My research involves the heavy use of machine learning and natural language processing in novel ways to interpret big data, develop privacy and security attacks, and gain insights about humans and society through these methods. I do not use machine learning only as a tool but I also analyze machine learning models' internal representations to investigate how the artificial intelligence perceives the world. This work [4] has been recently featured in Science where I showed that societal bias exists at the construct level of the machine learning models, namely semantic space word embeddings which are dictionaries for machines to understand language. When I use machine learning as a tool to uncover privacy and security problems, I characterize and quantify human behavior in language, including programming languages by coming up with a linguistic fingerprint for each individual. The methods I developed in this realm [1] are being used by the Federal Bureau of Investigation to identify so called doppelängers to link the accounts that belong to the same identities across platforms, especially underground forums that are business platforms for cryber criminals. By analyzing machine learning models' internal representation and linguistic human fingerprints, I am able to uncover facts about the world, society and the use of language. The combination of these novel research methods, scientific findings, and insights about humans and society place me at a unique spot. My work that is grounded in computer science also draws on computational social science, human computer interaction, and behavioral economics, and has applications to public policy.

My work builds upon the fundamental elements of machine learning by extracting features from natural language and joint language vision models. I have been inspired to investigate the construct level of machine learning models after being able to uncover new facts about authors and programmers by processing big data. If machine learning provides ways to analyze big data in a short amount of time by detecting significant statistical patterns about human behavior, the various models generated by very large scale human data should intuitively be imbibing facts about the world and culture in aggregate. In recent years, the new research area of fairness, accountability, and transparency in machine learning has been a subject of debate with speculations of unfairness and problems in interpretability. I came up with a method to analyze state-of-the-art word embeddings in the semantic space that has been trained on billions of sentences from the web. My results proved that all types of societal bias and stereotypes exist at the construct level of language models that form the foundations of any tool to perform text related tasks on machines and on the Internet. Word embeddings are numerical vectors for machines to understand language's semantics, syntax, and word relations via mathematical vector operations. Accordingly, these embeddings are used for tasks such as web search, text generation, machine translation, web page rank, automated speech generation, sentiment analysis, and named entity recognition. I introduced a new way of analyzing machine learning models that enhanced the interpretability and transparency of the models. My ongoing research on the subject focuses on multi-modal visual and language embeddings to understand bias in human and computer vision. This emerging area of research requires a long term research program to uncover fairness and transparency problems introduced by artificial and also natural intelligence so that we can find ways to deal with bias and unfairness at the mathematical, ethical, and policy level.

Machine learning aids humans by finding statical patterns in large amounts of data which is not a feasible task for expert humans to perform by themselves in a reasonable amount of time. By extracting linguistic features from natural language or programming language texts of humans, I show that humans have unique linguistic fingerprints since they all learn language on an individual basis. Based on this finding, I can de-anonymize humans that have written a certain text, source code, or even executable binaries of compiled code. This is a serious privacy threat for individuals that would

aylinc@princeton.edu
www.princeton.edu/~aylinc

like to remain anonymous, such as activists, programmers in oppressed regimes, or malware authors. Nevertheless, being able to identify authors of malicious code enhances security. On the other hand, identifying authors can be used to resolve copyright disputes or detect plagiarism. Accordingly, by using machine learning, I am able to develop privacy infringing and security enhancing methods. Once I show what exactly causes the privacy problem, I can tweak the machine learning methods for example to anonymize pieces of text as a countermeasure. My research in this area started with authorship attribution of translated text and micro-text in underground forums and then I showed ways to anonymize writing [5, 1, 2] as a privacy enhancing countermeasure. Afterwards, I started analyzing artificial languages to de-anonymize programmers from their source code and executable binaries [3, 6]. The next and most challenging task is de-anonymizing malware authors by automatically reverse engineering malware that might be encrypted or packed, and extracting stylistic features again via machine learning. I have received a two year grant from DARPA to work on this challenging problem which is of great importance to governments, corporations, and individuals.

Machine learning has been a powerful tool in my research and it has also been subject of my research itself. There remains a long term research agenda to systematically understand how machine learning and its applications affect the world to find ways for dealing with bias embedded in models and bringing more transparency and interpretability to machine learning. Some questions I am currently trying to answer are do different types of grammar of different languages have certain effects on bias in machine learning models or how does bias in machine learning evolve? Does artificial intelligence perpetuate stereotypes and how can we uncover nontrivial properties of machine learning to maximize its benefit to society and minimize the harm it might cause.

[1] Sadia Afroz, **Aylin Caliskan-Islam**, Ariel Stolerman, Rachel Greenstadt, and Damon McCoy. Doppelgänger finder: Taking stylometry to the underground. In *IEEE Symposium on Security and Privacy*, 2014.

[2] Andrew WE McDonald, Sadia Afroz, **Aylin Caliskan**, Ariel Stolerman, and Rachel Greenstadt. Use fewer instances of the letter i: Toward writing style anonymization. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 299–318. Springer Berlin Heidelberg, 2012.

[3] **A. Caliskan-Islam**, R. Harang, A. Li, A. Narayanan, C. Voss, F. Yamaguchi, and R. Greenstadt. De-anonymizing programmers via code stylometry. In *24th Usenix Security Symposium*, 2015.

[4] **Aylin Caliskan**, Joanna Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[5] **Aylin Caliskan** and Rachel Greenstadt. Translate once, translate twice, translate thrice and attribute: Identifying authors and machine translation tools in translated text. In *IEEE Sixth International Conference on Semantic Computing (ICSC)*, pages 121–125. IEEE, 2012.

[6] **Aylin Caliskan-Islam**, Fabian Yamaguchi, Edwin Dauber, Richard Harang, Konrad Rieck, Rachel Greenstadt, and Arvind Narayanan. When coding style survives compilation: De-anonymizing programmers from executable binaries. *arXiv preprint arXiv:1512.08546*, 2015.

aylinc@princeton.edu
www.princeton.edu/~aylinc