Ethics of Artificial Intelligence in Society

Emma Johnson[a], Eloy Parrilla[a], & Austin Burg[b]*

[a] *Department of Industrial and Systems Engineering, College of Engineering, NC State University, Raleigh, NC*
[b] *Department of Computer Science, College of Engineering, NC State University, Raleigh, NC*

https://doi.org/10.33697/ajur.2023.070

Students: ekjohns3@ncsu.edu, ehparril@ncsu.edu, acburg@ncsu.edu*
Mentor: Chang S. Nam, csnam@ncsu.edu

ABSTRACT
Every day, artificial intelligence (AI) is becoming more prevalent as new technologies are presented to the public with the intent of integrating them into society. However, these systems are not perfect and are known to cause failures that impact a multitude of people. The purpose of this study is to explore how ethical guidelines are followed by AI when it is being designed and implemented in society. Three ethics theories, along with nine ethical principles of AI, and the Agent, Deed, Consequence (ADC) model were investigated to analyze failures involving AI. When a system fails to follow the models listed, a set of refined ethical principles are created. By analyzing the failures, an understanding of how similar incidents may be prevented was gained. Additionally, the importance of ethics being a part of AI programming was demonstrated, followed by recommendations for the future incorporation of ethics into AI. The term "failure" is specifically used throughout the paper because of the nature in which the events involving AI occur. The events are not necessarily "accidents" since the AI was intended to act in certain ways, but the events are also not "malfunctions" because the AI examples were not internally compromised. For these reasons, the much broader term "failure" is used.

KEYWORDS
Ethics; Artificial Intelligence; Agent-Deed-Consequence (ADC) Model; Principles of Artificial Intelligence; Virtue Ethics; Deontology; Consequentialism; AI Systems

INTRODUCTION

Artificial Intelligence (AI) is becoming more prevalent in society. As its outreach is becoming greater, the need for understanding how to avoid AI shortcomings is pertinent for developing quality applications in the future. Three contemporary examples of AI-related failures and failures are detailed. The relation of the examples to the principles of ethics in AI, the three ethical theories, and the Agent Deed Consequence Model (ADC Model) are analyzed. The discussion and analysis will lead to a greater understanding of ethics in AI, as well as provide useful societal applications. Throughout this research, there is a determination of how each failure occurred and how the failures could be prevented in the future. A system of several steps is proposed with the intention of ensuring AI systems remain ethical. It is hoped that this model will lead to a greater understanding of why AI commits mistakes in order to allow knowledge to lead towards better outcomes and the prevention of AI failures.

METHODS AND PROCEDURES

There is a growing demand for products with integrated AI. This demand brings out a need to ensure AI acts ethically. AI is a system that can operate and complete its designated tasks without direct human input. As AI does not have the same capabilities for morality and ethics as humans do, it is always possible that AI may unintentionally violate ethical boundaries. To determine whether AI acted ethically or not, three ethical theories used to determine the morality of humans were explored: virtue ethics, deontology, and consequentialism. In addition, Dubljević's nine proposed ethical principles for AI were used, 'fairness and non-discrimination', 'privacy, safety and security', 'human control of technology', 'transparency and explainability', 'accountability', 'promotion of human values', 'professional responsibility', and 'sustainable development'.[1,2] The nine principles relate back to aspects of the three ethical theories as well.[2] Using these guidelines, case studies were conducted on three AI failures detailing how they violated specific principles of ethics in AI. An analysis is conducted to show how similar incidents could be avoided in the future using the ADC model.

ETHICAL THEORIES

Three well-known ethical theories that are used to evaluate the ethics and morality of a situation were investigated. These three theories are virtue ethics, deontology, and consequentialism.

Ethical Theories		
Virtue Ethics: Emphasizes the character of a person. It does not attempt to identify singular moral principles to any situation, rather, each person and situation should be evaluated individually.	**Deontology:** Emphasizes whether the actions of a person are right or wrong, and whether those actions respect obligations, duties, and rights in a given situation.	**Consequentialism:** Focuses on whether the outcomes of a situation are morally correct or not. Related to utilitarian ethics which favors the option that will result in the most good.

Table 1. Ethical Theories defined within a table.

Virtue Ethics

Virtue ethics is a broad model that emphasizes the agency or character of a person.[3,4] The theories of virtue ethics, "do not aim primarily to identify universal principles that can be applied in any moral situation," unlike deontology and consequentialism theories.[3,4] Virtue ethics would be concerned with examining virtues such as "responsibility", "righteousness", and "justice". These virtues would serve as motives for why an agent may have acted the way they did. Another way of thinking about virtue ethics would be to consider more than just the actions one took, but also the principles behind one's actions.[5] A virtue ethics philosopher would believe that the reason for which a moral agent goes along with certain actions would be more important than the actions themselves.

Deontology

There are two components of deontology. The first is concerned with the actions of agents themselves and whether those actions are simply right or wrong.[6,7] The second component is concerned with whether the actions of the agent respect the obligations, duties, and rights given the situation.[8] Deontology claims that an agent is moral if it follows these two aspects. Ways to explore deontology ethics would be to identify the specific actions committed by an agent or certain expectations being followed through. Immanuel Kant was one such philosopher who supported this moral basis.[5] Servicing patients in medicine and dentistry can serve as an example to further explain the point. In Washington, debates occurred on whether healthcare workers should be required to get the Covid-19 vaccine for work and if patients should also feel obligated to have the vaccine before going to a healthcare facility.[5] Deontological ethics would examine whether it is right or wrong or set vaccines as mandatory in the healthcare system. To a deontologist, it would be wrong to restrict freedom for the sake of the greater good.[5] In this sense, it would be wrong to mandate vaccinations to healthcare workers and patients.

Consequentialism

Consequentialism focuses on the results of an action, reasoning that an agent is moral if it chooses the most ethical consequence.[8] Also known as utilitarian ethics, consequentialism seeks for a situation to yield the most positive outcomes.[8] Using the same example as before, it is possible to view the same problem from a different ethical lens. Since the consequentialist viewpoint favors the "greater good", a philosopher would argue that it is right to set vaccines as mandatory in the healthcare system. As long as the vast majority of people who received the vaccine did not experience major negative side effects to it, then it would be reasonable to conclude that the greater good prevails over the individual, from the perspective of a consequentialist philosopher.

Philosophers and ethicists have used these three theories to determine the most ethically correct solution regarding specific ethical dilemmas. There cannot be a "most ethical course of action" in certain events because people would have their own opinions on what an AI should do given the context of the situation. A well-known example would be the Trolley Problem: philosophers such as Kant and Mill would have their own opinion on whether the AI driving the Trolley should steer one way or stay on course. Kant would argue that it is up to humans to decide whether the trolley should continue on its course or should be intercepted by a person while Mill would most likely argue that it is better for five people to be saved than one person as is in accordance with the utilitarian point of view.[9] These theories will be utilized through the analysis of the ADC model in section five of the paper to decide what is right and wrong in different aspects of a situation.

PRINCIPLES OF ETHICAL AI

There are nine principles regarding how ethics should be approached in AI: 'fairness and non-discrimination', 'privacy, safety and security', 'human control of technology', 'transparency and explainability', 'accountability', 'promotion of human values', 'professional responsibility', and 'sustainable development'. These principles are further divided into three categories: 'avoiding undesired results', 'liability/acting responsibly', and 'ameliorating the lack of ethics in AI'. These categories and their respective principles will be explained further in this section. **Table 2** shows a breakdown of the three categories with a short description of each principle.

Avoiding Undesired Results	**Fairness and Non-Discrimination**	AI algorithms that are non-discriminatory, fair, inclusive, representative, and free from human biases.
	Privacy	AI use that enables consent, protection from surveillance, and right to control the use of the data gathered.
	Safety and Security	AI that does no harm to humans and resists external threats.
	Human Control of Technology	AI that remains under human control and enables review by those impacted.
Liability/Acting Responsibly	**Transparency and Explainability**	AI that enables oversight and can be explained, understood, and recognized.
	Accountability	Continuous assessment and evaluation of AI use, as well as the creation of new regulations and subsequent liability for failure to meet these regulations.
Ameliorating the Lack of Ethics in AI	**Promotion of Human Values**	AI that is used to benefit society, human civilization, and human rights.
	Professional Responsibility	AI that is designed purposefully and collaboratively with relevant stakeholders.
	Sustainable Development	AI that benefits or does not hinder the development of sustainable societies and objectives.

Table 2. Categorization of the nine principles of AI.

Avoiding Undesired Results

'Avoiding undesired results' is composed of "principles concerned with avoiding the dangers of AI when used for unethical or immoral purposes, whether intentionally or unintentionally".[1,2] The first principle in this category is 'fairness and non-discrimination'. Dubljević's paper, *Ethics of AI in Organizations*, states that "care needs to be taken during both the creation and use of AI-based systems to ensure that human prejudice is not ingrained into the system before or after its initial deployment".[1] 'Fairness and non-discrimination' suggest that "AI should utilize only representative and high-quality data, be used impartially and equally across demographics, and consider a diverse array of stakeholders in its design and implementation".[10] The next principle is 'privacy'. It includes the right to consent to AI-based data collection, analysis, and measures of control over the subsequent use of the data.[1,2] After 'privacy', 'safety and security' are the next principles. 'Safety and security' propose that AI be protected from internal and external threats,[10] as well as there be an element of predictability to the AI for the protection of society and individuals' safety.[10] The last principle in this category is 'human control of technology'. Dubljević states that AI has to remain under human control and enable review by those the technology impacts.[1] The outcomes of AI are ultimately within human governance and "results and decisions stemming from AI technologies should be able to be reviewed, opted out of, challenged, or otherwise managed by people".[10]

Liability/Acting Responsibly

The next category, 'liability or acting responsibly', concludes that AI needs to be designed and utilized under appropriate scrutiny and within legal boundaries.[1] The first principle within this category is 'transparency and explainability'. This is defined by developing AI-based systems that may be easily managed and understood by experts and non-experts.[1] It can also be defined as AI that enables oversight and can be easily explained, understood, and recognized.[10] The second principle categorized as liability and responsibility is 'accountability'. This refers to who or what is accountable for a decision made by an AI-based system and can be further divided into before, during, and after the use of the AI.[1] In *Ethics of AI in Organizations*, Dubljević states that "after an AI-based system's deployment, regulatory systems should exist to rectify unjust decisions made by the AI-based system, in addition to legal liability for those that cause harm using AI".[1]

Ameliorating the Lack of Ethics in AI

After liability, the final category is 'ameliorating the lack of ethical values in AI'. The definition relating to this category is that AI is inherently amoral, therefore rules and laws are needed to guide the way that it is used, making sure it is ethical.[1,11] One principle in this category is the 'promotion of human values'. The principle suggests that AI-based systems should be used for the common good and be deployed/developed consistent with human values.[1] AI systems should be widely available and distributed as equally as possible.[1,10] Along with the equal distribution of AI systems, there needs to be a set of values that are agreed upon based on every culture and idea in the world. The most common values being considered include human dignity, human rights, and fundamental freedoms, leaving no one behind, living in harmony, trustworthiness, diversity and inclusiveness, and protection of

the environment.[11] The second principle in this category is 'professional responsibility'. It suggests that "AI be designed meticulously, purposefully, and with the input of stakeholders across a variety of levels".[1,10] Designers of AI-based systems must consider the long-term effects of their creations, and must therefore ensure they are used in a reliable and valid manner.[10] The final principle is 'sustainability/sustainable development', and it refers to "creating AI technologies that enable maintainable solutions to global problems such as healthcare and equality, minimizing resource waste, and environmental responsibility".[1,10] *The Ethics of AI in Organizations*, also states that it is important that the ethical principles of AI advocate for the avoidance of potentially disastrous outcomes of global warming.[1]

AGENT-DEED-CONSEQUENCE (ADC) MODEL

The ADC Model predicts that moral judgment consists of three components: the character of a person (agent), their actions (deed), and the consequences of the situation (consequence).[12] The ADC model implicitly applies the three moral theories to evaluate different aspects of a situation, as each component of the ADC model is tied to a moral theory. The model predicts that moral judgments are positive if all three of its components are considered positive, and negative if all three of its components are considered negative.[12] This model is useful because it can help one categorize the ethical and unethical elements of a situation.

Ethical Theories	ADC Model
Virtue Ethics: Emphasizes the character of a person. It does not attempt to identify singular moral principles to any situation, rather, each person and situation should be evaluated individually.	**Agent Component:** Is related to virtue ethics, as it emphasizes the traits of the person in a given situation.
Deontology: Emphasizes whether the actions of a person are right or wrong, and whether those actions respect obligations, duties, and rights in a given situation.	**Deed Component:** Is related to deontology, as it emphasizes the actions committed by a person and whether those actions follow moral principles.
Consequentialism: Focuses on whether the outcomes of a situation are morally correct or not. Related to utilitarian ethics which favors the option that will result in the most good.	**Consequence Component:** Is related to consequentialism, as it focuses on the end results of a situation.

Table 3. Ethical theories compared with the Agent-Deed-Consequence Model.

An example event will be explored to better understand each component of the ADC model. This scenario consists of emergency responders arriving at a collapsed building with people stuck under the debris. The emergency responders have a robot integrated with AI to assist them in deciding the most effective and efficient method for helping survivors out from underneath the rubble. The robot will analyze the conditions of the victims, assess the surrounding scenario, and make assumptions to determine how to act when coming across a victim. The victims would have to reach medical care as soon as possible, so the time for retrieval is critical.

Agent Component & Virtue Ethics

Virtue ethics emphasize the agency or character of a person, similar to the agent component of the ADC model.[3,13] The theories of virtue ethics "do not aim primarily to identify universal principles that can be applied in any moral situation," unlike deontology and consequentialist theories, which do.[4]

In relation to the scenario, the Agent component in question would be the robot itself. Since the robot is proactively deciding who to save or who not to save based on specific circumstances, it acts as the individual who is in charge of making ethical decisions. The robot would utilize ethical virtues and the circumstantial information available to it in order to make the most appropriate decision.

Deed and Deontology

Deontology claims that an agent is ethical if "it respects obligations, duties, and rights related to given situations".[8] Deontology specifically relates to the actions of a person, similar to the deed component of the ADC model. In the case of the example, the decision made by the robot to save certain people over others, or how the robot saved certain people, would make up the Deed component. A robot might decide to prioritize saving one person over another because it believes one has a much greater probability of survival. A robot might also decide that to save a person, it would have to cut off someone's leg if it is stuck under rubble and there is no quick way to remove it. The robot may have to be put in a position in which consent from the victim is "implied" if they are unconscious, and this would affect the ethical perception of human operators.

Consequence and Consequentialism

Consequentialism relates to the results of actions that are performed and defines virtues as traits that yield good consequences.[8] It focuses on judging the moral worth of the results of actions, related to the consequences component of the ADC model. Regarding the scenario, the result of the assessment and decisions made by the robot would contribute to the Consequence component of the ADC model. If the people who were prioritized survived and those who weren't prioritized didn't, negative

backlash could arise against utilizing the robot. On the other hand, if the people the robot brought out of the rubble did not survive, a different negative backlash could occur due to the inability to save anyone.

CASE STUDIES OF AI FAILURES

While there has not been much research done on the ethical consequences of artificial intelligence, there are plenty of examples of failure of AI. In this section, case studies were conducted on three different AI failures to analyze how they violated ethical principles of AI. To prevent similar cases from occurring, a summary of the events, a discussion of the implications, and recommendations were given.

The three events reviewed in this section are an Amazon AI hiring tool, a company's AI system that carries out automatic tasks such as renewing contracts and access, and a Microsoft chatbot AI sent out on Twitter. These three AI systems failed to meet their expectations regardless of whether it was the programmers' fault or human input. In turn, there is something to be learned from each failure.

Case 1: The Sexist Amazon AI Hiring Tool

An AI recruiting tool from Amazon showed significant negative bias towards women when analyzing resumés. The experimental tool used an AI system to give candidates scores from one to five stars, based on their resumés, with the goal of selecting the top five applicants.[14] The company soon realized that the AI tool was not selecting candidates for software development and other technical jobs gender-neutrally. Since the system was programmed to select resumés similar to those submitted to the company over a ten-year period, it inherited a bias towards men, as the company had mostly employed men in the past.[14] Male dominance in the technological industry also played a role in this failure. The program took points away from resumés that included words referencing women and even "downgraded graduates of two all-women's colleges,".[14] When trying to solve this problem, Amazon worked to edit the program, making the program remain neutral to terms referring to women but the system would continuously find a way to circumvent this, based on what it was originally taught.

Analysis

This particular case of AI mistake falls under the Ethics of AI category of 'Fairness and Non-Discrimination'. Dubljević states in his *Ethics of AI in Organizations* paper that "care needs to be taken during both the creation and use of AI-based systems to ensure that human prejudice is not ingrained into the system before or after its initial deployment".[1] The 'fairness and non-discrimination' principle, as detailed above, suggests that AI should only use highly representative data. Despite having a lack of diverse records, the AI system needs to consider that there are other participants in society outside of the data provided.. . The system needs to be able to consider the diverse demographics of the real world.[10] In this case, the AI system did not obtain the original resources to have a highly representative set of data. The program was given the resumés that humans had gone through before which were biased as a result of the time period, i.e. not many women in the workforce or focused on the technological industry. Furthermore, Amazon's hiring tool did not promote human values by disregarding the common good of its applicants. Since the tool did not look at applicants objectively and give a fair chance to each one, based on sex, it did not promote the common good of each applicant. Amazon failed to follow two additional principles as well, 'professional responsibility and accountability'. Amazon declined to comment on the incident,[14] showing indifference and lack of accountability for the tool they implemented, declining applicants without human interaction. The refusal to comment on the incident also highlights a lack of diversity and misguided judgments on the implementation of ethical guidelines for this tool. This, in turn, opposes the principle of 'professional responsibility'. In reference to the three main ethical theories, deontology, virtue ethics, and consequentialism, the bias incident discussed, has a stake in each. Relating to deontology and action in the ADC model, in the beginning, an AI system should be created that is unbiased and fair in its selections and ideas. Taking the time to develop the program correctly and without bias will save time in the long run. The virtue ethics key to this issue is to create algorithms that are fair, inclusive, non-discriminatory, and representative of all people; this is because the AI could be considered as an agent in making decisions, if the AI is treated as its own entity and considered capable of making its own decisions beyond what the developers had in mind. This is important because it not only makes places more diverse, but it shows a realistic portrait of the world right now. Finally, in reference to consequentialism, know the consequences of a biased program and the unfairness that comes from it. Everyone should be represented equally in a system. If there is fairness and non-discrimination in the development of the program, there will be less time spent trying to fix the system.

Recommendation

As companies are integrating AI into their systems and workplace, the indifference to consequences and possible errors associated with AI needs to be addressed. Using the Agent component of the ADC model, each applicant's character for the company should be analyzed. In relation to the Deed component of the ADC model, the action that is taken by the system should always be extensively researched before implementation and there should always be the ability to have human intervention. There should be a responsible party, making sure that the AI is following the correct guidelines, reevaluating each applicant based on their own

merits. The consequence component of the ADC model would be associated with the AI realizing that individuals should not be voted down when certain words are used in their resumé, especially when the initial sample of applicants was not representative of society as a whole. Through evaluating and implementing these components and characteristics of AI, a world will be created that promotes human values and equality.

Case 2: The Man fired by AI

Mr. Ibrahim Diallo became the victim of an AI failure when a system controlled by the company he worked for terminated his employment status with the company upon the expiration of his contract, without knowing of this expiration or termination.[15] Diallo's false termination was first discussed after his key card would not allow him access to the building or the logins for his computer. Shortly after noticing, he seemed to have resolved the issue with his manager; nevertheless, security escorted him out of the building.[16] Looking into the reason for his termination, Diallo found that the contract, originally signed when he was hired, had not been renewed. The AI system promptly terminated his employment at the company due to prior obligation and programming. This resulted in Mr. Ibrahim losing three full weeks of pay and the eventual resignation from his position.[15]

Analysis

Diallo's story shows a need for ethics in AI and how automated, intelligent systems can result in harm not only to the employee but the employer as well. The company could lose effective employees through scenarios such as this one. This AI failure fits under the ethical principles of 'human control of technology', 'professional responsibility', 'transparency, and explainability', as well as 'privacy'.[1] Humans remaining in control of technology connects with deontology as being able to review the impacts of technology and artificial intelligence. Without humans taking correct and decisive action against such failures, AI can continue to make these mistakes. Also, humans remaining in control of technology is aligned with consequentialism because if AI is not observed and inspected for mistakes by humans, there will be consequences in the future. Diallo's story connects with virtue ethics in that humans must continue to have 'professional responsibility' when it comes to AI. There needs to be a chain of command when it comes to the blame of AI, when it makes a mistake, especially in the professional world. It was the professional responsibility of Diallo's bosses to correctly reinstate his credentials within the computer system, instead of only reassuring him vocally that he had not lost his job. The lack of professionalism is eventually what made him have to leave his job. The company not taking the time to review Diallo's problem shows a lack of 'transparency and explainability'. It is the employer's responsibility to ensure the employee remains informed of their contract status. Nevertheless, the employers failed to do so. The privacy of Mr. Ibrahim was also breached when the company was required to escalate the situation to higher and higher levels of management. Diallo continuously had to ask others for their access to buildings and software because his had been revoked. This sharing of information not only hindered Diallo's privacy but the privacy of those around him. The consequentialist view of this situation is that artificial intelligence and other technology could get out of control and humans would not be able to stop it from making unethical decisions. The human view of the ethical decision-making process would possibly be obsolete.

Recommendation

Ibrahim Diallo's experience with his employer exacerbates the need for guidelines in reference to AI. His situation could have been avoided if the system prompted a human supervisor to approve of any major tasks/decisions such as completely revoking an employee's status with the company. The agent component of the ADC model connects to this AI failure by allowing for analysis and rechecking of decisions made by nonhumans. The deed component of the ADC model is that the system should have been double-checked during the affair as well as being equipped with a course of action for situations such as these. The employer should be able to stop the process of termination if need be. The consequence aspect of the AI failure is that Diallo's employer lost a good employee who was continuously regarded as "receiving constant praises" and whose "work spoke for itself,".[16]

Case 3: Microsoft Chatbot Learns Prejudice and Discrimination

Microsoft released a chatbot named "Tay" in 2016 with the intention of learning from humans on Twitter. The bot was meant to reflect the account of a teenage girl, learning from humans by interacting through tweets.[17] Just twenty-four hours after the chatbot was released, the tweets changed from, "humans are cool" to supporting neo-nazism, racism, misogyny, and genocide in its following interactions.[17] Microsoft took down the chatbot swiftly and made an apology statement to Twitter users. Microsoft revealed that Tay was targeted by hate groups and Microsoft employees should've prevented this from happening.[18]

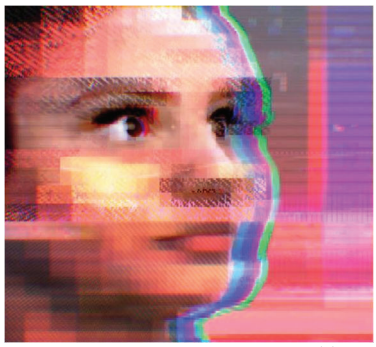

Figure 1. Profile picture of chatbot Tay.

Analysis

Through the quick progression of Tay learning prejudice and discrimination, it is shown how human interactions can have negative effects on AI when not programmed keeping ethical considerations in mind. Tay was programmed to learn from interactions with users, without discerning between moral and immoral values, leading to failure in adhering to the principle of 'promotion of human values'.[1] Since the ethical principle proposes that AI should be used for the common good, Tay and other chatbots would need to be created with the capacity to differentiate between positive and negative human values, so that it may only learn the former. Tay is also an example of an AI that had an unintended and unpredictable outcome. Microsoft did not foresee Tay becoming discriminatory and prejudicial, this shows a level of unpredictability that breaks the ethical principles of 'fairness and non-discrimination' and 'safety and security'.[1] This major shortcoming of Tay shows how AI and machine learning, without proper safety nets and protocols, can lead to catastrophic events. Though Tay could not physically harm Twitter users, an AI that interacted with people personally in real-life applications could have had devastating effects on users and non-users alike. Take for instance an autonomous vehicle that could learn misogyny from its user. If a situation arose when a crash was inevitable between hitting a man or a woman, the AI may end up targeting the woman since the user demonstrated a negative bias towards women in general. If AI is created with the ability to learn characteristics about its user or users it interacts with, it must also be created with guidelines to not learn human notions such as misogyny, racism, or other hateful beliefs.

Recommendation

One method that can be employed to ensure AI only learns a certain set of values is through the use of the ADC Model. As the AI views and interacts with users, it has the opportunity to study their beliefs, attitudes, and behaviors to learn how to be better suited to users. Using the agent component of the ADC Model, the AI would have to recognize the intrinsic characteristics of the users it interacts with, as well as recognize how its agency may be viewed by others. Concerning the deed component, the AI would need to recognize actions that are inappropriate and ensure it does not copy them. In relation to the consequence component, the AI would need to realize how the results of its actions could impact others negatively. Through evaluating these characteristics, the AI would ensure its learned characteristics do not match negative human values and fulfill the three components of the ADC Model. Through AI only learning positive human virtues and values, AI will have the potential to benefit users and humanity as a whole.

Case	Principles	Overall Impact
The Sexist Amazon AI Hiring Tool	- Fairness and Non-Discrimination - Accountability - Professional Responsibility - Promotion of Human Values	AI needs to be created with the goal of treating everyone equally no matter the gender, race, ethnic background, sexual orientation, etc.
The Man Fired by AI	- Transparency and Explainability - Human Control of Technology - Professional Responsibility - Privacy	AI systems may need human input before carrying out major tasks.
Microsoft Chatbot Learns Prejudice and Discrimination	- Fairness and Non-Discrimination - Safety and Security - Promotion of Human Values	AI must not inherit human biases that can harm humans or other AI. Additionally, it is important for humans to realize the impact they have on AI intelligence and learning.

Table 4. Table analyzing each of the cases of AI failures and their overall impact on society.

REVISED PRINCIPLES & DEVELOPED MODEL

As was the intention from the beginning, a model was created to help identify the unethical aspects of a situation in which AI failed or committed failures. Through the analyses of three unethical AI failures, a model was developed that would help guide society to recognize specific unethical aspects of AI. Going through the steps in the model, one can determine what components of the ADC model were unethical, what principles of ethical AI were violated, as well as how AI was involved in the failure. The model is named the "AI Failure Analysis Model (AIFAM).

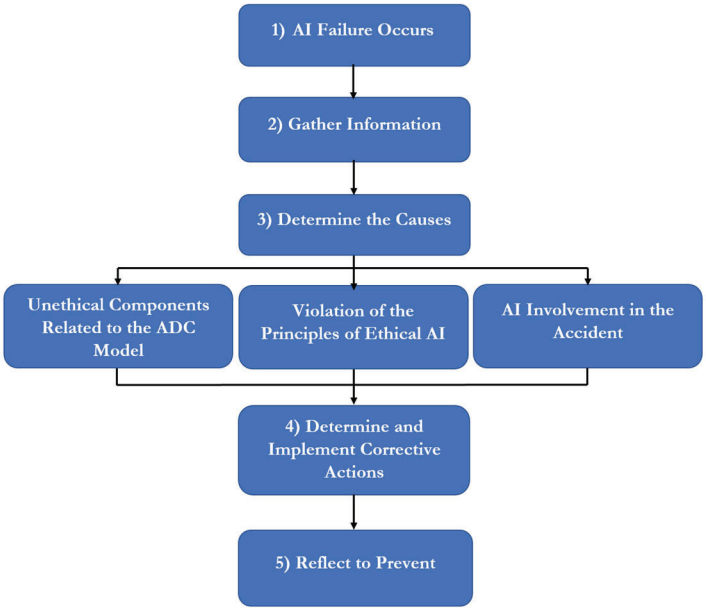

Figure 2. The AI Failure Analysis Model (AIFAM).

It is recognized that this model could be useful for creators or users of AI to look back and ensure their systems operate ethically after a failure occurs. The model could be utilized at companies such as Tesla, Amazon, or Google to objectively reevaluate their AI, and make necessary changes to it if necessary. Using the model, companies would demonstrate, to the public, how they plan on identifying and repurposing their AI after a failure occurs.

Steps of the Model

In this subsection, the steps of the AIFAM model are described and explained.

1. AI Failure Occurs
    a. AI fails and does not fulfill its purpose; the AI deviates from its programmed intention.
2. Gather Information
    a. The person/group conducting research on the failure obtains as much information about the failure, the AI used, and any helpful contextual information.
3. Determine the Causes
    a. At this step, the factors that led to the failure are examined. Each component has sample questions to help identify the causes and unethical aspects of the failure.
    b. ADC Model:
        i. What aspects of the situation were unethical? Agent, Deed, and/or Consequence?
        ii. How were they unethical?
        iii. Why were they unethical?
    c. Violation of the Principles of Ethical AI
        i. What principles of ethical AI were not followed?
        ii. How were they not followed?
        iii. Why were they not followed?
    d. AI Involvement in the Failure
        i. What was the AI originally intended/programmed to do?
        ii. How did the AI fail to meet its programming?
        iii. Why did the AI not follow its protocols?
4. Determine and Implement Corrective Action
    a. By knowing the causes for which the AI acted unethically, it is possible to determine how to tackle those causes to ensure they do not cause more failures. One should ensure these failures are not carried out by implementing corrective actions on the AI.
5. Reflect to Prevent
    a. Using the information collected, the analysis made, and the corrective action determined and implemented, one is able to reflect on the AI failure as a whole and ensure similar occurrences do not occur in the future.

DISCUSSION & CONCLUSION

The different failures involving AI showcase how ethical principles can be infringed upon in real-world applications. Though the failures of imperfect AI pale in comparison to science-fiction movies such as *The Terminator*, where the AI evolves on its own and attempts to carry out human extinction, the effects of these failures were still felt by real people such as Diallo who resigned his position and women who could have been hired by Amazon. The creators of AI must adopt ethical mindsets if they hope for their systems to operate ethically and have little to no possibility of harming human beings. Ethical principles are designed to protect humans from the dangers of AI, and they are further used to advance humanity as a whole to greater heights. Specific instructions on how to program or teach AI to respect and adhere to the ethical principles of AI are not provided, as that is outside the scope of this paper. However, guidance is given to ensure AI remains ethical in its growth and development.

Through exploring widely different examples of failures in AI, ethical principles of AI were observed in real-world applications. The Amazon AI hiring system trespassed on the principle of fairness and non-discrimination. The AI wrongfully firing a man violated privacy and professional responsibility. Microsoft's Tay chatbot could not fulfill the promotion of human values. These failures show why ethics in AI is important, and they encourage the creators of AI to prevent these failures from occurring again by taking closer looks at the causes of the events and why the AI might've acted the way it did. The AIFAM was created to analyze these events, and prevent them from happening again in the future, as it incorporated information from the ADC model, the principles of ethical AI, and the AI failure itself.

Nevertheless, the AIFAM should not be seen as a definitive and final guide to analyzing past failures and preventing future failures. Rather, it should be regarded as a starting point for ensuring AI remains ethical as it develops throughout the years. Future research is needed in areas such as analyzing more specific details of why AI behaves the way it does and how certain acts could be viewed as ethical or not to AI. These two topics of research may enhance the AIFAM or may need their own model to more closely examine the causes or why AI acts unethically. Having the AIFAM serve as a starting point for more research, it is hoped that AI remains ethical in terms of both its use and development.

ACKNOWLEDGMENTS

The authors thank Dr. Chang Nam for his guidance and assistance in completing this research and his mentoring at every step of the way. Thanks go to NC State University's Industrial and Systems Engineering Department as well for their help and support throughout this process. The authors also thank the reviewers of this work for their time and feedback.

REFERENCES

1. Dubljević, V. (2022) Ethics of AI in Organizations in Human-Centered Artificial Intelligence (Nam, C; Jung, J; Lee, S., Ed.) 221-239, Elsevier. *DOI:10.1016/B978-0-323-85648-5.00019-0*
2. Ouchchy, L., Coin, A. & Dubljević, V. (2020) AI in the headlines: The portrayal of the ethical issues of artificial intelligence in the media. *AI & Society*, Volume 35, 927–936. *https://doi.org/10.1007/s00146-020-00965-5*
3. Athanassoulis, N. (2007) Virtue Ethics, *Internet Encyclopedia of Philosophy: A Peer-Reviewed Academic Resource*, *https://iep.utm.edu/virtue/*
4. Statman, D., Trianosky, G.V. (1997) What is Virtue Ethics All About?, *Virtue Ethics* (Cambridge: Edinburgh University Press)
5. Dillon, J. (2021) Utilitarian vs deontological ethics in medicine and dentistry, *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology,* Volume 132, Issue 6, 617-618, *https://doi.org/10.1016/j.oooo.2021.08.025*.
6. Dubljević, V., and Racine, E. (2014) The ADC of moral judgment: Opening the black box of moral intuitions with heuristics about agents, deeds and consequences. *AJOB Neuroscience*, Volume 5, Issue 4, 3–20.
7. Dewey, J. (1966): Three independent factors in morals. *Educ Theory.*; Volume 16:198–209.
8. Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., & Yang, Q. (2018). Building ethics into Artificial Intelligence. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI 2018),* 5527–5533. *https://arxiv.org/abs/1812.02953v1*
9. Grincevičienė, V., Barevičiūtė, J., Asakavičiūtė, V., & Targamadzė, V. (2019) Equal opportunities and dignity as values in the perspective of I. Kant's deontological ethics: The case of inclusive education. *Filosifija Sociologija.*, Volume 30, Issue 1, 80-88. *https://www.proquest.com/scholarly-journals/equal-opportunities-dignity-as-values-perspective/docview/2213858860/se-2?accountid=12725*
10. Fjeld, J., Achten, N., Hiligoss, H., Nagy, A. C., & Srikumar, M. (2020) Principled Artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center for Internet & Society. https://dash.harvard.edu/handle/1/42160420*
11. UNESCO, (2020) Elaboration of a Recommendation on the ethics of artificial intelligence, Ad Hoc Expert Group, *https://en.unesco.org/artificial-intelligence/ethics*

12. Dubljević, V., Sattler, S., & Racine E. (2018) Deciphering moral intuition: How agents, deeds and consequences influence moral judgment, *PLOS One,* Volume 13, Issue 10, *https://doi.org/10.1371/journal.pone.0204631*.

13. Zizzo, N., Bell, E., & Racine, E. (2016) What Is Everyday Ethics? A Review and a Proposal for an Integrative Concept. *The Journal of Clinical Ethics. https://pubmed.ncbi.nlm.nih.gov/27333062/*

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

15. Wakefield, J. (2018) The man who was fired by a machine. BBC News. *https://www.bbc.com/news/technology-44561838*

16. Diallo, I. (2018) The Machine Fired Me. iDiallo. *https://idiallo.com/blog/when-a-machine-fired-me*

17. Wakefield, J. (2016) Microsoft chatbot is taught to swear on Twitter. BBC News. *https://www.bbc.com/news/technology-35890188*.

18. Reese, H. (2016) Why Microsoft's 'Tay' Ai Bot Went wrong. TechRepublic. *https://www.techrepublic.com/article/why-microsofts-tay-ai-bot-went-wrong/*.

19. BBC. (2018) Amazon scrapped 'sexist AI' tool. BBC News. *https://www.bbc.com/news/technology-45809919*.

20. Lee, P. (2016) Learning from Tay's introduction. The Official Microsoft Blog. *https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/#sm.00000gjdpwwcfcus11t6oo6dw79gw*.

21. Price, R. (2016) Microsoft's genocidal AI chatbot is broken again. Business Insider. *https://www.businessinsider.com/microsoft-ai-tay-twitter-racist-genocidal-breaks-down-repeats-too-fast-2016-3*.

22. United States National Transportation Safety Board (1995) Marine Accident Report: Grounding of the Panamanian Passenger Ship *Royal Majesty* on Rose and Crown Shoal Near Nantucket, Massachusetts June 10, 1995. *https://permanent.fdlp.gov/websites/www.ntsb.gov/publictn/1997/MAR9701.pdf*

23. Wallach, W., Allen, C. & Smit, I. (2008) Machine morality: bottom-up and top-down approaches for modelling human moral faculties. *AI & Society*, Volume 22, 565–582. *https://doi.org/10.1007/s00146-007-0099-0*

24. Andersen, S. (2001) Theological Ethics, Moral Philosophy, and Natural Law. *Ethical Theory and Moral Practice,* Volume 4, 349–364. *https://doi.org/10.1023/A:1013318824823*

25. Knoll, M. (2019). Machiavelli's Consequentialist Ethics of Responsibility. *History of Political Thought,* Volume 40, Issue 4, 631-648. *https://proxying.lib.ncsu.edu/index.php/login?url=https://www-proquest-com.prox.lib.ncsu.edu/scholarly-journals/machiavellis-consequentialist-ethics/docview/2312965948/se-2*

26. Stanziani, A. (2021). Utilitarianism and the question of free labor in russia and india, in the eighteenth and nineteenth centuries. *International Journal of Asian Studies*, Volume 18, Issue 2, 153-171. *doi: https://doi-org.prox.lib.ncsu.edu/10.1017/S1479591420000583*

ABOUT STUDENT AUTHORS

Emma Johnson is currently an undergraduate student at North Carolina State University working towards a B.S. in Industrial and Systems Engineering with a focus in Health and Human Systems. She has worked as a research assistant under Dr. Chang Nam in the Department of Industrial and Systems Engineering, assisting with AI and ethics research. Emma has also worked as a research and development engineering intern for Adhezion Biomedical.

Eloy Parrilla is an undergraduate student at North Carolina State University working towards a B.S. in Industrial and Systems Engineering. He previously worked as a research assistant under Dr. Chang Nam where he aided in research of ethics and artificial intelligence. He also worked as an engineering intern for Altec Industries.

Austin Burg is currently working towards a B.S. in Computer Science at North Carolina State University. He works as a research assistant for the Neuro-computational ethics group under Dr. Veljko Dublević, simulating ethical scenarios in virtual reality. He is also a co-head of the Service Raleigh web committee, a volunteer organization, where he facilitates web development.

PRESS SUMMARY

Have you wondered how Artificial Intelligence (AI) makes decisions? What about the ethics of those decisions? How does AI help or harm different people? Artificial intelligence is not always perfect and is known to cause failures that impact a multitude of people. The purpose of this study is to explore how ethical guidelines are followed by AI when it is being designed and implemented in society. Three ethics theories, along with nine ethical principles of AI, and another model associated with ethics were investigated to analyze failures involving AI. When a system fails to follow the models and theories, a set of refined ethical principles are created. By analyzing these AI failures, an understanding of how similar incidents may be prevented can be gained. Additionally, the importance of ethics being a part of AI programming is demonstrated, followed by recommendations for the future incorporation of ethics into AI.