Arathi Sethumadhavan, PhD
Samuel Levulis, PhD
Ethics & Society
Microsoft

ETHICS + SOCIETY

Designing Transparent AI

Ethical Principles

FAIRNESS

Treat all stakeholders equitably and prevent undesirable stereotypes and biases.

TRANSPARENCY

Create systems and outputs that are understandable to relevant stakeholders.

RELIABILITY

Build systems to perform safely even in the worst-case scenario.

PRIVACY & SECURITY

Protect data from misuse and unintentional access to ensure privacy rights.

INCLUSION

Empower everyone, regardless of ability, and engage people by providing channels for feedback.

ACCOUNTABILITY

Take responsibility for how systems operate and their impact on society.

Background

Artificial Intelligence (AI) has the potential to improve many areas of our lives, but the users of AI-based systems typically have a limited understanding of how these systems make decisions.

Transparency: "The descriptive quality of an interface pertaining to its abilities to afford an operator's comprehension about an intelligent agent's intent, performance, future plans, and reasoning process."

AI is seldom 100% reliable. To minimize unintended consequences associated with reliance on AI, and calibrate the right level of trust in the AI, it is important to convey the reliability of the AI to the users of the system.

Chen et al., 2014

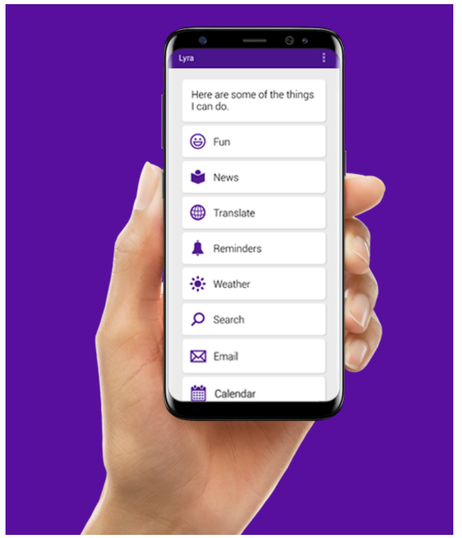
Design principles for increasing AI transparency

1. Make the capabilities and purpose of the AI clear.
2. Set appropriate expectations about what the AI can and cannot do.
3. Provide accurate and timely feedback.
4. Present contextual information about AI reliability.
5. Display information about the source of an AI failure and what users should do in a given situation.
6. Group and isolate less reliable or vulnerable AI components/functions so that user trust of reliable components/functions does not erode.
7. Provide information about how the AI algorithm works.
8. Provide information about the uncertainty of the AI's predictions and decisions.
9. Allow users to provide feedback to improve the accuracy of the AI system
10. Balance AI transparency and information overload wisely.

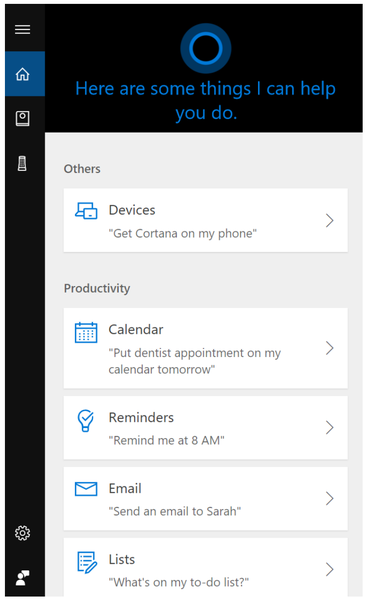
PRINCIPLE 1

Make the capabilities and purpose of the AI clear

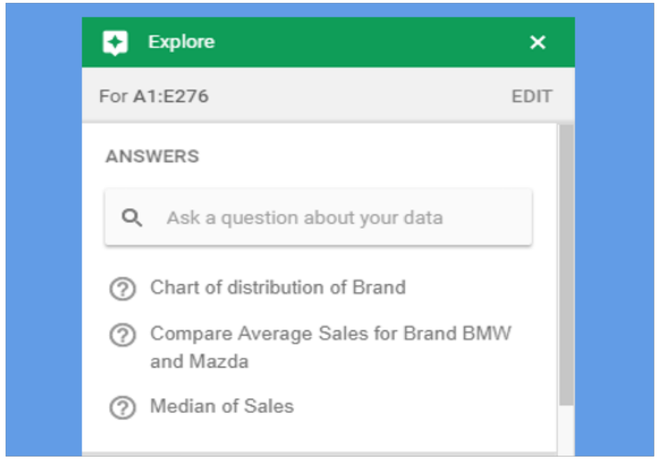
To the extent possible, information about the AI's purpose, process, and performance should be made transparent to users to help them establish an appropriate level of trust in the AI.



The Lyra personal assistant app provides information about some of the things it can do, such as translation or looking up the weather.



Cortana provides examples of some of the types of tasks that it can perform, such as putting an appointment on the user's calendar or setting a reminder.



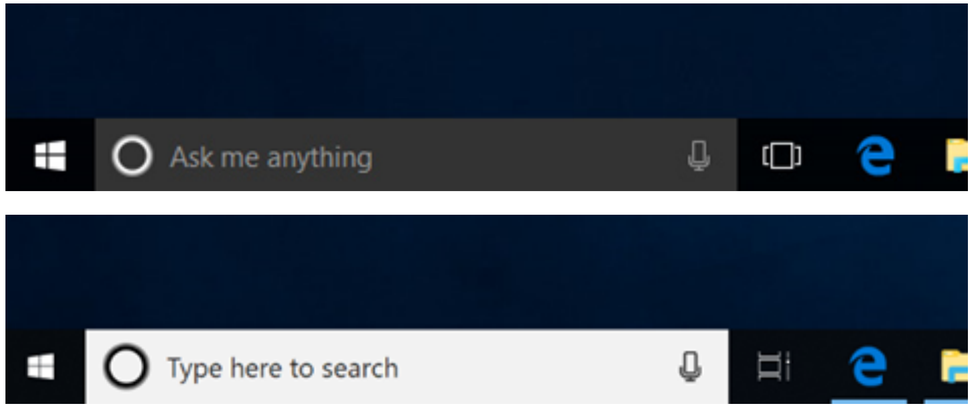
GSuite's Explore provides sample questions that users can ask it about their data.

Lee & Moray, 1994; Lee & See, 2004; https://www.heylyra.com

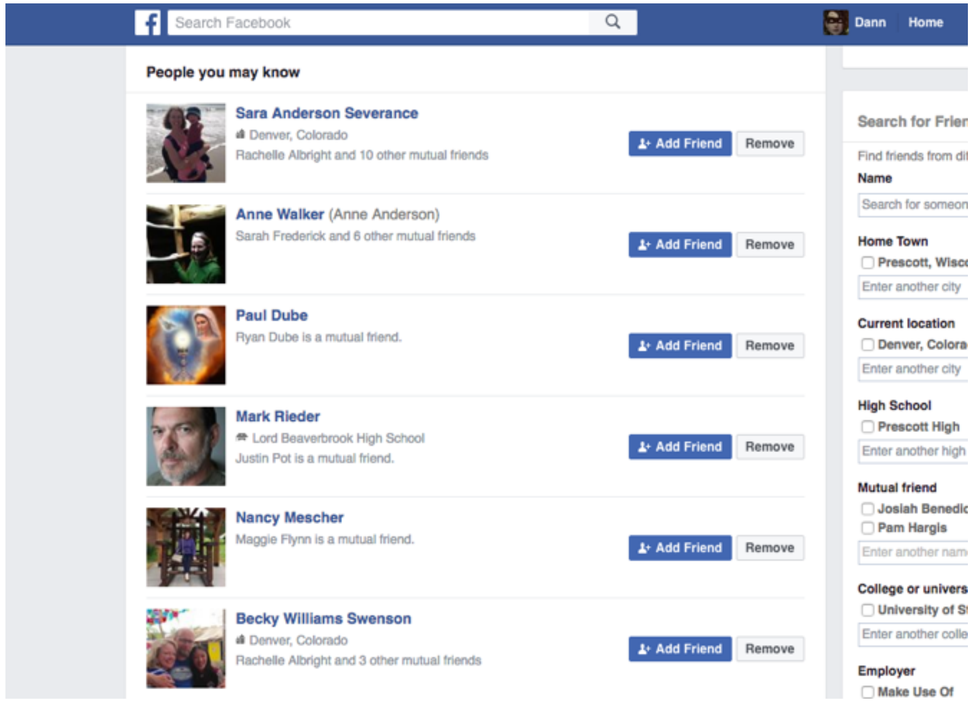
PRINCIPLE 2

Set appropriate expectations about what the AI can and cannot do

Explaining the behavior of the AI in ways that users understand helps them create a mental model of system capabilities, and more appropriately calibrate trust in the system.



"The "Ask me anything" in Cortana's task bar was changed to "Type here to search".



Facebook phrases their recommendations as "People you may know" rather than "People you know."

de Graaf & Malle, 2017; Collisson et al., 2017

PRINCIPLE 3

Provide accurate and timely feedback

Google Home provides feedback that it is listening for a query via the four colored LED lights that spin when the device hears a hotword, “Ok Google” or “Hey Google.”

Designing systems that provide users with accurate feedback about their reliability or how they operate facilitates appropriately-calibrated trust.

Feedback can typically be provided at various stages of interaction with a system.

Siri displays a transcription of the user’s query (“Who won the Seattle Seahawks game?”) as a means of providing feedback regarding what was heard.

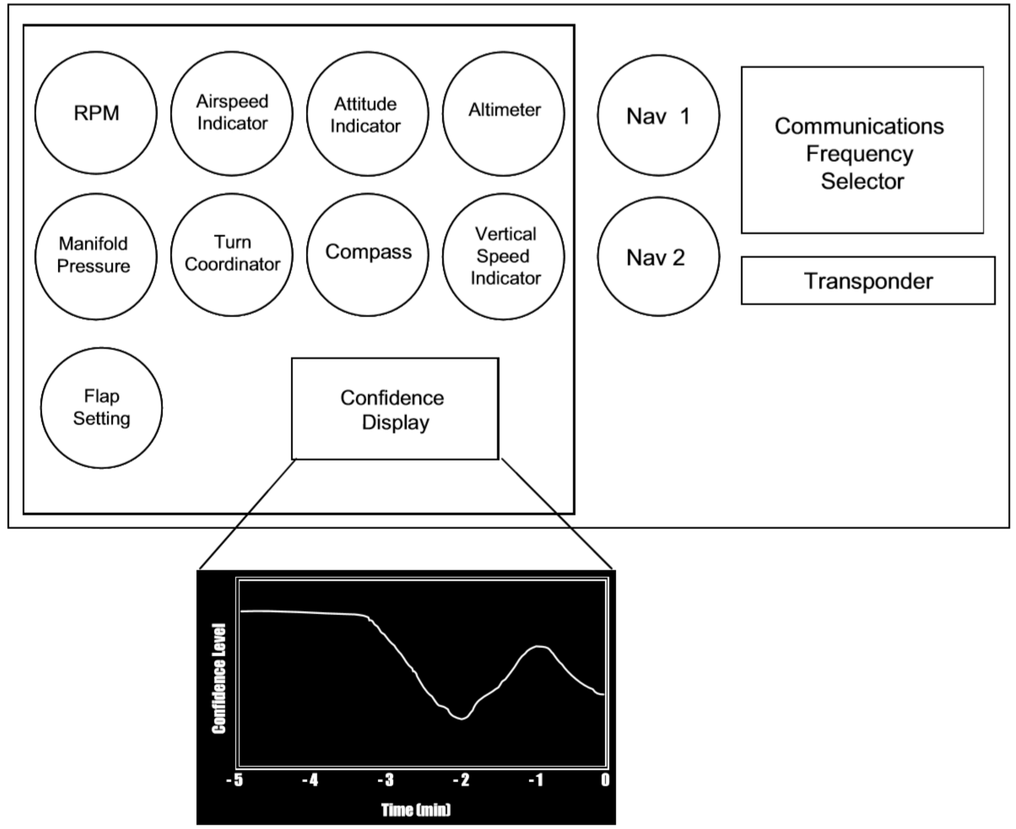
Dadashi, Stedmon, & Pridmore, 2012; Endsley, 2017; Gao & Lee, 2006; Hoff & Bashir, 2015; Jamieson, Wang, & Neyedli, 2008; Neyedli, Hollands, & Jamieson, 2011; Seong & Bisantz, 2008; Parasuraman & Riley, 1997; Wang, Jamieson, & Hollands, 2009; Source: Google Home Help

PRINCIPLE 4

Present contextual information about AI reliability

An AI system's reliability is usually context-dependent, as functions are typically sensitive to system settings or changes in external conditions.

To ensure appropriate trust, contextual information about changes in a system's reliability should be made apparent to the user.



Confidence trend display that assists pilots with detecting and handling in-flight icing encounters, and surrounding instruments.

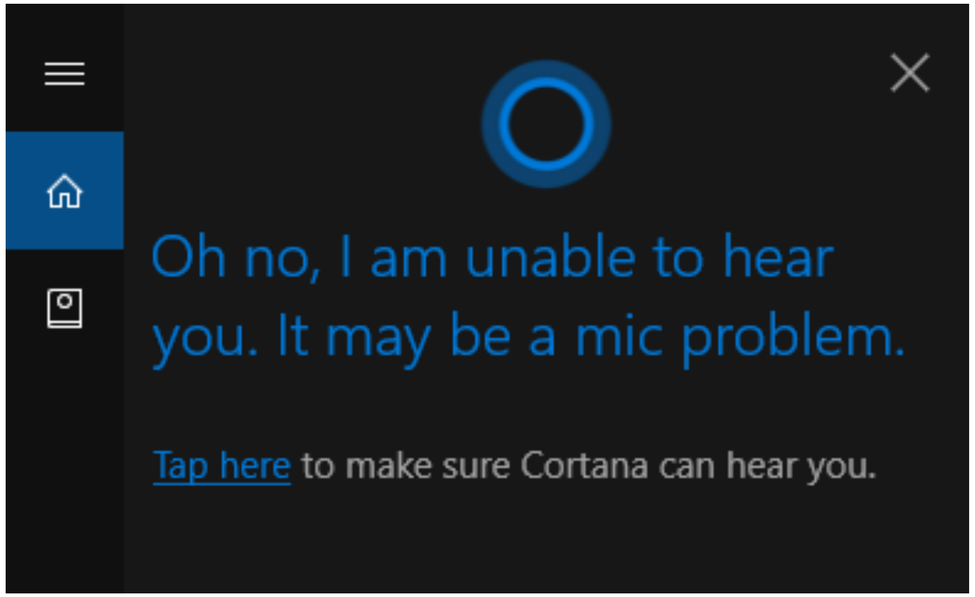
Rovira et al., 2014; McGuirl & Sarter, 2003; Muir, 1988; Lee and Moray, 1994

PRINCIPLE 5

Display information about the source of an AI failure and what users should do in a given situation

During situations in which an AI system does fail due to less than perfect system reliability, information about the source of the failure should ideally be presented to the user.

The automated functions/decisions that are more susceptible to error should be made apparent to users, along with the factors that likely contributed to any given failure.



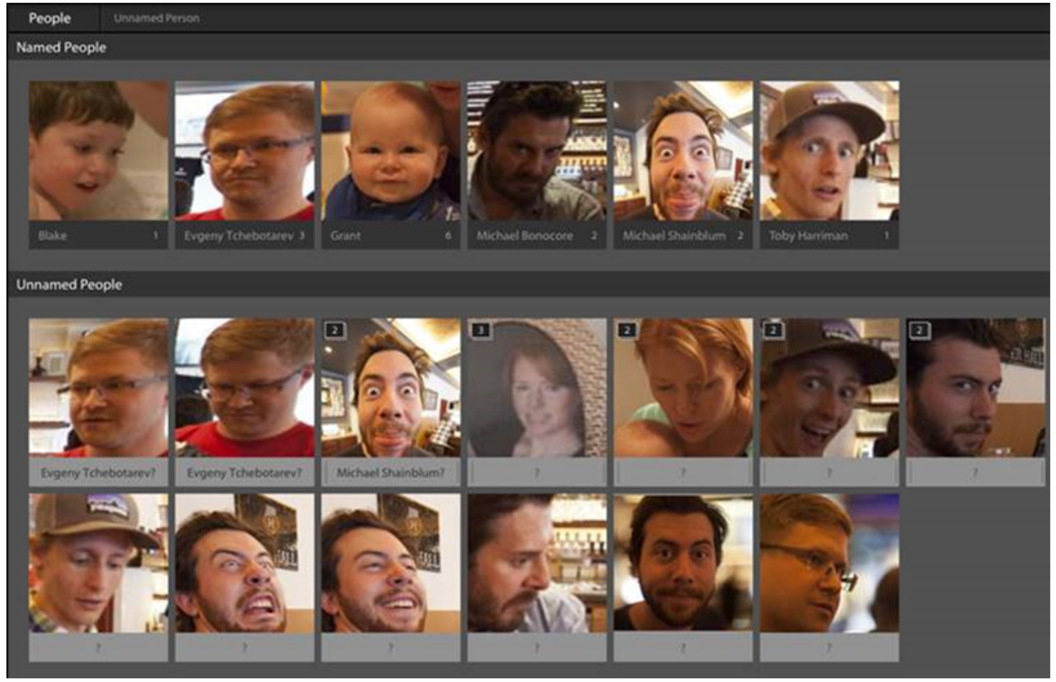
The feedback Cortana provides in Windows 10 when it cannot hear the user.

Atoyan et al., 2006; Billings, 1997; Helldin, 2014

PRINCIPLE 6

Group and isolate less reliable or vulnerable AI components/functions so that user trust of reliable components/functions does not erode

To reduce the likelihood of distrust spreading to similar but functionally independent systems, a distinction between reliable and unreliable system components should be made apparent to the user when AI fails.



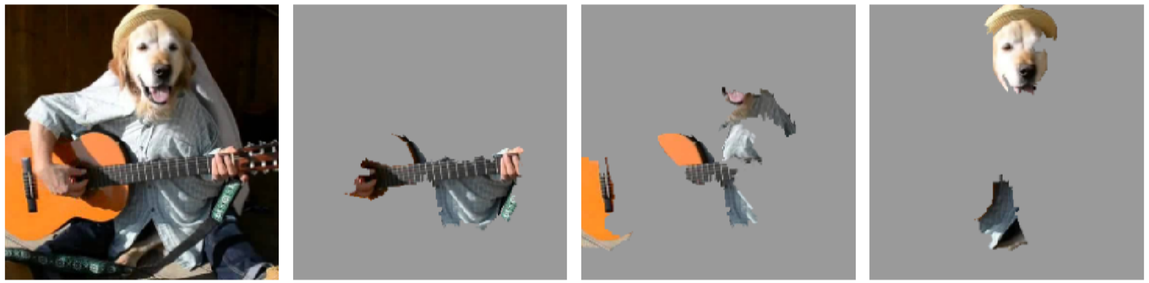
Adobe Lightroom's automatic naming functionality separates pictures that it is confident in naming from those that it is unsure of.

Collisson et al., 2017; Lee & Moray, 1992; Muir & Moray, 1996

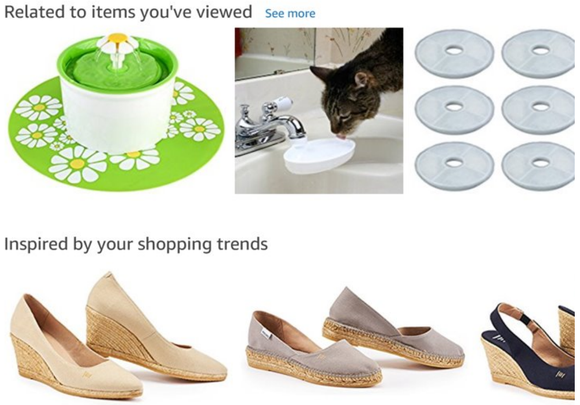
PRINCIPLE 7

Provide information about how the AI algorithm works

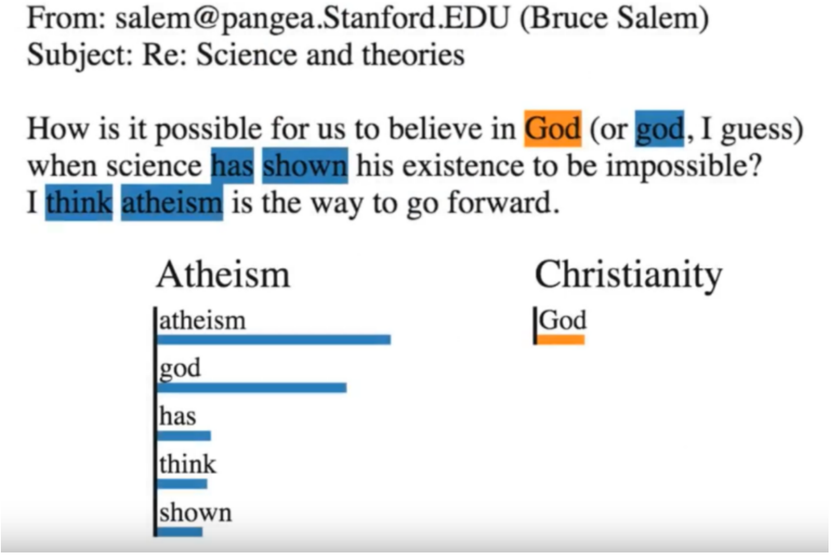
When feasible, users should be given access to information about an AI system's algorithm because users tend to trust and rely on the AI more appropriately if they understand how the system makes its decisions.



A LIME explanation of an image classification prediction made by Google's Inception neural network. The top three classes predicted for the original image are "Electric Guitar" (p = 0.32), "Acoustic guitar" (p = 0.24), and "Labrador" (p = 0.21).



When providing recommendations, Amazon gives users information about why it made particular recommendations.



A LIME explanation of the basis of a text-classification platform's prediction of whether or not an email is about atheism.

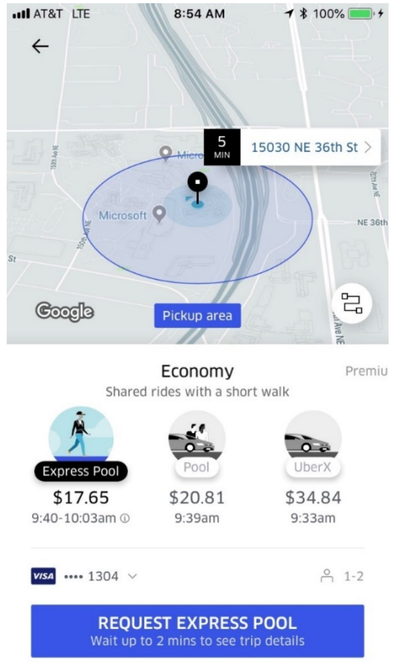
de Graaf & Malle, 2017; Collisson et al., 2017; Glass et al. 2008; Madumal et al., 2018; Pazzani 2000; Rovira et al., 2014; Koh & Liang, 2017; Ribeiro et al., 2016a, 2016b

PRINCIPLE 8

Provide information about the uncertainty of the AI's predictions and decisions

Research has found that displays that notify users of prediction uncertainty can preserve user trust towards the display and increase user performance.

Netflix presents information about the uncertainty of its suggestion.



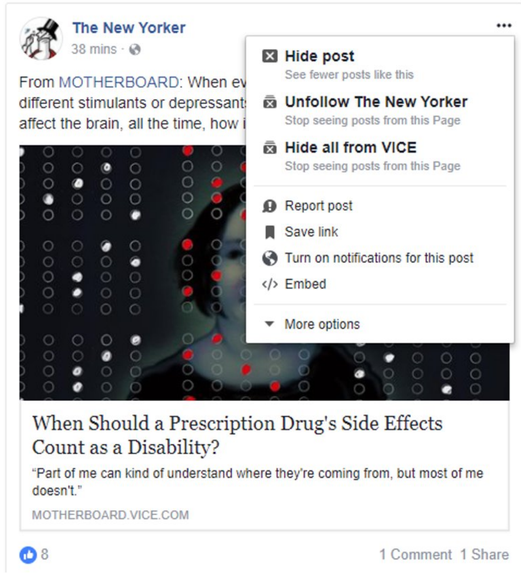
Uber's Express Pool feature provides information about the uncertainty in its estimate of the rider's arrival by presenting a range of potential arrival times.

Jung et al., 2015; Kay et al., 2016; Wunderlich et al., 2017

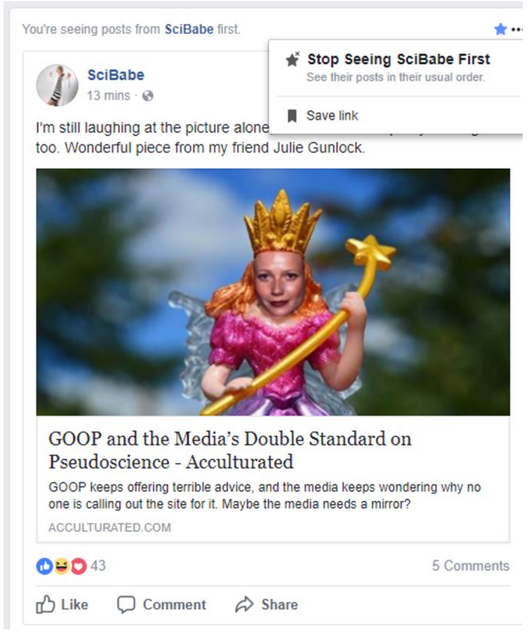
PRINCIPLE 9

Allow users to provide feedback to improve the accuracy of the AI system

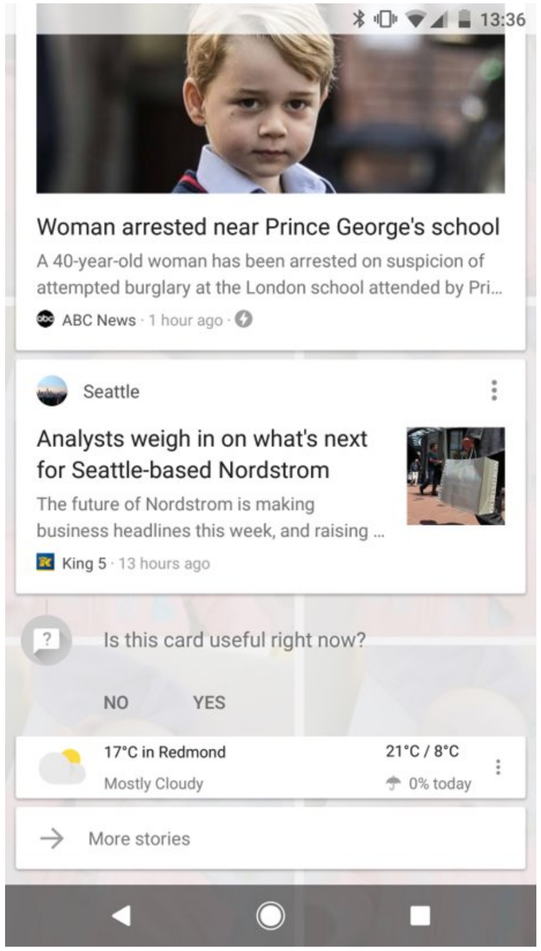
Providing users with the opportunity to provide feedback to an AI system has the potential to increase the accuracy of algorithms.



Facebook lets users give feedback to its algorithms with the option to "see fewer posts like this."



Facebook lets users give feedback regarding how posts should be weighted in their feed.



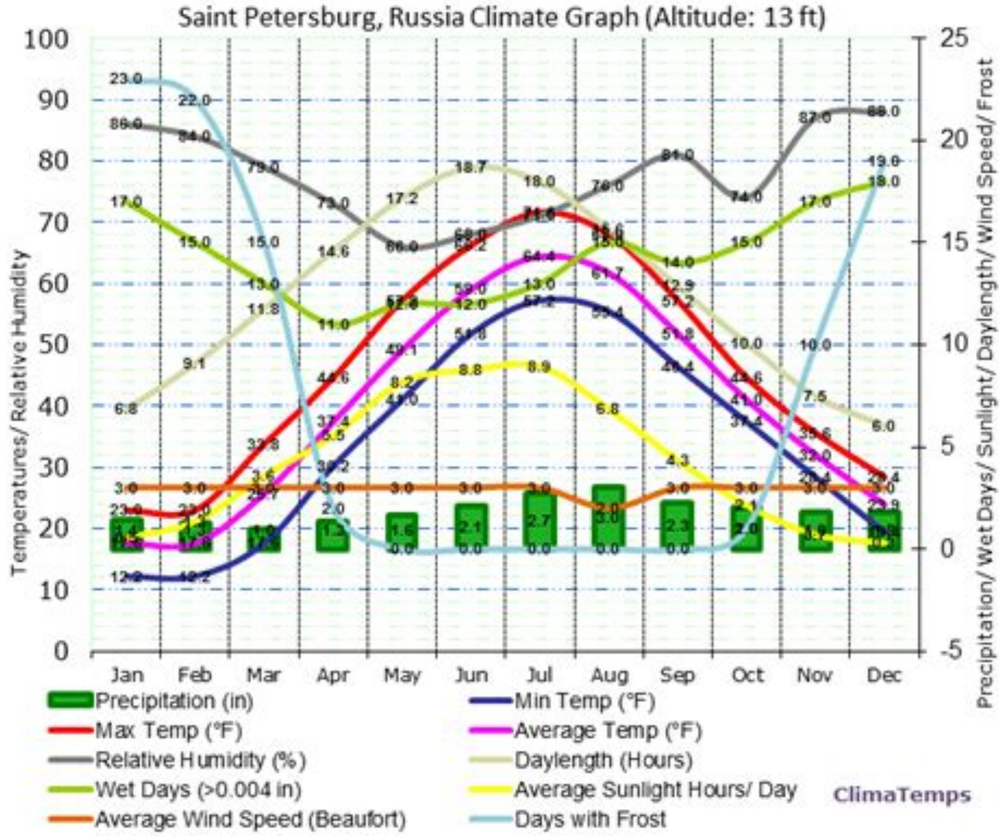
News in the Google App allows feedback through yes/no selections.

Collisson et al., 2017; Stumpf et al., 2007

PRINCIPLE 10

Balance AI transparency and information overload wisely

The goal of system transparency is not to present all of the AI's capabilities, behaviors, and decision-making rationale to the user; ideally the system should present information as succinctly as possible to allow users to maintain situation awareness without becoming overloaded.



A weather chart that overloads users with information

Chen et al., 2014; Lee & See, 2004; Medlock, 2016; Mercado et al., 2016; Perry, 2016

DISCUSSION

Additional Considerations

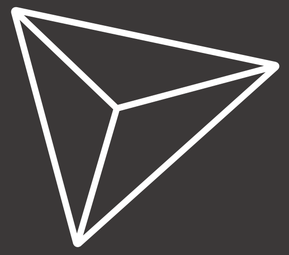
The principles to be implemented to increase AI transparency may be dependent on a number of factors, including:

- Who needs to understand the AI model?
- What do they need to understand? Why?
- What do they already know?
- What type of data is it?
- What is their context of use?

Methods for increasing transparency should be validated through user testing.

Tradeoffs between increases in transparency and other constraints (e.g., technical requirements, policy, standards,) should be examined.

Kronos (2018)



ETHICS + SOCIETY

Microsoft

Thank you

ARSETHUM@MICROSOFT.COM - SALEVULI@MICROSOFT.COM