Explaining AI to Laypeople through New Metaphors from the Arts

Lay Summary

My project brings together **artists**, **writers, researchers** and **everyday people** in order to develop new **metaphors** for Artificial Intelligence. AI is becoming a bigger and bigger part of our daily lives. It effects hiring decisions, what information we see on social media, and how banks assess our credit scores. Now more than ever, it’s important for all of us to understand how these AI systems work.

The problem is, AI systems are complicated and their decisions are difficult to understand. This is where **interface metaphors** are crucial. These metaphors enable people to understand complex technological concepts using intuitions from their daily lives. For instance, the interface metaphors of “folders” and “trashcans” on computers are standins for more complicated Computer Science topics like “directories” and “disk reformatting.” In the 1980s these metaphors were crucial to bring computers to a broader audience. We need similar metaphors for AI to make this technology more open and accessible.

Artists and **writers** can help us find new metaphors for AI because they are constantly coming up with new representations and metaphors for the world. In fact, one of the best metaphors for AI recently came from SciFi author Ted Chiang in a viral *New Yorker* article, in which he compared popular chatbot ChatGPT to a fuzzy image of the Internet. Through **interviews**, **large-scale surveys**, and a **workshop with artists and writers**, this project will find more metaphors that will give everyday people a working understanding of AI.

Description of Project

In this project, I will facilitate more intuitive and interpretable **Artificial Intelligence design for the public** through **community arts workshops**, **ethnography**, and **AI system design**. In doing so, I will collaborate with professors from 4 departments (Computer Science, History, Literature, and Environmental Science) and work with at least 2 Toronto community arts organizations (Coach House and Inter/Access). By the end of this project, I will have produced **2 creative-non fiction essays** for one of the international publications I write for (e.g. *LitHub, Los Angeles Review of Books)*, submitted **2 academic papers** to top-tier Human-Computer Interaction conferences, and have hosted a **gallery exhibit and literary reading for the public**.

This project addresses the key need for new **metaphors** for AI. There has been recent public outcry about the ways in which, unbeknownst to us, AI is affecting the ways in which we consume information and make decisions about hiring and credit ratings. One of the reasons AI hasn’t been subject to more legal and ethical scrutiny until recently is that AI is not capable of explaining its decision making. To address this, CS has largely focused its efforts on ‘**Explainable AI’** which attempts to provide explanations about AI decisions to users. However, these explanations are often too complex for laypeople to understand. To address this, my research seeks instead to develop **metaphors** that provide users with **“mental models”** for AI systems that give an approximate, working understanding of AI. These mental models should help the public engage in conversations about artificial intelligence and pave the way for more **ethical design** of AI.

Metaphors hold an important place in the history of Human-Computer Interaction. The visual “desktop” metaphor enabled people to understand the computer as a physical work table, complete with “files,” “file folders,” and a “trash,” when in reality these are stand-ins for more complex computer science topics like “directories” and “disk reformatting.” Recently, the science-fiction author Ted Chiang authored a widely publicized article in *The New Yorker* stating that the Internet is a reasonable metaphor for **ChatGPT**: it possesses an incredible breadth of knowledge, but is also laden with inaccuracies, follies, and biases. This metaphor doesn’t fully capture how ChatGPT works, but reasonably helps the public develop a mental model for the system.

Artists and writers are uniquely poised to help us devise new metaphors because they are constantly coming up with visual and linguistic representations to think through human experience. Similar Human-Computer Interaction papers have already invited artists to help develop new metaphors for climate change (see: *The Disaster and Climate Change Artathon*). In this project, I will host a workshop with Toronto artists and writers to explore new metaphors for AI. This project will use several common Human-Computer Interaction research methods 1) **Ethnographic** field work with laypeople and policymakers in order to understand their intuitions about AI, 2) **Experimental workshops** with Toronto artists and writers that explore new metaphors for AI, and 3) **System design** of AI which puts some of these new metaphors into practice.

Description of public impact of work, expected outcomes/learning, and future directions

One of the most pressing social questions of today is how the public will deal with **Artificial Intelligence**. There has been recent public outcry about the ways in which AI is affecting the ways in which we consume information and make decisions. The Biden Administration recently released a **“Blue Print for an AI Bill of Rights”** to address the challenges posed by AI systems that **“threaten the rights of the American public,”** noting that AI can affect hiring and credit decisions. My work addresses the need for more public policy for AI by helping laypeople develop new **“mental models”** and **metaphors** to understand AI systems. This work will empower laypeople to have a voice in today’s crucial discussions about AI and paves the way for more appropriate public policy.

By coming up with better metaphors to articulate AI to users, this project will also enable software developers to **better design AI systems** for **positive social impact**. While AI currently facilitates cancer diagnosis, helps avoid led poisoning, and suppresses hate speech, many remain wary of implementing AI because they are unsure it can be trusted. With appropriate metaphors, we can reach these skeptical populations and engage in conversations about how AI may or may not be able to help.

This research involves an exciting **transdisciplinary synthesis** of **Visual Arts, Literary Arts**, **Human-Computer Interaction**, and **Ethnography**. A major component of this project is conducting an artists' workshop with legendary Toronto literary indie press **Coach House** and local digital media arts organization **Inter/Access** to devise new metaphors. As a published novelist with a novel from prominent literary press **Dalkey Archive**, a musician with 5 albums out on **75OrLessRecords**, and a media artist with awards from **HackPrinceton** and **HackCooper**, I am uniquely poised to take on this work. In addition, a previous paper of mine—*Fostering Interdisciplinary Exchange in STEM–Artist Collaborations*—well prepared me to design this collaboration.

I intend to publicize this work in several ways. First, the workshop with artists will culminate in a **gallery exhibit and literary reading open to the public**. Here, the public will be given the opportunity to think through AI with sight, sound, and touch. I will then write **2 essays** about this work. The first will be published in one of the internationally distributed art journals I write for, like ***Los Angeles Review of Books, LitHub,*** **or** ***Hazlitt*** and focus on the gallery exhibit. The second will focus on the AI system I build based on these metaphors and will be published in a popular technology magazine like ***Wired*** or ***MIT Tech Review***. I will also write **2 academic articles** for publication at a **top-tier Human-Computer Interaction conference**. Lastly, after the fellowship, I intend to **adapt my dissertation** into a book about the future of Human-AI interaction that I can **hand to my agent** for publication.

Description of work during the one-year timeline of fellowship, including work on project and proposed forms of public engagement

September-November

Phase 1: Interviews with Laypeople

In this stage, I will use ethnographic techniques to interview laypeople about their existing intuitions for AI. I will actively seek out people from underrepresented communities who do not have a professional understanding of AI.

December

Phase 2: Broad survey study

In this stage of the project, I will gain a broader sense of existing intuitions laypeople have for AI by conducting large-scale surveys based on the results of the interviews.

January-April

Phase 3: Artist Workshop

I will host 1-2 workshops with local Toronto digital media artists and literary writers in which they will attend lectures about AI, work with AI systems, and develop their own intuition for these systems. They will then be asked to develop artworks and stories which develop new intuitions and metaphors for AI. These artworks will be displayed at a local gallery (potentially Hart House or Inter/Access) with a literary reading and opening event to the public. I will then publish an article about this work in *LitHub, Los Angeles Review of Books,* or one of the other major publishing venues for which I am a contributor.

May

Phase 4: Design Considerations

Based on this workshop and the results from the interviews and surveys, I will develop a series of design considerations and new metaphors for the development of AI systems.

June-August

Phase 5: Building a New AI System

Finally, I will build a new AI system based on some on one or several of the metaphors that were found during our analysis. I will make this system available for public consumption and run a small scale user study with members of the UofT community. I will publish an essay about these metaphors and the system in a public tech publication like *Wired, MIT Tech Review,* or *Motherboard*.

Equity, Diversity, and Inclusion

This work is designed to promote a more **inclusive conversation** around AI which **invites traditionally underrepresented groups in STEM**. For example, while persons of certain socioeconomic backgrounds may not have access to the educational and financial resources to pursue a deeper understanding of AI, this project will give these groups a working intuition for the technology so that they can decry AI when appropriate. Other underrepresented groups in STEM like **women, queer people, and people of colour** have also thus far been largely left out of the conversation around AI. By inviting members of these traditionally underrepresented communities to explore AI with their own intuitions, we invite them to consider how AI can be described with the language of their communities rather than the language of STEM. In order to achieve this, I and the research team will 1) **actively seek out members of underrepresented populations** to interview in the ethnographic study, 2) actively seek out members of underrepresented populations to join the artist workshop and 3) make clear to both interviewees and the artist workshop that we are interested in how participants can use existing intuitions from their **cultural backgrounds** to describe AI.

Further, the research team and list of collaborators for this project are diverse in terms of discipline, race, and gender. I am White Hispanic, and have consciously included other minorities in the project team. Four different departments are featured in the project team: **Computer Science, Environmental Science, English, and History of Science**. This project also provides a unique synthesis between STEM and the arts. The local community arts organizations with which we will work, Coach House Press and Inter/Access, have a long history of working with and supporting queer artists, Indigenous artists, and artists of colour, and will help ensure the inclusivity of the artist workshop.

Project Collaborations

- Daniel Wigdor, Computer Science
- Robert Soden, Environmental Science and Computer Science
- Ishtiaque Ahmed, Computer Science (focus on ethics of AI in Global South)
- Adam Holland, English Literature
- Jean-Olivier Richard, History of Science
- Tristan Sauer, Inter/Access (local digital media arts org)
- Alana Wilcox, Coach House Press (famed local literary press)