1-Hour Collaborative Learning Activity for Responsible Human-AI Design

Evan M. Peck
evan.peck@bucknell.edu
Bucknell University

Course Human-Computer Interaction

Programming Language None

Resource Type Lab

CS Concepts User-Centered Design; Design Principles and Heuristics; Design Ethics

Knowledge Unit N/A

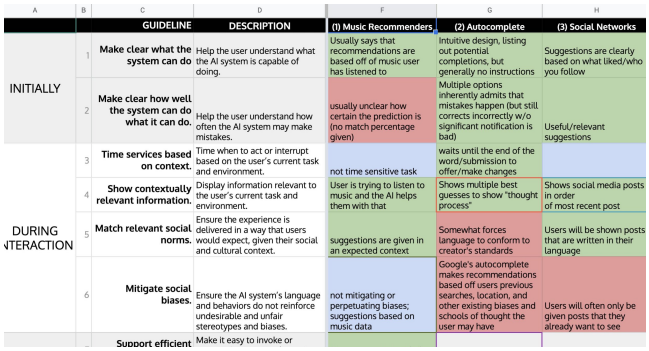
SYNOPSIS

It is challenging to concisely and effectively expose students to the social and practical considerations of designing Human-AI systems. But due to curricular or staffing constraints, Human-Computer Interaction (HCI) is often relegated to a single course (or less!) within CS curriculum, leaving little room for some instructors to integrate applied responsible design into existing CS topics. Still, the societal impact of Human-AI interaction deserves both attention and time. To navigate these tensions, I designed a self-contained activity that considers how responsible Human-AI design can fit into existing course structures. The goal of this 1-hour collaborative learning activity is to (1) give students hand-on experience applying ethical design considerations to Human-AI systems, and (2) be highly portable to fit a variety of contexts and time constraints.

The activity: Students use Microsoft's 18 Human-AI guidelines[1] across four phases (*initially*, *during interaction*, *when wrong*, *over time*) to evaluate real applications [1]. They complete the following tasks (60 minutes):

- Microsoft's Human-AI guidelines are presented and distributed to students via printable cards[2]
- Student groups of 3-5 students are tasked with applying the guidelines to a familiar applications (music recommendations in Spotify, auto-complete in text messaging, Instagram, voice assistants, etc.). Unknown to the students, these applications match those evaluated by experts in Amershi et al [1].
- Within each group, individual students are responsible for categorizing a subset of guidelines as a *violation or clear violation*, *application or clear application*, or *not applicable*.
- Annotations are made within a collaborative spreadsheet (see https://bit.ly/hai-activity) to facilitate virtual classroom environments.

[1]Guidelines: https://www.microsoft.com/en-us/research/project/guidelines-for-human-ai-interaction/
[2]Printable cards: https://www.microsoft.com/en-us/research/uploads/prod/2019/04/AI-Design-guidelines_041519.pdf



		GUIDELINE	DESCRIPTION	(1) Music Recommenders	(2) Autocomplete	(3) Social Networks
INITIALLY	1	Make clear what the system can do	Help the user understand what the AI system is capable of doing.	Usually says that recommendations are based off of music user has listened to	Intuitive design, listing out potential completions, but generally no instructions	Suggestions are clearly based on what liked/who you follow
	2	Make clear how well the system can do what it can do.	Help the user understand how often the AI system may make mistakes.	usually unclear how certain the prediction is (no match percentage given)	Multiple options inherently admits that mistakes happen (but still corrects incorrectly w/o significant notification is bad)	Useful/relevant suggestions
	3	Time services based on context.	Time when to act or interrupt based on the user's current task and environment.	not time sensitive task	waits until the end of the word/submission to offer/make changes	
	4	Show contextually relevant information.	Display information relevant to the user's current task and environment.	User is trying to listen to music and the AI helps them with that	Shows multiple best guesses to show "thought process"	Shows social media posts in order of most recent post
DURING NTERACTION	5	Match relevant social norms.	Ensure the experience is delivered in a way that users would expect, given their social and cultural context.	suggestions are given in an expected context	Somewhat forces language to conform to creator's standards	Users will be shown posts that are written in their language
	6	Mitigate social biases.	Ensure the AI system's language and behaviors do not reinforce undesirable and unfair stereotypes and biases.	not mitigating or perpetuating biases; suggestions based on music data	Google's autocomplete makes recommendations based off users previous searches, location, and other existing biases and schools of thought the user may have	Users will often only be given posts that they already want to see
		Support efficient	Make it easy to invoke or			

Figure 1: An example of students collaborating on evaluating applications using Human-AI design guidelines

- After an individual work period, students come back to their groups to explain, debate, and settle on a full analysis of the group's application.
- (Optional) Pairs of groups that independently evaluated the same application meet to settle on a final categorization of each guidelines (*violation or clear violation*, *application or clear application*, or *not applicable*)
- Coming together as a class, expert analyses from Amershi et al. are revealed and compared with student evaluations [1].
- The instructor guides group discussions about why expert decisions may differ from student evaluations, and whether the guidelines adequately capture the full scope of student experience with the applications.

While relatively simple, the applied nature of this activity yields rich discussions about ambiguities and challenges surrounding the design of Human-AI systems in a relatively short amount of time (a single class period). Extensions of this assignment can be used applied to emphasize corporate responsibility and industry design incentives (see *Recommendations*).

PREPRINT

KEYWORDS

Student Activity, Human-AI Interaction, Ethical Design

1 ENGAGEMENT HIGHLIGHTS

This activity touches on a number of evidence-based learning approaches within a relatively short amount of time.

- **collaborative learning:** individual students use peer-learning to apply guidelines to an application area and then share those guidelines with larger groups. This is reinforced when student groups meet with each other to resolve differences in evaluations

- **using meaningful and relevant context:** students consider and critique AI within the context of applications that they have meaningful interactions with on an everyday basis.
- **culturally relevant pedagogy:** guidelines such as *Mitigate Social Bias* or *Match Relevant Social Norms* necessitate discussions surrounding issues of diversity and social responsibility, and can guide conversation about the sociopolitical role of AI and design.

Potential modifications: There are a number of modifications that can be made depending on the structure of the course or goals of the activity

- *Corporate comparisons:* Google and Apple have also created publicly-available guidelines to navigate machine learning or AI interactions with people [2]. One modification of this assignment would be to assign different guidelines to different groups for the same application (for example, music recommendation is evaluated by Microsoft's guidelines in one group and Google's guidelines in another group). Subsequent group discussions can focus on why evaluations differ between industry guidelines and how those guidelines may nudge design.

2 RECOMMENDATIONS

- **introduction:** I found that giving a light introduction to the guidelines prompted more significant peer-learning opportunities between students. Depending on the background of students, instructors might want to consider more significant initial scaffolding.
- **class size**: This activity has primarily been used in classes that contain between 25-30 students. However, due to the collaborative structure, it should be able to scale to larger (or smaller classes) by changing the number of applications or the number of groups who are devoted to each application. For example, Amershi et al. contains 10 different applications [1]. If two student groups (of 4) are assigned to each application, the structure would still hold for a class of 80.
- **background:** This activity has been run in a Human-Computer Interaction course that targets 3rd and 4th year undergraduate Computer Science students. This background enabled me to move quickly through initial introduction of the guidelines. Students with less background may need slower introduction or a rephrasing of the terminology.
- **individual component:** I would encourage instructors *not* to skip the assignment of roles within groups. I found that having each student within a group be specifically responsible for a subset of guidelines yielded more student participation in both group and class discussions.
- **context:** This activity was originally designed for a remote classroom experience (necessitated by the COVID-19 pandemic), and groups were facilitated using breakout rooms in Zoom. A physical classroom may yield other opportunities for collaboration on shared tables/whiteboards beyond the spreadsheet shared here.

3 ADDITIONAL SECTIONS

3.1 Assessment

This activity was designed to provoke low-stakes engagement with a challenging topic. As a result, it was not formally assessed. Formal assessments could focus on the effectiveness of peer-learning in understanding the basics of each design criteria.

3.2 Limitations and Cautions

This activity works best when the instructor is already aware of the dangers of AI when uncritically launched into diverse communities and cultures. It is framed to be an introductory activity to students who are already engaged with design processes. However, it should be understood that the activity's concise format makes a trade-off with conceptual depth.

One of the primarily dangers of a short activity is that rather than highlighting the tensions and challenges of design (which this should!), it instead gives students false confidence in their evaluations, and nudges towards techno-solutionism. Sometimes, struggles to land on a *"correct"* design exemplifies maturity in understanding the true human complexity of a design problem. That is certainly the case here.

For instructors who have not spent time reading about some of the problems in AI, I might encourage minimally familiarizing themselves with some of the impacts on historically excluded groups (see *Related Online Resources*), and to challenge students during the discussion section to think critically about their decisions.

This activity should *not* be seen as a replacement for teaching students issues of cultural competency and identity. I've found that students with a background in those areas prior to my class often act as a catalyst to some of our richest class discussions (and criticisms) surrounding the design guidelines.

4 RELATED ONLINE RESOURCES

- The project page for Microsoft's *Guidelines for Human-AI Interaction* can be found at https://www.microsoft.com/en-us/research/publication/guidelines-for-human-ai-interaction/. This includes newer resources as well, such as a *Human-AI eXperience* toolkit.
- To make an alternative comparison with Google, you can also reference Google's People + AI Guidebook Patterns https://pair.withgoogle.com/guidebook/patterns
- AI's impact on historically excluded groups (focusing particularly on race) has been covered by many scholars in recent years, including Ruha Benjamin, Sofiya Noble, and Joy Buolamwini (among others!). For instructors unfamiliar with this space but hoping to augment conversations in class, the documentary *Coded Bias* offers a good starting point: https://www.codedbias.com/

5 MATERIALS

The following files are contained in the `.zip` folder:

- `human-ai-guideline-cards.pdf` - a PDF of the cards originally found at https://bit.ly/hai-activity that should be distributed to students.
- `human-ai-activity-spreadsheet.pdf` - a PDF version of the spreadsheet that students collaborate on to evaluate different systems.
- `human-ai-instructor-guide.pdf` - a detailed explanation of how the activity is run, along with discussion prompts.
- `guidelines-for-human-ai-interaction.pdf` - a PDF of the research paper by Amershi et al. [1] that introduced the human-ai guidelines. Expert evaluations that are compared with student evaluations are also in this paper.

REFERENCES

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[2] Austin P Wright, Zijie J Wang, Haekyu Park, Grace Guo, Fabian Sperrle, Mennatallah El-Assady, Alex Endert, Daniel Keim, and Duen Horng Chau. 2020. A Comparative Analysis of Industry Human-AI Interaction Guidelines. *arXiv preprint arXiv:2010.11761* (2020).