Artificial Intimacy
An Exploration of the Personal and Intimate in Natural Language Processing Models

Mirabelle Jones
University of Copenhagen
msd@di.ku.dk

Irina Shklovski
University of Copenhagen
ias@di.ku.dk

Nastasia Griffioen
University of Twente
n.griffioen@utwente.nl

Obaida Hanteer
University of Copenhagen
obaida.hanteer@hum.ku.dk

ABSTRACT
Artificial Intimacy is an AI art installation that explores what natural language processing (NLP) models in our everyday lives would feel like if they were to be personalized to match our own personalities and values. We explored the possibility of fine-tuning NLP models using personal social media data. Our selected data sources—Leslie Foster and Gorjeoux Moon—have offered their own social media data to fine-tune the models. We present a video capturing their conversations with their social media selves. The interactive portion of the installation invites the audience to engage with Foster’s and Moon’s chatbots and explore interactions with NLP models that are personalized in this way.

CCS CONCEPTS
• Empirical Studies in HCI, • Empirical Studies in Collaborative and Social Computing;

KEYWORDS
NLP, Model finetuning, Values, Artificial Intelligence, Personalization

ACM Reference Format:

1 INTRODUCTION
Natural language processing (NLP) models are used to analyze spoken and written human language, and output human language in ways that would make sense to end-users. NLP based systems find their way into medical settings (e.g., hospital patient check-ups), commercial settings (e.g., consumer service chatbots), and our private home lives (e.g., home assistants such as Alexa) to great effect. Yet there have been many examples of NLP systems exhibiting problematic—and sometimes even dangerous—behaviors. From Microsoft’s unfortunate Twitter bot Tay [1] to Amazon’s unsuccessful job screening tool [2], NLP-based systems demonstrate output marked by biases and violations of social norms. Large datasets on which NLP models are typically trained are at the core of these problems (e.g., [3]) as these data encode cultural biases and reflect under-representation of minorities, impacting the resulting models and their behaviors [4], [5]. In response to repeated public embarrassments and scandals, the NLP community has sought a range of solutions, some focusing on fine-tuning models with selected data to mitigate general training problems. For instance, OpenAI has made efforts to finetune their NLP models GPT-2 and GPT-3 to be more value-sensitive using “values-targeted datasets” [6] in response to problems that emerged after the release of these models to the general public [7].

Artificial Intimacy offers the audience an opportunity to engage with two fine-tuned chatbots, exploring what happens when such fine-tuning occurs on social media data from particular individuals. The goal of the installation is to question what kind of experience can fine-tuning NLP models achieve and how might we experience the world populated by NLP-based systems that can become intimate portraits of the diverse selves.

2 DESIGN AND IMPLEMENTATION
Artificial Intimacy is composed of a sculptural installation, a chatbot system, and videos of two people chatting with “their” fine-tuned chatbots (see Figure 1 for installation view). The sculpture housing that represents the chatbots is 3D printed in a two-toned glossy filament and houses a combination microphone and speaker, which constitute the hardware for the chatbots. The sculpture’s shape was designed based on the form of diatoms, a type of plankton whose disk-shaped, perforated attributes closely resemble the rounded, perforated structure of PIA’s such as Google Home and Alexa. The resulting appearance is meant to resemble traditional commercial products such as Google Home and Alexa while maintaining an extraterrestrial aesthetic. The chatbots were designed first by creating the datasets for each bot. Our chat-bot models Leslie Foster and Gorjeoux Moon were selected for their non-majoritive identities and non-normative values. The first author used Foster and Moon’s Facebook messenger history to create a corpus of prompts and responses to fine-tune GPT-3. Both Foster and Moon then engaged in conversation with their bots for several weeks and these conversations were recorded. To produce the videos for the project, Foster and Moon selected questions and responses they found resonant.
Mirabelle Jones et al.

Figure 1: Installation view of Artificial Intimacy with interactive voice assistant sculpture, video screens with headphones, decorative wall hangings.

The chatbots, once approved by Foster and Moon, were transitioned through programming in Python into Personal Interactive Assistants (PIAs) for use with the sculpture.

**Interactive component:** we invite visitors to engage in conversation with Foster and Moon’s bots. Questions can be asked directly into the sculpture and responses will be almost instant. The bots can be asked a broad variety of questions, providing responses based on value-inclusive models. After the interaction, visitors will be asked to reflect on their experience.

3 IMPLICATIONS

Artificial Intimacy builds on an emergent artistic practice of challenging AI systems through design [8]–[10]. The installation was designed and created to prompt the audience to consider the experience of personalized, value-embedded AI models. While social media are indeed an obvious online space for people to share texts that would express their values, the specific values brought forward by the fine-tuned models prompt us to consider the different contexts and identity aspects that may be expressed by people even within one social medium. Fine-tuning OpenAI’s GPT-3 model on personal social media data can produce models that are at least in part in keeping with the values and core beliefs of our two subjects. Yet these models still resorted to gibberish at times resulting in inconsistent experiences. When we consider what interacting with an AI version of the self might mean in the real world, we are forced to think about whether such personalization of AI models is a viable approach.

ACKNOWLEDGMENTS

We thank Leslie Foster and Gorjeoux Moon for their willingness to share their social media data and their intimate NLP models with audiences. Miguel Sicart and Anders Søgaard gave valuable advice and input. Sarah Newman, Kim Albrecht and metaLAB Harvard offered collegial debate, space and support for initial explorations.

REFERENCES


