and eventually compete on equal terms." [35] In 1956, the term 'artificial intelligence' was coined and the field coalesced. Leading researchers forecast in the 1960s that ultra-intelligent computers would appear by 1980 or 1985 [8, 11, 17].

They didn't, but early efforts such as ELIZA in 1966 and PARRY in 1972 mimicked human beings, conversing by teletype and keyboard. Ongoing interest in anthropomorphic conversational software is

…e, Zo, Hugging Face … these are found on

… has more limited … the subfield artificial …rtheless, intelligent …used chatbots could …eal the nature and …ired to realize it.

…ad precursors in …tworks and on the …low-latency internet …e agents such as … in the 1990s [22], …nguage.

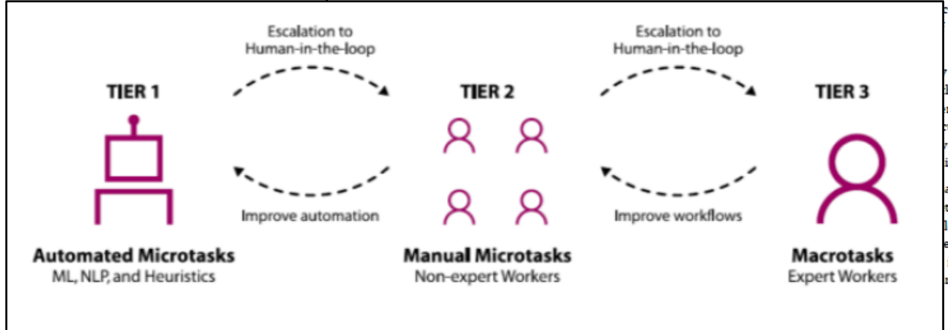


Figure 1: A three-tier architecture for handling tasks.

Challenges in current conversational agents

Interview study of 14 users of conversational agents

"Like Having a Really bad PA": The Gulf between User Expectation and Experience of Conversational Agents

Ewa Luger
Microsoft Research, UK
ewluge@microsoft.com

Abigail Sellen
Microsoft Research, UK
asellen@microsoft.com

ABSTRACT
The past four years have seen the rise of conversational agents (CAs) in everyday life. Apple, Microsoft, Amazon, Google and Facebook have all embedded proprietary CAs within their software and, increasingly, conversation is becoming a key mode of human-computer interaction. Whilst we have long been familiar with the notion of computers that speak, the investigative concern within HCI has been upon multimodality rather than dialogue alone, and there is no sense of how such interfaces are used in everyday life. This paper reports the findings of interviews with 14 users of CAs in an effort to understand the current interactional factors affecting everyday use. We find user expectations dramatically out of step with the operation of the systems, particularly in terms of known machine intelligence, system capability and goals. Using Norman's 'gulfs of execution and evaluation' [30] we consider the implications of these findings for the design of future systems.

Author Keywords
Conversational Agents; mental models; evaluation

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous

INTRODUCTION
Framed as "dialogue systems often endowed with 'humanlike' behaviour" [43 p.357], conversational agents (CA) are becoming ever more common human-computer interfaces. The launch of Siri (Apple, 2011), Google Now (2012), Cortana (Microsoft, 2015), and Alexa (Amazon, 2015) indicate a spike in mainstream market commitment to this form of experience and, in a departure from their traditional services, even Facebook have thrown down the gauntlet by launching 'M'; a hybrid dialogue system that employs both artificial intelligence and human responses to task requests. Equally, such products are no longer solely tied to the handset. Both Siri and Cortana are now core components of their respective operating systems and Alexa finds its home in the form of Amazon Echo, giving us every reason to believe that spoken dialogue interfaces will become the future gateways to many key services.

Whilst the past 4 years have clearly seen a reinvigoration of such systems, this is very much a return to an old idea; that conversation is the next natural form of HCI. It has also long been argued that "when speech and language interfaces become more conversational, they will take their place along with direct manipulation in the interface" [6]. Moreover, they will have the potential to enhance both the system usability and user experience [43]. However, despite these expectations, the weight of research has veered away from such single modalities and tended towards multimodal developments, with a focus upon embodiment and anthropomorphism rather than voice alone. Indeed, our fascination with computers that converse can be traced back as far as 1964 when, seeking to create the illusion of human interaction, Joseph Weizenbaum of MIT created Eliza [10], a computer program that responded on the basis of data gleaned only from human respondents' typed input. Whilst script-based, it is considered the first convincing attempt to simulate natural human interactions between a user and a computer. This chatterbot, rudimentary by today's standards, was designed in the form of a Rogerian psychotherapist and, due to the high level of emotional involvement exhibited by users, was hailed as the beginnings of an automated form of psychotherapy [45]. Fast-forward 50 years and, whilst psychotherapy-bots for the time being remain the stuff of science fiction, HCI is again seeing moves towards serious adoption of naturalistic human-computer dialogue systems.

However, despite tech giants vying to develop the most compelling experience, the field of HCI has developed little empirical knowledge of how such agents are used in everyday settings. Whilst CA research exists, it tends towards either technical papers related to architecture [37], CAs studied in experimental settings, or systems created for specific contexts, such as guiding users around a space [24], delivering information [41], or for the support of language learning [40]. Whilst each study brings us closer to understanding effective design, without concurrent knowledge of the pragmatics of everyday use, we fail to truly understand dynamics such as how and why such systems are used and "which factors influence acceptance and success in such scenarios" [24 p.329]. In light of this deficit, our paper seeks to understand user experience of CA systems by answering two simple questions; (a) what factors currently motivate and limit the

CHI'16, May 07 - 12, 2016, San Jose, CA, USA

ACM 978-1-4503-3362-7/16/05…$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858288

Challenges

Learning – from talking to chatbot as person to use restricted language

Require user effort – effective use require continuous investment

Lack of feedback – difficult to see capabilities and opportunities

Expectations not met – mismatch expectations and experience

"Like Having a Really bad PA": The Gulf between User Expectation and Experience of Conversational Agents

Ewa Luger
Microsoft Research, UK
ewluge@microsoft.com

Abigail Sellen
Microsoft Research, UK
asellen@microsoft.com

ABSTRACT

The past four years have seen the rise of conversational agents (CAs) in everyday life. Apple, Microsoft, Amazon, Google and Facebook have all embedded proprietary CAs within their software and, increasingly, conversation is becoming a key mode of human-computer interaction. Whilst we have long been familiar with the notion of computers that speak, the investigative concern within HCI has been upon multimodality rather than dialogue alone, and there is no sense of how such interfaces are used in everyday life. This paper reports the findings of interviews with 14 users of CAs in an effort to understand the current interactional factors affecting everyday use. We find user expectations dramatically out of step with the operation of the systems, particularly in terms of known machine intelligence, system capability and goals. Using Norman's 'gulfs of execution and evaluation' [30] we consider the implications of these findings for the design of future systems.

Author Keywords

Conversational Agents; mental models; evaluation

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous

INTRODUCTION

Framed as "dialogue systems often endowed with 'humanlike' behaviour" [43 p.357], conversational agents (CA) are becoming ever more common human-computer interfaces. The launch of Siri (Apple, 2011), Google Now (2012), Cortana (Microsoft, 2015), and Alexa (Amazon, 2015) indicate a spike in mainstream market commitment to this form of experience and, in a departure from their traditional services, even Facebook have thrown down the gauntlet by launching 'M'; a hybrid dialogue system that employs both artificial intelligence and human responses to task requests. Equally, such products are no longer solely tied to the handset. Both Siri and Cortana are now core components of their respective operating systems and Alexa finds its home in the form of Amazon Echo, giving us every reason to believe that spoken dialogue interfaces will become the future gateways to many key services.

Whilst the past 4 years have clearly seen a reinvigoration of such systems, this is very much a return to an old idea; that conversation is the next natural form of HCI. It has also long been argued that "when speech and language interfaces become more conversational, they will take their place along with direct manipulation in the interface" [6]. Moreover, they will have the potential to enhance both the system usability and user experience [43]. However, despite these expectations, the weight of research has veered away from such single modalities and tended towards multimodal developments, with a focus upon embodiment and anthropomorphism rather than voice alone. Indeed, our fascination with computers that converse can be traced back as far as 1964 when, seeking to create the illusion of human interaction, Joseph Weizenbaum of MIT created Eliza [10], a computer program that responded on the basis of data gleaned only from human respondents' typed input. Whilst script-based, it is considered the first convincing attempt to simulate natural human interactions between a user and a computer. This chatterbot, rudimentary by today's standards, was designed in the form of a Rogerian psychotherapist and, due to the high level of emotional involvement exhibited by users, was hailed as the beginnings of an automated form of psychotherapy [45]. Fast-forward 50 years and, whilst psychotherapy-bots for the time being remain the stuff of science fiction, HCI is again seeing moves towards serious adoption of naturalistic human-computer dialogue systems.

However, despite tech giants vying to develop the most compelling experience, the field of HCI has developed little empirical knowledge of how such agents are used in everyday settings. Whilst CA research exists, it tends towards either technical papers related to architecture [37], CAs studied in experimental settings, or systems created for specific contexts, such as guiding users around a space [24], delivering information [41], or for the support of language learning [40]. Whilst each study brings us closer to understanding effective design, without concurrent knowledge of the pragmatics of everyday use, we fail to truly understand dynamics such as how and why such systems are used and "which factors influence acceptance and success in such scenarios" [24 p.329]. In light of this deficit, our paper seeks to understand user experience of CA systems by answering two simple questions; (a) what factors currently motivate and limit the

CHI'16, May 07 - 12, 2016, San Jose, CA, USA

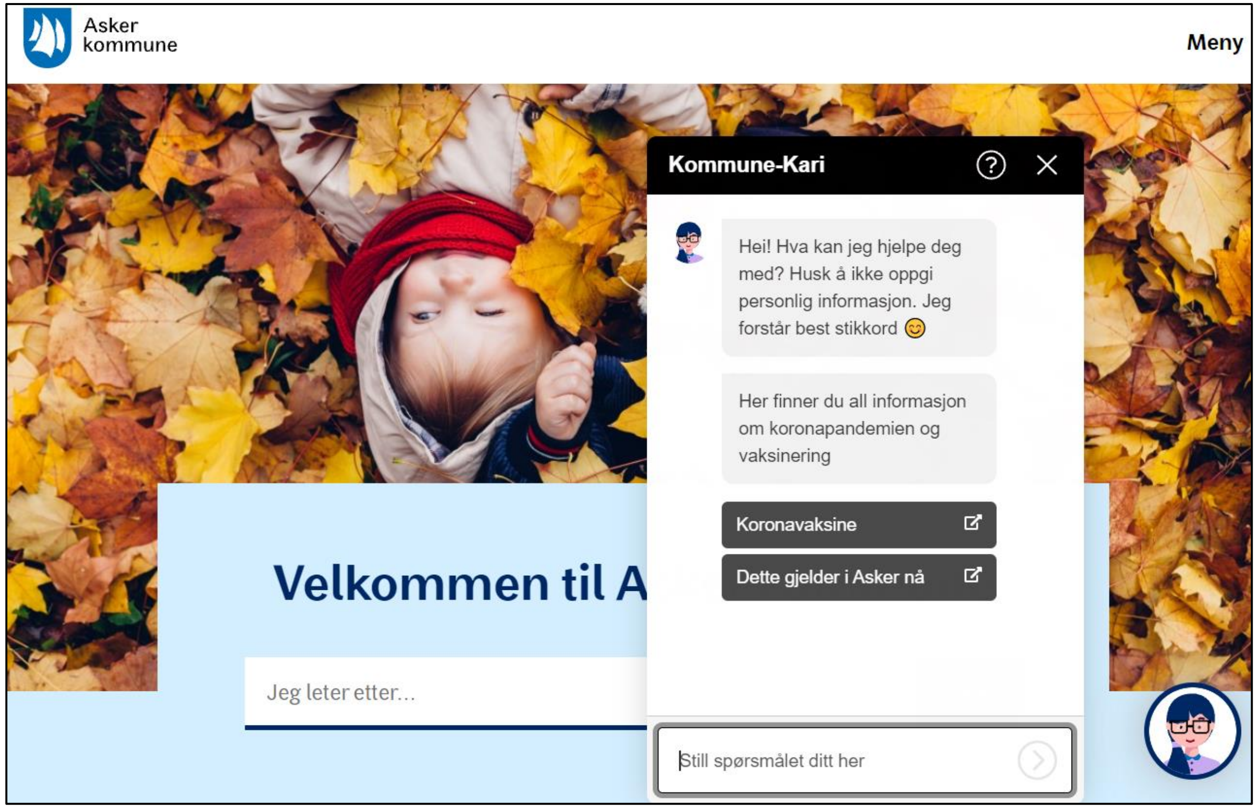
ACM 978-1-4503-3362-7/16/05…$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858288

H-AI-I guidelines and chatbot design

Discussion case – Kommune Kari

G1: Make clear what the system can do

G2: Make clear how well the system can do what it can do

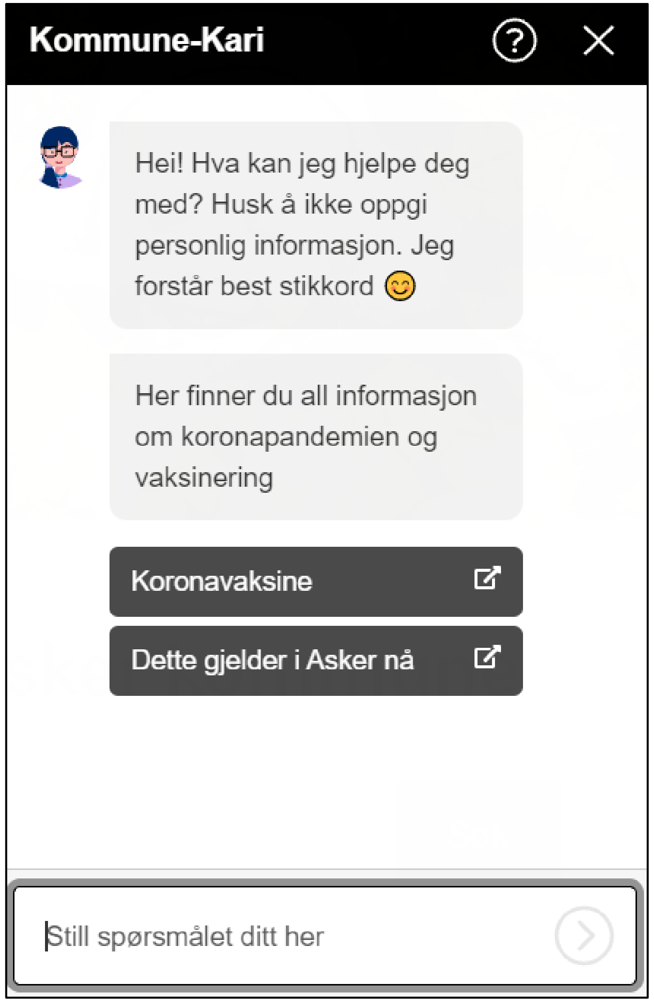


H-AI-I guidelines and chatbot design

Discussion case – Kommune Kari

G1: Make clear what the system can do

G2: Make clear how well the system can do what it can do



H-AI-I guidelines and chatbot design

Discussion case – Kommune Kari

G1: Make clear what the system can do

G2: Make clear how well the system can do what it can do