

# Guidelines for Human-AI Tnteraction

Dinh Duy Kha

# Outline

# Paper: Guidelines for Human-AI interaction

- ▶ Authors: Saleema Amershi et. al., Microsoft Research
- ▶ Published in: CHI 2018

## Overview

- ▶ 18 design guidelines for human-AI interaction is composed from over 150 design recommendations from over 20 years of learning in AI design
- ▶ The guidelines are validated through multiple rounds user study
  - ▶ The results verify the relevance of them and reveals gaps in our knowledge about HAI

# Motivation

- ▶ *AI-infused systems* ( System that have AI features ) have uncertainties that violates traditional UI design principles
- ▶ Over 20 years, numerous guidelines and recommendations has been proposed for HAI in the industries and academia
  - ▶ However, there are still mistakes made in various AI interfaces
  - ▶ Which shows that designers and developers still struggle with creating effective AI-infuse systems
- ▶ A shared guidelines is useful for people to design and evaluate AI-infused systems

# Phase 1: Consolidation of guidelines

- ▶ Guidelines are gathered from three sources:
  - ▶ Review of industry AI products
  - ▶ Recent public articles about AI design
  - ▶ Relevant papers about AI design
- ▶ 168 design guidelines are obtained, which are consolidated into 20
- ▶ The guidelines are organized into four categories based on when during the user's interaction they are applied:
  - ▶ Initially
  - ▶ During interaction
  - ▶ When wrong
  - ▶ Over time

## Phase 2: Modified Heuristic Evaluation

- ▶ 11 team members participated
- ▶ The evaluators examine 13 AI-infused products and try to identify both applications and violations of the guidelines
- ▶ The findings are reviewed, and the number is further reduced to 18

## Phase 3: User study

- ▶ A user study with 49 HCI participants is conducted

### Procedure

- ▶ A heuristic evaluation: each participant is assigned to an AI product and asked to find applications and violations of each guideline

# Phase 3: User study

## Products

- ▶ Products are selected using a maximum-variance sampling strategy:
  - ▶ Top ranking apps, software and websites in the U.S. is searched
  - ▶ Products are grouped by their use case, resulting in 10 categories, 2 product each
  - ▶ Select prominent AI-driven feature to evaluate per product

Product Category	Feature	Participants
E-commerce (Web)	Recommendations	6
Navigation (Mobile)	Route planning	5
Music Recommenders (Mobile)	Recommendations	5
Activity Trackers (Device)	Walking detection and step count	5
Autocomplete (Mobile)	Autocomplete	5
Social Networks (Mobile)	Feed filtering	5
Email (Web)	Importance filtering	5



# Phase 3: User study

## Participants

- ▶ People at large software company with at least 1 year experience in HCI
- ▶ 49 participated
- ▶ 2-3 participants is assigned to each product

## Adjustment and Misinterpretation

- ▶ The responses are reviewed in the cases of:
  - ▶ Duplication (55 instances)
  - ▶ The participant use “Does not apply” to indicate that they cannot find an example of the guideline (73 instances)
  - ▶ The participant use “Does not apply” to indicate a violation (20 instances)

+ke clear The participant misinterpretes one guideline to another

## Phase 3: User Study (Results)

### Clarity and Clarifications

- ▶ Some guidelines are rephrased for more clarity
- ▶ Examples:
  - ▶ G1: “Make capabilities clear” → “Make clear *what* the system can do”
  - ▶ G2: “Set expectations of quality” → “Make clear *how well* the system can do what it can do”

## Phase 3: User Study (Results)

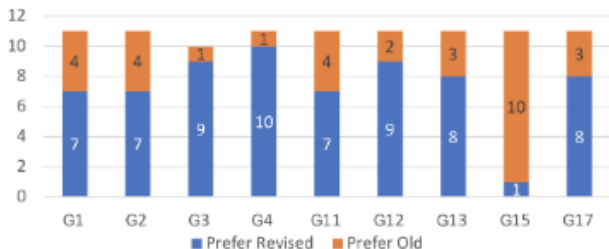
### Evolution of guidelines 1 and 2

<b>Phase 1: Consolidating guidelines</b> Set appropriate expectations. Set accurate expectations to give people a clear idea of what the experience is and isn't capable of doing.
<b>Phase 2: Internal evaluation</b> Set appropriate expectations.
<b>Phase 3: User study</b> G1: Make capabilities clear. Help the user understand what the AI system is capable of doing. G2: Set expectations of quality. Help the user understand what level of performance the AI system is capable of delivering.
<b>Phase 4: Expert evaluation of revisions</b> G1: Make clear what the system can do. Help the user understand what the AI system is capable of doing. G2: Make clear how well the system can do what it can do. Help the user understand how often the AI system may make mistakes.

Table 3: Evolution of Guidelines 1 and 2.

## Phase 4: Expert Evaluation

- ▶ Experts: people who have experience in UX/HCI who are familiar with discount usability methods
- ▶ 11 experts are recruited: 6 UX designers, 3 UX researchers, 2 in research and product planning roles
- ▶ Experts are asked to review 9 revised guidelines and chose what they prefer



**Figure 3: Number of experts out of 11 who preferred the revised or the old version. One participant suggested their own alternative for Guideline 3.**

# Discussion

There is are tradeof between generality and specialization

- ▶ The guidelines might not be able to address all types of AI-infused system
  - ▶ For example, voice-based AI, activity trackers
- ▶ Design guidelines that can be easily evaluated from the interface are focused on.
  - ▶ Ex: Broader principles such as “build trust” is excluded

# Some guidelines

## G1: Make clear what the system can do

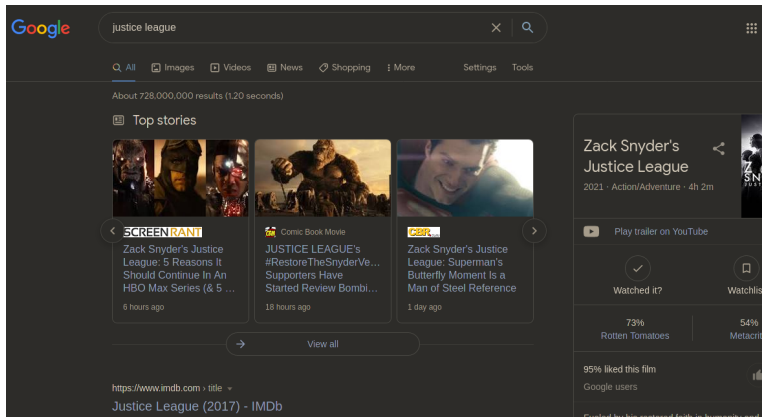
- ▶ Help the user understand what the AI system is capable of doing
- ▶ Category: Initially
- ▶ Example: Activity Trackers
  - ▶ All metrics that it tracks is displayed and explained how



# Some guidelines

## G4: Show contextually relevant information.

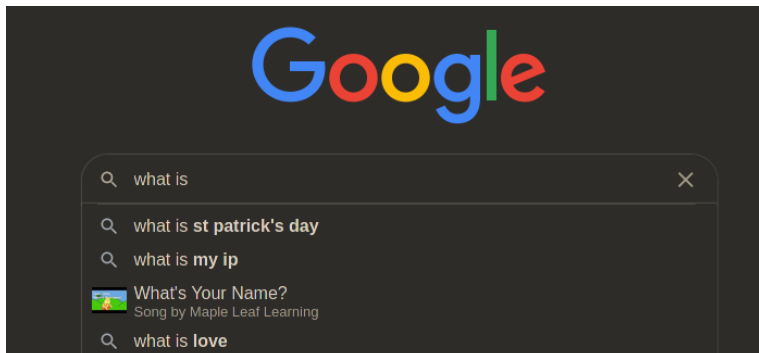
- ▶ Display information relevant to the user's current task and environment
- ▶ Category: During interaction
- ▶ Example: Web Search



# Some guidelines

## G10: Scope when in doubt.

- ▶ Engage in disambiguation or gracefully degrade the AI system's services when uncertain about a user's goal
- ▶ Category: When wrong
- ▶ Example: Autocomplete
  - ▶ Usually 3-4 suggestion is provided instead of directly completing





# Some guidelines

## G13: Learn from the user behaviour.

- ▶ Personalize the user's experience by learning from their actions over time.
- ▶ Category: Overtime
- ▶ Example: Music Recommenders, Video Recommenders

