Journal of Usability Studies
Vol. 14, Issue 4, August 2019 pp. 181–186

Intentionally Ethical AI Experiences

Carol Smith
Sr. Research Scientist - Human Machine Interaction
Carnegie Mellon University, Software Engineering Institute
Pittsburgh, PA, USA
carologic@gmail.com

Books, movies, and culture, in general, have elevated artificial intelligence (AI) as the answer to everything, while simultaneously training generations of us to dread AI. Recent events have brought some of our worst technology fears to light. Those now infamous AI systems likely began with poor or heavily biased content, inadequate training, and most significantly a lack of mitigation planning and maintenance. AI systems are as imperfect as the humans making them and retain the biases, the inconsistencies, and the flaws inherent in the initial data provided to them. Looking for patterns in flawed data magnifies the flaws and identifies patterns that may not be valid. These poorly designed systems can create dangerous situations for humans.

AI must be designed responsibly and intentionally. AI systems require humans to create data that supports and trains the AI, and then to manage and supervise the AI's progress. These systems require weeks to months to build, and that work must be done in a thoughtful, purposeful way. The resulting system should be at least as accurate (if not more) than a colleague doing the specific job or task the AI is trained to do.

Machine learning (ML) is used to describe the collecting of large amounts of data (for example, a city's parking ticket data including vehicle information, locations, and more), the creation of algorithms (a set of rules), and the training of systems to create an AI based on patterns and inference instead of explicit commands. Machine learning and artificial intelligence are frequently used synonymously, though ML is a subset of AI.

In this essay, I make the assumption that the situations being considered include a specific type of person who has a specific problem in mind, and that there has been some examination into whether an AI system is the right solution for this problem (and it is).

Given these assumptions, the systems we make **should be less biased** than our colleagues and still clearly communicate the inherent bias that all systems (and people) have. To do this, as UX professionals, we must take additional responsibility at three key decision points in the creation of AI systems (at least for now).

The three key decision points are (a) Content and Curation, (b) Training, and (c) Maintenance. None of these are about the UI. They are about the experience with the system, and each of these decision points makes a significant impact on the experience. These decision points do not happen one at a time in order, but rather, are continuous throughout the life of an AI system. Decisions in one area will affect the other decisions made. This is true for whatever type of AI you are creating.

Content and Curation

AI systems require a massive amount of data (the terms content and data are used interchangeably in this essay). Often organizations decide they want an AI solution ("We need an AI for that!") before determining if they have access to the amount of content required to create the AI system they desire. Taking a step back and asking critical questions about the content is our responsibility so that people are able to get the accurate information they need from the system.

AI systems can become more accurate with more consistent data on specific topics. AI systems only know what you tell them and can only learn a very narrow topic area. Due to a variety of constraints, they cannot currently contain the knowledge for even the relatively narrow breadth of a pre-school education. Everything you share with an AI system becomes salient and influential to the AI and needs to be prepared for that purpose.

The quality of the data and trustworthiness of its authors must also be considered. What do we know about the provenance of the data? Will the intended audience recognize the source and the authors' credentials as reliable? Do the authors come from diverse enough backgrounds that multiple points of view are considered (education, industry, employer, area of focus, process, etc.)? Is that important? What is missing from the data? Does the data have a cohesive point of view? All data is biased—how is this data biased?

Bias. Your perception of everything around you is biased by your experiences—by the resource availability you had as a child, your social class, race, gender, sexuality, culture, theology, tradition, and other experiences you have had. We are each unique in our experiences and who we are, but often, where we live, who we live with, and what we do affects us much more than we can objectively be aware of. For example, until recently in the US it was rare for young children who were asked to draw a scientist, to draw a representation of a woman. It isn't that no women were scientists or that these children were sexist, rather they grew up in a society where the norm for a scientist was a man.

In 2018 Reuters reported about Amazon's automated employment screening AI system that "taught itself that male candidates were preferable. It penalized resumes that included the word 'women's,' as in 'women's chess club captain.' And it downgraded graduates of two all-women's colleges" (Dastin, 2018, para 7). Amazon eventually lost hope and shut down the project, disbanding the team.

This is not the only time that a company has had a biased set of data. There have been multiple examples where AIs created to recognize faces, having only been provided with sets of faces with low levels of melatonin, once trained, could not recognize darker faces (or tattooed skin). A lack of diversity in the original data set will create a biased (racist) AI.

The cause may seem innocent—the developer (statistically, a white male) looked for a data set. He found a data set online that claimed to be well regarded, reviewed a sample of images and they seemed fine, and selected the dataset. He may have even shown it to others on the team. It is likely that no one on the team had a priority to check for diversity in the data, and (even

more likely) the team was not diverse themselves. Their inherent bias is to see white people as "people" and not necessarily notice a lack of racial or ethnic diversity.

Bias isn't the only concern with data sets. For example, if the content turns out to be contentious, is not uniform, lacks consistency, or raises concerns that may create distrust, it is probably not worth the effort of creating an AI. As previously mentioned, current AI systems are only able to learn very narrow areas of content, and if there is disagreement about how humans make decisions or describe that work, the AI will not do a better job.

Training

Once the data is selected, a model can be created through the selected training method. Sadly, this is often done in the opposite order, where algorithms (programmatic rules) are created before the data is gathered. In these situations, where data is an afterthought, the training and resulting model are unlikely to meet the needs of the problem. Humans must be involved in this process in order to ensure that the training is done accurately and that the model created meets a human's need. While the training is not often visible to the UX team members, it will be incredibly important to the people using the system.

There are different methods of training and each of them have risks and benefits. Being careful about selecting the training that will be most beneficial to the work is important. Our work is to understand the needs of people and ask pertinent questions about those decisions to best advocate for their needs. Most UX professionals will not necessarily have the knowledge to differentiate between algorithms or training, but being able to ask well-formed questions will help the ML specialists make the right decisions.

There are two primary types of ML: supervised and unsupervised learning. Supervised learning involves carefully vetted specialists and subject matter experts (SMEs) spending the time to classify and categorize the content into computational taxonomies (ground truth) to share with the system. The system then uses this information as an input and uses feedback about the accuracy (regression) to further inform the model that is created.

To translate the ground truth for the AI, experts will either need to (a) pair with an ML specialist or (b) use a GUI system to create the ground truth themselves. An ML specialist can be more resourceful than a GUI system, but the process can be difficult to manage due to scheduling conflicts. A GUI is much faster and enables the expert to directly teach the system, but may not be flexible and/or complex enough for the needs of the model.

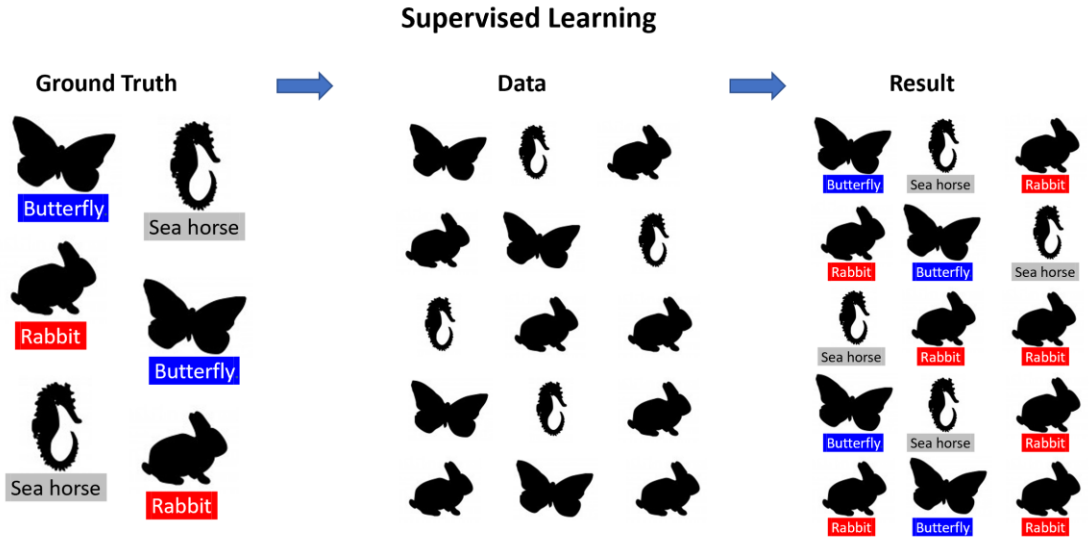

Figure 1. Simplified example of successful supervised learning.

Unsupervised learning is the process of taking the data provided and having the system cluster and narrow the information. The AI system uses forms of association analysis to look for patterns in the data. Assuming the system finds patterns in the data, those patterns may or may not have previously been seen by humans, and they may or may not be helpful.

Unsupervised learning is used in cases where exploration of the data is desired or where the data needs to be reduced and grouped. While "unsupervised" implies no one is watching, it's a misnomer, as a human must still review the work and determine if the model that results is good. For both supervised and unsupervised learning, when there is complex subject matter, the subject matter experts need to make a commitment to be available in order to ensure the model created is valid.

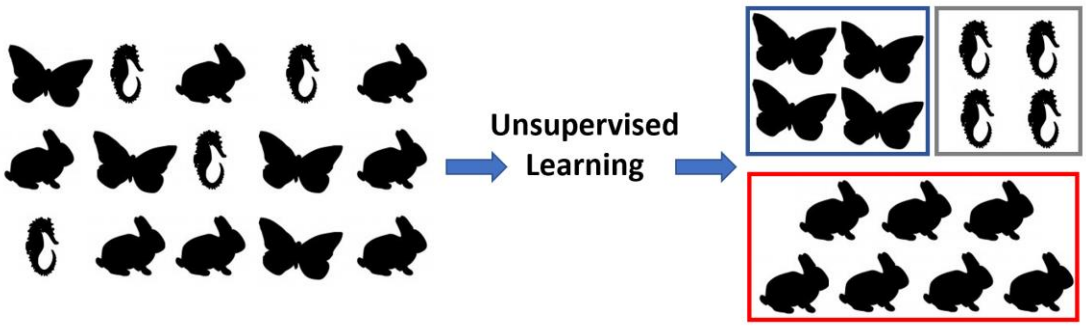

Figure 2. Simplified example of successful unsupervised learning.

In every AI system, we must continuously consider the implications on humans and their situation. There are tools to determine accuracy, but they do not work in every situation, and depending on the needs for the model, may not be helpful. Can the experts tell if it is sufficiently accurate? How? Understanding what the needs are of the people using the system to define a good model is key to helping the team understand their constraints.

It is sometimes a matter of realizing a problem is extremely complex and there may not be a simple solution. Self-driving is one of those very complex spaces. Another is image recognition. While progress has been made on identifying eye diseases with image recognition (Vincent, 2018), other situations, such as Google's Photos, relying on a racist algorithm that identified dark skinned people as gorillas (Simonite, 2018), continue to be frustrating with no real answers available.

With any of this work, there is risk of building the wrong thing. As the work progresses over weeks and months, it is possible that the area of focus will change and rework may be needed. Minor inconsistencies can create significant frustrations and potentially result in a fragile model. For example, if you were creating an AI to distinguish between cereals, in the time it takes to make the model, a new cereal type, a cereal rebranding, or change in ingredients may occur, potentially significantly affecting the already completed work. Whatever is created will only be as good as the data and the time spent making and improving the model.

Maintenance

Maintenance of the system is where the most UX related thinking needs to be done, and where the least effort is typically focused. For example, an AI system must be examined from an ethical point of view for potential harm to humans, particularly to disadvantaged groups and for opportunities for misuse. How might this system cause harm to humans? It is rare that a system cannot do any harm, so going through the worst-case scenarios at a high level, prior to starting work, will help the team consider what mitigation strategies are needed. This is uncomfortable, unpleasant work and many will want to shy away from it. This is where your expertise and advocacy for people becomes priceless.

UX teams need to take responsibility initially (and perhaps long-term) to ensure that this work gets done and that people are kept safe. Creating a Code of Ethics may be helpful to ensure there are clear guidelines with regard to what your organization values, your position with regard to helping people (e.g., humans will always be responsible for making decisions regarding human life), what lines your AI won't cross (no human harm), and how you will track your progress. There are many sources that can be used to model these types of codes, including UXPA's Code of Professional Conduct (https://uxpa.org/resources/uxpa-code-professional-conduct), the ACM Code of Ethics and Professional Conduct (https://www.acm.org/code-of-ethics), and many corporations have created their own versions including IBM and Microsoft.

Monitoring the system is a major consideration for any organization taking on the creation of an AI. It seems many organizations assume they can do the work to set it up and then it will auto-magically work. That is never the case. **AI systems must be constantly monitored**—particularly when new data is introduced to the system—it is extremely important to corroborate that the AI is doing what it is supposed to do with that data. What warnings should be made apparent? How can humans identify issues? Can plain language indications be used? Using undesirable scenarios as models to help everyone identify issues and improve the system may be helpful. Who can they report those to once identified? When and how is undesirable data expunged from the system?

Once we recognize the potential negative effects our AI can have on people and society, how do we ensure that humans can turn it off in a timely and safe manner? For example, in 2016 Microsoft created a chatbot, Tay, that would chat with people on Twitter. Unfortunately, it seems the team did not realize that there would be bad actors who would train Tay to hurt people (Vincent, 2016). By not anticipating this bad behavior, they failed to create a mitigation plan and the Internet watched in awe as the chatbot tweeted racist, bigoted, and mean words to the world. We don't know much about the team, but certainly having a more diverse group that included people who are typically marginalized, might have resulted in anticipating undesirable behavior and then creating good mitigation strategies.

As Grady Booch famously said we must "ensure humans can unplug the machines." This isn't a game, nor a sci-fi movie. **Real information, real lives are at stake and we must remain in control.** When systems are not acting as expected, when people involved in its creation are unsure why things are happening, it is time to shut it down. These are not sentient or unknowable systems, and while it may be difficult to determine the root of decision making, it is not impossible. Once the system is shut off, who needs to be notified about this unfortunate event? What are the unintended consequences of turning off (who loses access, who loses service)? These issues will continue to be controlled by organizations for the near future and may become a point for government regulation.

Intentionally Ethical AI

AI systems are not able to solve every problem, nor should they. There are areas of life where AI's should either not be used or used only with extreme caution, such as judicial decisions and complex medical care. These are areas where context matters and a formula answer may not suffice. AI systems need a village of diverse people asking questions so that all potential outcomes are considered.

The future of AI is still fuzzy. As UX professionals become more comfortable with the technology and ask difficult questions, we can reduce the situations that put humans in harm's way. Ensuring that an AI meets the needs of the people using it, is accurate, consistent (as appropriate), and has a high level of overall information quality will make it effective. Providing people with information regarding content provenance, training methodology, and system expectations will provide the transparency needed to engender trust and increase engagement with the AI.

I have great hope that AI will become an extremely useful and helpful technology to humans. If you aren't familiar with AI, partner with someone and explore the tools. By asking hard questions about the provenance of content, the training the system undergoes, and ensuring that the AIs are managed in responsible, ethical, transparent, and fair ways, we can help keep humans safe.

References

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

Simonite, T. (2018). When it comes to gorillas, Google Photos remains blind. *Wired*. Retrieved from https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/.

Vincent, J. (2016). Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. *The Verge*. Retrieved from https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.

Vincent, J. (2018). DeepMind's AI can detect over 50 eye diseases as accurately as a doctor. *The Verge*. Retrieved from https://www.theverge.com/2018/8/13/17670156/deepmind-ai-eye-disease-doctor-moorfields.

About the Author

Carol Smith

Ms. Smith is a Senior Research Scientist in Human-Machine Interaction at Carnegie Mellon University's Software Engineering Institute and an adjunct instructor for CMU's HCII program. She has been conducting UX research to improve the human experience across industries for 18 years and working to improve AI systems since 2015. She has served two terms on the UXPA international board and is currently an Editor for the Journal of Usability Studies. She holds an M.S. in Human-Computer Interaction from DePaul University.