DATA BIAS CONSIDERATIONS FOR ARTIFICIAL INTELLIGENCE SYSTEMS IN HIGHER EDUCATION

A Research Paper submitted to the Department of Engineering and Society
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By

Cristian Scruggs

March 28, 2022

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISOR
Catherine D. Baritaud, Department of Engineering and Society

Adoption of Artificial Intelligence (AI) is rapidly increasing; however, results of AI systems are deemed more important than the data used to produce said results (Olson, 2021, para. 13). This proliferation coincides with an increasing reliance on web applications in higher education (Khalid et. al., 2012, p. 2). Many of these apps include the most state-of-the-art features, with AI enhanced ones likely not far behind.

The technical project, Satori, is a modern web application for handling the office hours queue for CS 2150 at the University of Virginia (UVA). Development of Satori will occur under the supervision of Aaron Bloomfield, a professor of Computer Science at UVA. The development team includes Ramya Bhaskara and Joshua Mehr, fourth and third year undergraduate students respectively, both studying Computer Science at the University of Virginia: School of Engineering and Applied Science. Satori will have a feature set that allows students to be helped as efficiently as possible in order to eliminate the wait times of over 2 hours currently seen in office hours. Furthermore, this technical project has applications beyond just CS 2150, as it is being developed for adoption by future course instructors and to be improved upon by future developers.

As AI in U.S education has grown by almost 50% in the last decade, it is not unlikely that some AI systems could be in Satori's future (Marr, n.d., para. 1). However, research has shown that as AI systems grow in use, a concerning lack of attention is given to the underlying data used by said systems. This leads to AI systems being trained on historically inaccurate and biased data sources (Campolo and Crawford, 2020, p. 11). Due to this trend, the STS research aims to answer the question of, "What considerations should be made to collect unbiased data for AI systems in U.S. universities?". The research is concerned with exploring the socio-technical nature of AI systems in a higher education environment and is tightly coupled with the technical

project. Based on an application of Actor-Network-Theory (ANT) by Law and Callon (1988), this paper will contextualize the issue using the Actor-Network-Theory framework and then identify the major considerations needed for unbiased data collection.

ARTIFICIAL SUPERIORITY

At a very high level, a basic classification AI system (1) collects data about users, groups, or the social world, (2) determines the likelihood of some outcome occurring based on a set of notable characteristics of the data, and then (3) makes some sort of determination/classification about a person, group, place, etc. based on similarities with other items in the data set while reasoning and adapting based on the determinations (Sarkar, 2019, para. 4-10). Points (2) and (3) relate to the machine and deep learning models which are the algorithms that enable determinations/classifications of new data points into a set. The success of these algorithms are almost entirely based on the data used but the results of the algorithms have traditionally received the most attention from AI researchers and scholars (Campolo & Crawford, 2020, p. 1). As these AI systems take an increasing role in making deterministic decisions about humans, concerning patterns have emerged that suggest that the technology is not as free from human bias as one may hope.

As reported by Harwell (2019), in 2019, a federal study into the AI facial recognition system used by the FBI to identify over 390,000 suspects since 2011 found that the system often exhibits social and racial biases. According to the study, people of color were misidentified far more by the system than white people, with false-positive rates of up to 100 times in the case of Asian and African Americans (Harwell, 2019, para. 1-2). Despite cases like this, according to AI researchers Campolo and Crawford (2020), the concerning philosophy of enchanted determinism has become mainstream in some AI development circles (p. 1). Enchanted determinism is a

philosophy that treats AI systems with a mystical reverence, where the techniques that drive AI models are not fully understood but their results are held as greater than what can be achieved by normal human interpretation (Campolo and Crawford, 2020, p. 11). This philosophy is held by AI developers and users alike and is used to defend cases such as the one highlighted by Harwell as incidental while diverting blame for any of AI's shortcomings away from any human actors, instead the algorithm itself is blamed (Campolo and Crawford, 2020, p. 12).

ENCHANTED DETERMINISM AND EDUCATION

Enchanted determinism particularly becomes an issue when solutions derived from AI systems are held as objective truth for the classification of people (Bechmann and Bowker, 2019, p. 4). Although a fair amount of research has been conducted looking into AI biases, since AI itself is an emerging technology, the ethical AI movement that aims to devalue enchanted determinism has struggled to gain the necessary force behind it to critically impact the culture surrounding AI (Zeitchik, 2022, para. 26). Enchanted determinists believe that AI systems are so complex in their decision making that they take on a mind of their own; however, these systems are inherently data driven and their results are highly contingent on what data the models were trained on (Olson, 2021, para. 13).

When the primary focus of AI development is on the results rather than the data behind the results, the data used in the AI systems is often ignored. As a consequence, these systems tend to be trained on data sets that reflect historically inaccurate biases and underrepresent minority groups (Campolo and Crawford, 2020, p. 11). With AI in the U.S. education sector set to grow by 47.5% from 2017-2021, a critical analysis of the underlying causes of biased AI systems is crucial (Marr, n.d., para. 1). Furthermore, the research aims to address the uniqueness of education within this topic. Education is far too complex and broad a field to be lumped into

the general discussion of good data collection and the results of an AI system cannot possibly serve to provide a full understanding of a student (Zawacki-Richter et al, 2019, p. 21). Therefore, a focus will be made on the actor-network involved with AI in higher education using a similar Actor-Network-Theory approach outlined by Law and Callon (1988).

DEFINING THE ACTOR-NETWORK-THEORY APPROACH

As highlighted by Zeitchik (2022), AI researchers who are concerned with creating more equitable AI systems are promoting the idea that AI engineers must have a foundational understanding of the human elements of the world that their systems are being injected into in order to create fair AI systems (para. 14). Additionally, although data sets are currently being used in AI systems deployed in the social world, there are still ongoing debates as to what qualifies as "good data collection" and that the context of the social world AI systems are being deployed in are critical in making these assessments (Candelon et al. para. 2). The Actor-Network-Theory approach aims to address these two ideas as AI bias is examined in the network of higher education.

AI IN A SOCIAL WORLD

Actor-Network-Theory (ANT) is the socio-technical theory that an artifact exists within a network of actors which includes all humans, concepts, or organizations and their unique relationships (Law & Callon, 1988, 285). ANT is not a means of discovering an objective truth, rather it serves as a means to organize the socio-technical impact of an artifact and discuss them. Figure 1 on page 5 provides the ANT graph that is used to contextualize the research. With the ANT approach, the inherently socio-technical nature of AI systems in higher education is revealed which allows those involved in the network to identify the major ways in which social systems should be taken into account for data collection.

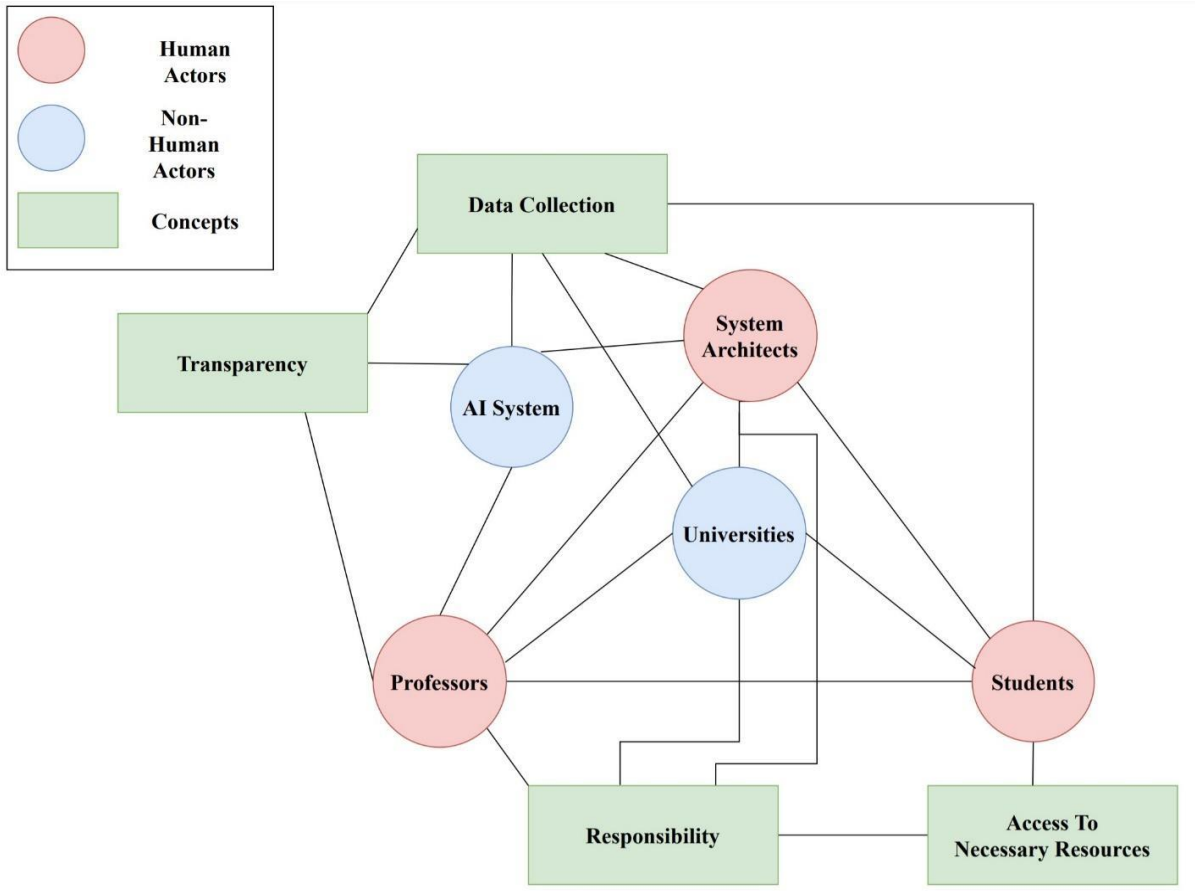

Figure 1: AI System in Higher Education Network. This graph visualizes the network that will be analyzed in the research to identify potentially biased relationships. (Scruggs, 2022).

IDENTIFYING THE CRITICAL RELATIONSHIPS

The human actors include students, AI system architects, and professors. These three actors have a triangular relationship where they all closely affect one another and individually have unique relationships with all other actors in the network. Next, the concept of data collection relates to the students the data pertains to, the underlying concepts behind said data, and the organizers of the data. An insight into how these relationships function and influence one another is key into assessing the potential biases of the network. Finally, the concept of responsibility is of particular interest. Responsibility in both collecting and verifying the data to

ensure no biases will be accessed along with accountability for any negative consequences of AI systems is a key consideration of this research.

The primary research question is: “What considerations should be made to collect unbiased data for AI systems in U.S. universities?”. Given this, the Actor-Network defined above serves to show all the elements involved when answering this question. By the nature of the ANT approach, no one actor is more important to the network and the relationships between each actor is vital to understanding the intricacies of the system (Law and Callon, 1988). According to similar research into general AI data transparency by European ethicist and historian Felzmann et al. (2019), even the most comprehensive of measures taken to address these concerns lacks awareness of the social context surrounding AI systems (p. 9). The following section will explore the current findings on the considerations that both developers and users must take in order to ensure that the social implications of AI systems are taken into account before they are deployed.

AN INFORMED APPROACH TO DATA COLLECTION AND USE

Artificial intelligence, a technology that is meant to mimic human behavior without human emotions, often reflects the same social biases humans do in their own decision making processes (Manyika et al., 2019, para. 2). Rather than pretending like AI systems are emotionless and free of human error, developers must take the responsibility to reveal the data involved in their systems and their reasons for using them in order to create more transparency between the process and the end result.

Rafanelli (2022) suggests that AI systems are highly dependent on the choices of developers such as which data to include and what categories to classify the data into (p. 5). Instead of taking power away from human hands, AI allows developers to make sweeping

decisions based on historical or even unconscious bias in data sets (Rafanelli, 2022, p. 3). In the context of education, the dangers of allowing these biases to be reflected in AI systems is particularly high. As shown in the actor-network above (p. 6), AI systems can affect the determinations professors make about students. Furthermore, Reuters reported that an AI system used by Amazon for recruiting had to be scrapped because the system tended to favor male candidates over female ones (Dastin, 2018, para. 6). If such a system is deployed to make a major decision in higher education, i.e. college admissions decisions, the potential consequences for students could be life-changing and may forever alter the image of a college or university.

DEVELOPER CONSIDERATIONS

It is important for developers themselves to be aware of some of the root causes of bias in AI systems. Bechmann & Bowker (2019) suggest that AI systems will often filter out important characteristics of individuals such as race and gender to better suit the expected results (p. 4). Although this practice may sound malicious, the reality usually is that developers do not give particular care into the process of data cleaning, often just focusing on results. On page 8, Figure 2 shows areas of potential discrimination in the entire AI development cycle. It's important to note that data drives all of these areas, suggesting that it's important to understand all of these possible discrimination areas.

AI and machine learning design phases	Classification and discrimination exemplified
Defining task and outcome variables	Selecting frequent patterns instead of marginal, setting number of clusters, setting limited number of stereotypic outcome variables
Data	Ignoring specific data as not containing meaning thus limiting model to be predisposed to only certain inputs, balancing sample ignoring classes on the margins and unconventional sampling classes, along with groups outside social media
Model selection	Have a subjective understanding of "meaningful clusters, use pretrained model with undocumented biased 'experience', sub-classifiers, weights and thresholds"
Data preparation	Translate into mono-language input and thereby limit native language nuances, reduce image information that potentially would correlate with marginal classes, taking out under/oversharers that could be of interest in terms of evaluating, e.g. validity of outcome variables or overall task
Model training and deployment	Interpret clusters manually and subjectively in dialogue with model logics, interpret false negatives and false positives to detect non-stereotypic sub-classes to include in retraining

Figure 2: Discrimination in The AI Process. Table demonstrating the different ways discrimination may creep into AI systems during the development process (Bechmann & Bowker, 2019).

Many of the situations described in Figure 2 (p. 8) relate to a narrowed focus in the development of AI systems. As shown in the "Data" and "Model Selection" rows, developers often only prepare for certain types of expected inputs into their systems that often favor the majority of people. With such a limited focus on the socio-technical nature of AI systems, those in the minority are often forgotten. Relating back to education, developers must adapt their systems to take into account a diverse student population. The research suggests that there are ways to adapt such complicated technology into a diverse social world, but it starts with commitments or obligations from developers to inform the public about how their systems work.

THE RIGHT TO INFORMED CONSENT

Given the various areas for potential discrimination, developers and users alike need to be aware of AI's potential for discrimination and bias. As mentioned, research on this topic is fairly new but extensive and the majority of action to address these issues has been seen internationally in Europe. The European Union's General Data Protection Regulation (GDPR) is meant to be a piece of legislation that protects users of any technology from having their personal data used without their consent or without their knowledge of what the data is being used for (Felzmann et.

al, 2019, p. 2). The legislation is known as the strictest privacy and security legislation to date and poses hefty fines for those who violate it (Clark, 2021, para 1).

Felzmann et. al (2019) noted that a key component of the GDPR is the complementing ideas of prospective and retrospective transparency (p. 3). Prospective transparency is the idea that before an individual's personal data is processed, those individuals must be informed in clear terms why their data is being processed, who is doing the processing, and how the data will be processed (p. 3). Similarly, retrospective transparency is the right of an individual to have the decision made based on their data explained to them and for the individual to challenge said decision (p. 3). However, since these two ideas do not stop harmful decisions from being made, individuals should have the right to reasonable inference which is a thorough justification of the necessity of the data used and the results made by the system (p. 3). With adding these further protections to the current ideas of the GDPR, an individual would have the information required to give informed consent for their data to be processed (p. 4).

Although the GDPR's policies on transparency and informed consent are miles ahead of any United States legislation (Clark, 2021, para 1), the regulation needs some adjustments in order to adequately adapt to the challenges of transparency in AI (Felzmann et. al, 2019, p. 4). Figure 3 on page 10, contextualizes the ideas of transparency in terms of AI and robotics. These ideas of transparency in AI directly contradict the ideas of enchanted determinism thus adopting the ideas presented in the figure is crucial in order to ensure equitable AI systems in higher

education.

Transparency in the context of robotics and AI	
For a...	To...
Developer	Understand whether their system is working properly in order to identify and remove errors from the system or improve it
User	Provide a sense for what the system is doing and why, to enable intelligibility of future unpredicted actions circumstances and build a sense of trust in the technology
	Understand why one particular decision was reached
	Allow a check that the system worked appropriately
	Enable meaningful challenge (e.g. credit approval or criminal sentencing)
Society broadly	Understand and become comfortable with the strengths and limitations of the system
	Overcome a reasonable fear of the unknown
Expert/Regulator	Provide the ability to audit a prediction or decision trail in detail, particularly (un)intended harmful actions, e.g. a crash by an autonomous car
Deployer	Make a user feel comfortable with a prediction or decision, so that they keep using the system

Figure 3: Technological Transparency in AI. This table highlights the levels of transparency necessary for the actors affected by AI development. (Felzmann et. al, 2019, p. 5).

The idea of transparency in AI acknowledges that the technology is inherently socio-technical in nature but leaves out how best to educate the masses on their AI system. As the technology is rather complex and diverse, there is no one size fits all solution to this. Rather, in the case of education, the network described in Figure 1 (p. 5) highlights the relationships between system architects, professors, and universities. If system architects were to adopt the ideas presented in Figure 3 (p. 10) and present their systems in a transparent and honest way to university professors and officials, they could work in tandem to best present their AI to a diverse student body. This allows professors, who know their students' backgrounds and abilities the best, to be involved with teaching their students about the system in a way that aligns with their understanding of other topics.

TAKING RESPONSIBILITY

Although the GDPR is landmark legislation for data protection and privacy, there is no counterpart in the United States (Clark, 2021, para 1). Although there have recently been state legislation aiming to adapt the ideas of GDPR in states like California and Virginia (Clark, 2021,

para 1-3), the speed of AI adoption is too rapid to wait for regulations like this to be adopted on the national level, thus AI developers need to self regulate themselves now. Specifically, in order to answer the question of: “What considerations should be made to collect unbiased data for AI systems in U.S. universities?”, developers must look to the steps taken in Europe with the GDPR and make the necessary adjustments to adapt the ideas of accountability through transparency. The potential applications of AI are immense and because of both public interest and company investment, the technology is seeing unprecedented levels of growth (McKendrick, 2021, para. 3). However, without careful consideration for the groups of affected actors an AI system can affect and set standards for mitigating the risks to those groups, AI will continue to be misunderstood, distrusted, and potentially harmful to the public.

REFERENCES

Bechmann, A., & Bowker, G. (2019). *Discrimination in The AI Process* [Figure 2]. Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media [Table 2]. *Big Data & Society, 6*(1), 1-11. https://doi.org/10.1177/2053951718819569

Bechmann, A., & Bowker, G. (2019). Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media. *Big Data & Society*, *6*(1), 1-11. https://doi.org/10.1177/2053951718819569

Campolo, A., & Crawford, K. (2020). Enchanted determinism: Power without responsibility in artificial intelligence. *Engaging Science, Technology, and Society, 6*, 1-19.

Candelon, F., Charme di Carlo, R., De Bondt, M., & Evgeniou, T. (2021, October 1). AI regulation is coming. *Harvard Business Review*. https://bit.ly/3bd3Fp8

Clark, B. (2021). GDPR in the USA? New state legislation is making this closer to reality. *The National Law Review, 12*(83). https://tinyurl.com/mr2jz6hh

Dastin, J. (2018, October 10). Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters* https://tinyurl.com/yc33vzt9

Felzmann, H., Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2019). *Technological Transparency in AI* [Figure 3]. Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns [Table 1]. *Big Data & Society, 6*(1), 1-14. https://doi.org/10.1177%2F2053951719860542

Felzmann, H., Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society, 6*(1), 1-14. https://doi.org/10.1177%2F2053951719860542

Harwell, D. (2019, December 19). Federal study confirms racial bias of many facial-recognition systems, casts doubt on their expanding use. *The Washington Post*. https://tinyurl.com/ycktf7nu

Khalid, S., Rongbutsri, N., & Buus, L. (2012). Facilitating adoption of web tools for problem and project based learning activities. *Proceedings of the Eighth International Conference on Networked Learning, 1*, 559-566.

Law, J., & Callon, M. (1988). Engineering and sociology in a military aircraft project: A network analysis of technological change. *Social Problems*, 35(3), 284–297. https://doi.org/10.2307/800623

Marr, B. (n.d.). How is AI used in education-real world examples of today and a peek into the future. https://bit.ly/3rtujU3

McKendrick, J. (2021, September 27). AI adoption skyrocketed over the last 18 months. *Harvard Business Review*. https://bit.ly/3pFMR2z

Olson, P. (2021, October 4). For Tesla, Facebook and others, AI's flaws are getting harder to ignore. *The Washington Post*. https://wapo.st/2ZllArt

Rafanelli, L. (2022). Justice, injustice, and artificial intelligence: Lessons from political theory and philosophy. *Big Data & Society*, *9*(1), 1-5. https://doi.org/10.1177/20539517221080676

Sarkar, S. (2019). *A high level overview of artificial intelligence, machine learning and deep learning*. IBM. https://ibm.co/3ExwCsz

Scruggs, C. (2022). *AI System in Higher Education Network*. [Figure 1]. *STS Research Paper:* Data bias considerations for artificial intelligence systems in higher education (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Zeitchik, S. (2022, January 17). Former Google scientist says the computers that run our lives exploit us - and he has a way to stop them. *The Washington Post*. https://tinyurl.com/5xd8chzt