Characteristics of AI-infused systems

AI-infused systems are ' systems that have features harnessing AI capabilities that are directly exposed to the end user' (Amershi et al., 2019). Identify and describe key characteristics of AI-infused systems. Draw on the first lecture of Module 2 and three of the mandatory articles (Amershi et al. (2019), Kocielnik et al. (2019), Yang et al., (2020)).

AI-infused systems are systems that have features harnessing AI capabilities that are directly exposed to the end user' (Amershi et al., 2019). Key characteristics for these kinds of systems are that uncertainty, inconsistency, and behind-the-scenes personalization (Amershi et al., 2019), and probabilistic, impacted by user action and transparency issues (Kocielnik et al, 2019).

Uncertainty implies that the AI-infused system may be more prone to errors than deterministic systems, both in the form of false positives, and false negatives. Inconsistency in this context implies that changes that may seem small to the user, can have big consequences for the output of the system.

Identify one AI-infused system which you know well, that exemplifies some of the above key characteristics. Discuss the implications of these characteristics for the example system, in particular how users are affected by these characteristics.

An example AI-infused system is my Outlook e-mail spam filter. It is both my bane and my blessing. There is visible inconsistency in how it sometimes filters my rejection letters for applications for summer jobs, which I wish to not delete but the same in a folder, thus getting a feeling of my time spent on them is worth something. Because of this I as a user need to regularly check my spam folder, as the content is deleted after a week if not recovered.

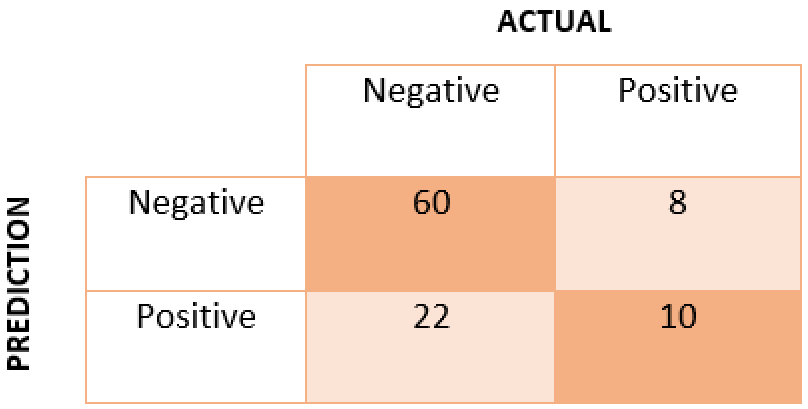
Another issue with the outlook spam filter is the transparency issues; as a user, I have no insight into how the email system picks out spam. As a result, I don't know if the email I'm waiting for, like a response to a summer job application I did not feel qualified for, will go to my spam folder or not. Thus, I must check two folders instead of one.

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.

Amershi proposes 18 guidelines for interaction with AI-infused systems with a GUI and backs the guidelines up with validation from user testing. The reason for proposing these is that there is a rapid inclusion of AI in computing systems, and designers need new guidelines to tackle these new tools.

Kocielnik et al explain that the general public has vastly different expectations of AI systems, and explore how these expectations are affected when two different focuses are given to two AI systems for meeting scheduling, avoiding false positives, and avoiding false negatives, with the same accuracy (See illustration).



		ACTUAL	
		Negative	Positive
PREDICTION	Negative	60	8
	Positive	22	10

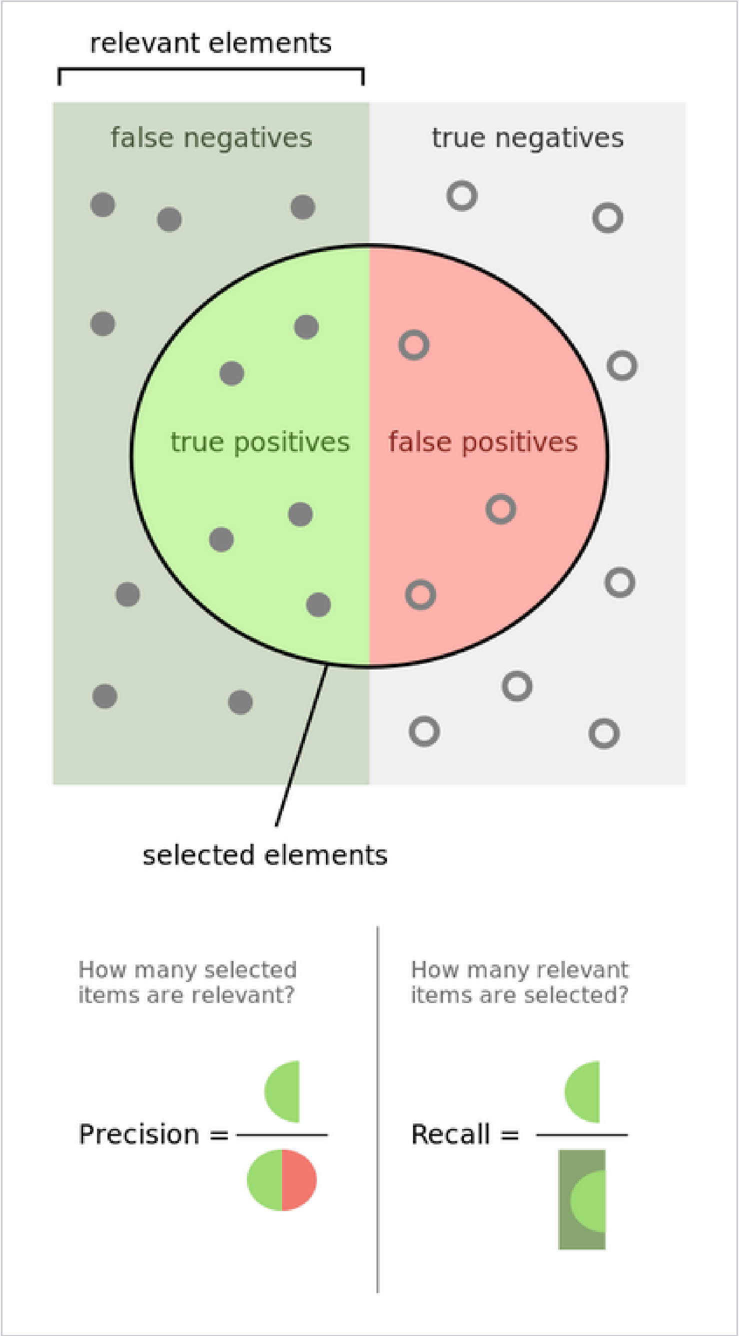
[1]

Here the accuracy would have been true (60+10)/(60+8+10+22) = 70/100 =70% accuracy.

The research shows that the system that provided high recall (of all possible meetings, it captures most of them) was more satisfactory than the high precision (how many of the proposed meetings were actually meetings) version (Helpful figure for differences between these further down). This means that for email meeting

[1] https://www.google.com/url?sa=i&url=https%3A%2F%2Ftowardsdatascience.com%2Fperformance-metrics-confusion-matrix-precision-recall-and-f1-score-a8fe076a2262&psig=AOvVaw1M95_AForKjhuh85lMZenY&ust=1634901050766000&source=images&cd=vfe&ved=0CAwQjhxqFwoTCJigyZav2_MCFQAAAAAdAAAAABAJ, 21.10.21

booking AI, it was better to suggest to many meetings and get them all than to propose only possible meetings that were *indeed meetings*, with the danger of missing some. With the high precision system, expectation adjustment techniques proved effective, but this may be because the users disliked it so much in the first place.



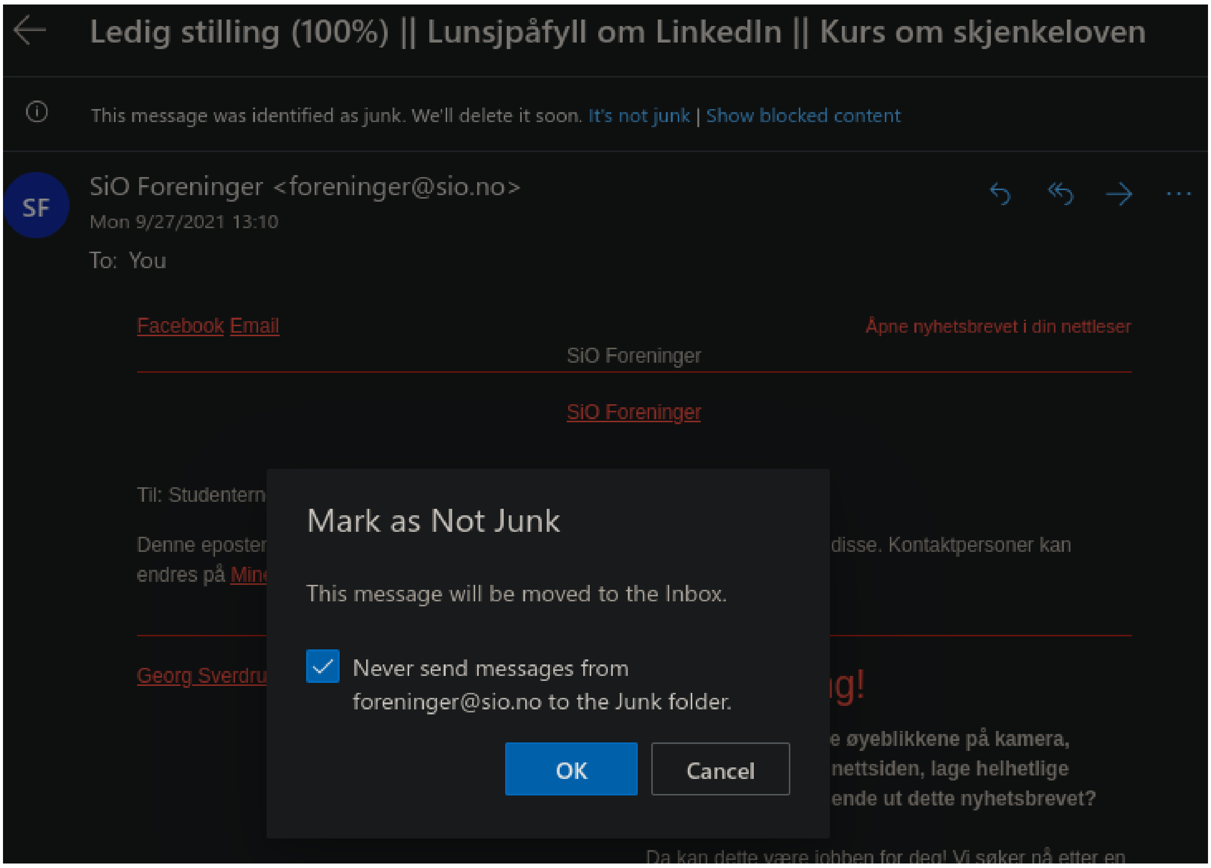
Precision and recall

https://en.wikipedia.org/wiki/Precision_and_recall, *21.10.2021.*

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.

Guideline 1 is “Make clear What the System can do”, and guideline 2 is “Make clear how well the system can do what it can do”. In the case of the Outlook email spam filter, the system clearly shows me what it can do: Recognize spam and filter it into its own directory. The system does not follow guideline 2, as there is no visible explanation for me as a user for how often or why the system makes the mistakes it makes.

A possible improvement using the system could be if it better explained to me how I could configure it to do what I wish it to do. When i mark a message as “not spam”, this pop-up is given:



Ledig stilling (100%) || Lunsjpåfyll om LinkedIn || Kurs om skjenkeloven

This message was identified as junk. We'll delete it soon. It's not junk | Show blocked content

SiO Foreninger <foreninger@sio.no>
Mon 9/27/2021 13:10
To: You

Facebook Email

Åpne nyhetsbrevet i din nettleser

SiO Foreninger

SiO Foreninger

Til: Studentern

Denne eposter ... disse. Kontaktpersoner kan endres på Mine

Mark as Not Junk

This message will be moved to the Inbox.

☑ Never send messages from foreninger@sio.no to the Junk folder.

Georg Sverdru

...g!

...e øyeblikkene på kamera, ...nettsiden, lage helhetlige ...ende ut dette nyhetsbrevet?

Da kan dette være jobben for deg! Vi søker nå etter en

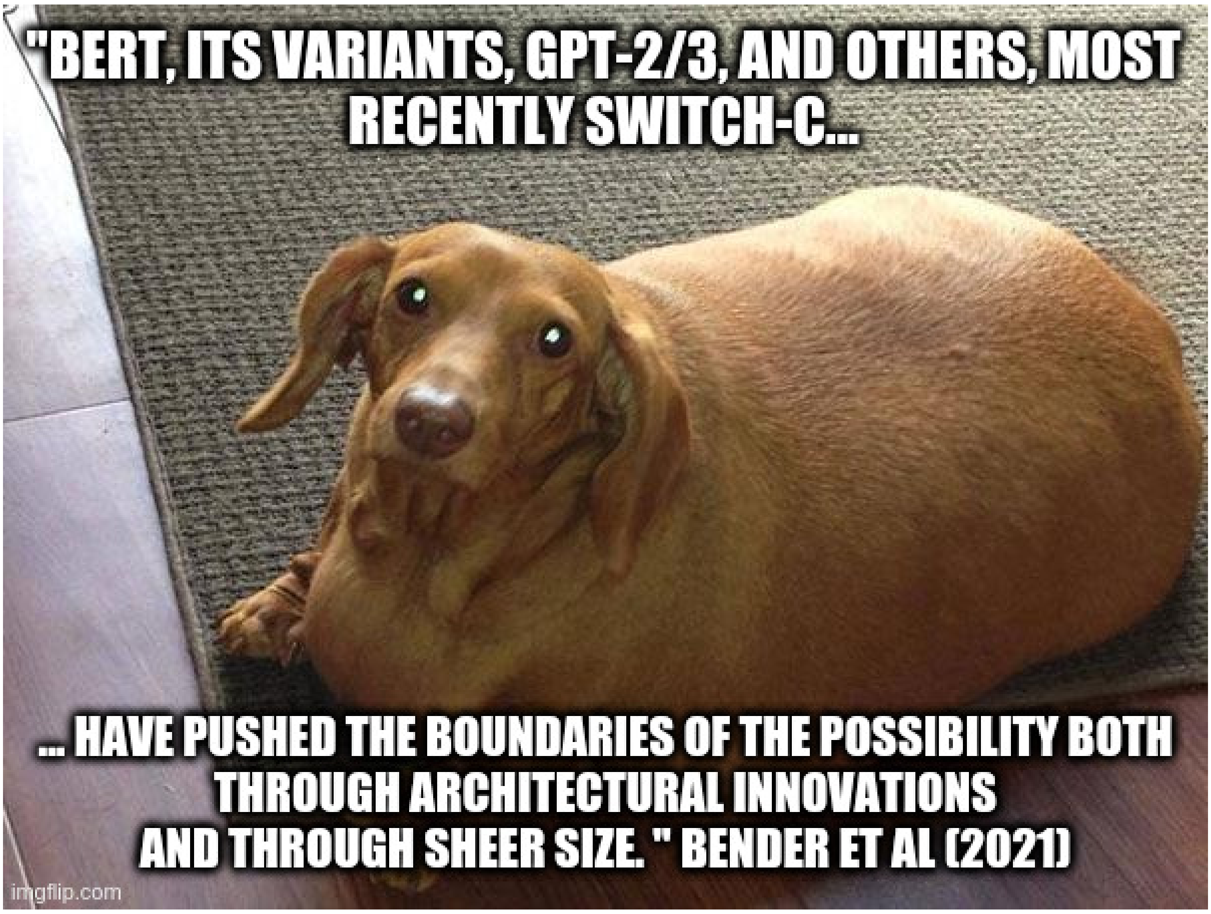
Private photo.

I want more of this kind of clear and easy configuration, that tells me how well the system will sort my email in the future. As for guideline 1, I only wish I could get an explanation for why all the shitty newsletters from Kiwi and Dressmann are not yeeted straight into the spam folder abyss where they belong.

Future outlook AI setting things right

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 language models.

As the illustration above suggests, Bender et al (2021) suggests that a language model based on larger datasets does not necessarily achieve the goals of a language model better than a smaller, better curated dataset.


"BERT, ITS VARIANTS, GPT-2/3, AND OTHERS, MOST RECENTLY SWITCH-C...
... HAVE PUSHED THE BOUNDARIES OF THE POSSIBILITY BOTH THROUGH ARCHITECTURAL INNOVATIONS AND THROUGH SHEER SIZE. " BENDER ET AL (2021)
imgflip.com

By focusing on curating and testing a smaller dataset, the researchers propose that inherent bias in the example texts can be combated. There is also the issue of extremist views the users may express being encouraged by the system. The climate impact of training larger and larger NLP models is also not negligible (see link to the cloud matters project, i recommend it highly: https://infopoetry.densitydesign.org/infopoetries/cloud-matters.html).

There is also a non-negligible financial cost in using ever larger datasets in the form of time, computing and requirements for hardware, that creates barriers for smaller researcher and developer teams ability to contribute to the field.

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.

I encourage you to Google "Tay chatbot" and go down the rabbit hole

By using conversations as a design object, the lecturer stressed the importance of moving from a UI design perspective into a service design perspective (Følstad, 2021). Chatbots will often be a link in a chain of services, intertwined with human provided services and computer-hosted tools.

Four key challenges in the design of chatbots are (Luger, E., & Sellen, A. (2016):

- Learning to talk to the chatbot in a way it understands, as opposed to talk to it like a person
- Effective use requires continuous investment by the users
- There is a lack of feedback from the system, which makes it difficult to understand the capabilities and limitations of the chatbot
- There is a large mismatch between expectations and the actual experience of using chatbots

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.

Guideline 1 is “Make clear What the System can do”, and guideline 2 is “Make clear how well the system can do what it can do”:

In relation to chatbots, if the users were informed in an unobtrusive, effective manner of what the chatbot could do, they would mabey faster learn to talk to the chatbots in a way they can understand. Expectations could also more easily be matched, so that the users don't ask the NAV Frida chatbot to approve their applications.

Guideline 2 could help mitigate the continuous investment needed by the users; if the users know that a certain level is about as good as it gets, they may be willing to train the AI to reach that point, as the goal will be defined. A clear explanation of how well the system could perform will also help lower expectations, thus mediating the disappointment felt when NAV Frida does not approve you application.

References:

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)

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)

“Precision and recall”, unknown, downloaded 21.10.2021, https://en.wikipedia.org/wiki/Precision_and_recall,

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 *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (*FAccT '21*). Association for Computing Machinery, New York, NY, USA, 610–623. DOI:https://doi.org/10.1145/3442188.3445922

Asbjørn Følstad, lecture 20. October 2021, UIO Oslo.

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. (https://www.microsoft.com/en-us/research/wp-content/uploads/2016/08/p5286-luger.pdf)