***Mini-Exam Two: Ethics of AI Name:* Gavin Lynch**

**atGPT(4)**

In this exam, **you will evaluate the short-essay answers that ChatGPT generated** to the questions below (in blue). Determine whether ChatGPT’s answer to each question is 1) exactly right, 2) partly right, or 3) wrong, and then, if the answer is not exactly right, explain why. Use these criteria:

1) exactly right: ChatGPT answered the question *precisely correctly* and also avoided adding *incorrect* or *incoherent* things or adding *lots of irrelevant* things;

2) partly right: CbatGPT answered the question correctly but *vaguely*; or ChatGPT answered the question precisely correctly but then also added some incorrect or incoherent things or added lots of irrelevant things, or

3) wrong: ChatGPT either gave an *incorrect or incoherent answer* to the question, or just *failed to answer* the question, even vaguely, despite perhaps saying lots of otherwise relevant and coherent correct things.

(Hint: I suggest you answer the questions yourself by consulting lecture slides and your notes, before reading ChatGPT’s answers. ChatGPT’s answers often sound good but can be misleading.)

Make a copy of this exam and type your answers in the spaces provided below. Please try to avoid answers that exceed the spaces allocated; you need not write complete sentences. Each numbered question is worth 4 points (for a total of 36).

* You are free to use notes, powerpoints, articles, etc., but the exam should be your own work. Do not work with other students.
* Save your completed exam as a pdf or Word document and upload it to the appropriate assignment box on Canvas.
* Good luck!

**I. Autonomous Systems**

1. What feature distinguishes the ethics of *autonomous* AI agents from that of other AI agents? How does this feature generate “responsibility gaps”? Give an example of an autonomous system that generates a responsibility gap and propose a way to close it.

(i - what distinguishes the ethics of autonomous AI agents): Autonomous AI agents are distinguished by their ability to make decisions and take actions without human intervention. They can assess situations, make choices based on programmed criteria or learned experiences, and act upon these choices independently.

**( ) exactly right, ( X ) partly right, or ( ) wrong? (check one) If not exactly right, explain why:**

This answer is almost wrong, as what distinguishes autonomous AI agents from other AI agents is their ability to learn from past experiences, and then act independently and make decisions that were not explicitly programmed in.

(ii – how this feature generates responsibility gaps): This independence in decision-making can lead to "responsibility gaps." Since autonomous agents act without direct human control, it becomes challenging to attribute responsibility for their actions. This gap arises because traditional legal and ethical frameworks are built around human decision-makers, and autonomous agents do not fit neatly into these frameworks.

**(X) exactly right, ( ) partly right, or ( ) wrong? (check one) If not exactly right, explain why:**

(iii – an example autonomous system with a responsibility gap): A self-driving car represents an autonomous system that can generate a responsibility gap. If a self-driving car is involved in an accident, determining liability becomes complex. Is it the manufacturer, the software developer, the owner, or the car itself at fault?

**(X) exactly right, ( ) partly right, or ( ) wrong? (check one) If not exactly right, explain why:**

(iv – proposal to close it): To address this gap, a new legal framework could be developed, specifically for autonomous systems. This could involve creating standards for assigning levels of responsibility to different stakeholders (manufacturers, programmers, users) and perhaps establishing a special insurance or compensation fund for incidents involving autonomous systems. Additionally, ensuring transparency in how these systems make decisions could aid in attributing responsibility.

**(X) exactly right, ( ) partly right, or ( ) wrong? (check one) If not exactly right, explain why:**

2. What does Professor Wright mean when she claims that autonomous machine agents should be “rightful machines?” Give an example of one of the ethical preferences found by the Moral Machine Experiment that a rightful machine should ignore and explain why.

(i - rightful machine): When Professor Wright talks about autonomous machine agents being "rightful machines," she refers to the idea that these machines should operate within the bounds of legal and ethical norms, respecting fundamental human rights and values. This concept implies that autonomous machines should not only be technically efficient but also align with societal and moral standards.

**( ) exactly right, (X) partly right, or ( ) wrong? (check one) If not exactly right, explain why:**

**The answer is correct in that a rightful machine should respect the bounds of the legal system, but this does not include ethical opinions. Rightful machines should uphold the ideas of public law and justice over all else**

(ii - example preference and why to ignore it): An example from the Moral Machine Experiment might be the preference for sacrificing older individuals to save younger ones. A "rightful machine" should ignore such a preference because it discriminates based on age, which goes against the principle of equal respect and worth of all human lives, regardless of age. Ethical principles like non-discrimination and the equal value of human life should guide the decisions of autonomous machines, rather than the utilitarian calculations that might emerge from aggregated public opinion.

**( ) exactly right, (X) partly right, or ( ) wrong? (check one) If not exactly right, explain why:**

While this could be argued as an example to skip, the real reason to skip it would be because of the legal ramifications and justice-based consequences of discrimination, which the rightful agent is trying to hold up.

**II. Algorithmic Justice**

3. Indicate which type of error (false positive, FP or false negative, FN) seems more important to avoid in each of the following applications (that is, which type of error seems worse than the other?). Put an X next to the error that is worse (FP or FN) and explain why.

LAWS that targets and kills enemy combatants:

(X) FP or ( ) FN

In the context of LAWS (Lethal Autonomous Weapons Systems) targeting and killing enemy combatants, avoiding false positives (i.e., mistakenly targeting non-enemy individuals or civilians) is more important than avoiding false negatives (i.e., failing to target enemy combatants). False positives can result in the loss of innocent lives and lead to significant ethical and legal consequences, while false negatives, while undesirable, may not have the same immediate and severe consequences in a military context.

**(X) exactly right, ( ) partly right, or ( ) wrong? (check one) If not exactly right, explain why:**

A system that identifies who is eligible for food stamps:

( ) FP or (X) FN

In a system that identifies eligibility for food stamps, avoiding false negatives (i.e., failing to identify eligible individuals who are in need) is more important than avoiding false positives (i.e., incorrectly identifying someone as eligible when they are not). False negatives could result in deserving individuals going without necessary assistance, potentially leading to hunger and poverty. False positives may result in a minor administrative burden but are generally less harmful in this context.

**(X) exactly right, ( ) partly right, or ( ) wrong? (check one) If not exactly right, explain why:**

A criminal legal system that judges who is guilty of a crime (and punishes them):

(X) FP or ( ) FN

In a criminal legal system, avoiding false positives (i.e., convicting innocent individuals) is more important than avoiding false negatives (i.e., failing to convict guilty individuals). Convicting innocent people not only violates their rights but also undermines the integrity and credibility of the justice system. While false negatives can lead to guilty individuals going unpunished, this is generally considered a lesser evil compared to the miscarriage of justice caused by false positives.

**( ) exactly right, (X) partly right, or ( ) wrong? (check one) If not exactly right, explain why:**

I feel that this depends a lot on the types of criminals, for serial murderers I would argue that releasing one would be of more harm than accidentally imprisoning someone wrong, but for many minor crimes, I agree with Chat GPT's answer.

A system that determines which passengers boarding an airplane to search for explosives:

( ) FP or (X) FN

In a system that determines which passengers to search for explosives at an airport, avoiding false negatives (i.e., failing to detect an actual threat) is more important than avoiding false positives (i.e., subjecting innocent passengers to additional security checks). The primary goal of airport security is to prevent potential acts of terrorism, and missing a real threat (false negative) can have catastrophic consequences. While false positives can be inconvenient for passengers, they are generally less severe than a security breach.

**(X) exactly right, ( ) partly right, or ( ) wrong? (check one) If not exactly right, explain why:**

4. (a) What are the two main ways that algorithmic systems created with machine-learning methods might make decisions that seem unfair? Which type of unfairness problem has an uncontroversial solution in theory, even if it is not always possible to achieve in practice? Why can’t the other one be solved without controversy?

(i - two main unfairness problems):

Bias in Training Data: This occurs when the data used to train the algorithm is not representative of the population or contains historical biases. The algorithm learns these biases and perpetuates them in its decisions. For example, if a hiring algorithm is trained mostly on data from male employees, it may favor male candidates, leading to gender bias.

Algorithmic Opacity and Complexity: Some machine learning models, especially deep learning models, are often described as 'black boxes' because their decision-making processes are not transparent. The complexity and lack of interpretability can make it difficult to understand how decisions are being made, which can be perceived as unfair, especially if those decisions have significant consequences for individuals.

**( ) exactly right, (X) partly right, or ( ) wrong? (check one) If not exactly right, explain why:**

While the first answer on the Bias in Training data is correct, the second answer on the algorithmic opacity and complexity is wrong. The second most common reason for unfairness is an unfair distribution in error. This could mean for certain groups, the algorithm selects more false positives or false negatives, or some other type of error.

(ii - which one has a solution, in theory): The issue of Bias in Training Data has an uncontroversial solution in theory. By ensuring that the training data is representative and free of historical biases, the fairness of the algorithm's decisions can be improved. While it can be challenging to achieve in practice, the approach to resolving this issue is conceptually straightforward.

**(X) exactly right, ( ) partly right, or ( ) wrong? (check one) If not exactly right, explain why:**

(iii – why the other type of unfairness cannot be resolved without controversy) The problem of algorithmic opacity and complexity in machine learning is difficult to solve, even in theory, because these models are often too complex and not transparent, making it hard to understand and fix how they make decisions. Additionally, fairness is a complex and context-dependent concept, which adds to the challenge.

**( ) exactly right, ( ) partly right, or (X) wrong? (check one) If not exactly right, explain why:**

Chat GPT has the wrong problem again. The issue of unfair distribution of error has controversial answers, the reason being it is unclear how error should really be distributed. People argue if it should be distributed evenly, what maximizes utility, or what you deserve for example.

(b) Provide an example drawn *from class lecture* of (i) an algorithmic decision-making system where the unfairness issue is relatively clear and easy to fix in theory, and then (ii) another example from class where it is not.

(i - unfairness is clear and fixable): An example could be a credit scoring algorithm that was trained on historical data showing that people from a certain neighborhood have higher default rates. This could lead to the algorithm unfairly denying credit to individuals from that neighborhood. In theory, this is relatively easy to fix by adjusting the training data to not include or downplay geographic location as a factor in credit scoring, or by ensuring that the data is representative of all demographics.

**( ) exactly right, ( ) partly right, or (X) wrong? (check one) If not exactly right, explain why:**

**The example discussed in class was that of an algorithm to predict the flips of coins based on the rooms people went into, and the biases of those coins.**

(ii - unfairness is not clear and may not be fixable, even in theory): An example might be a complex deep learning model used for predictive policing, which predicts where crimes are likely to occur. Such systems can inadvertently reinforce existing police biases if they are trained on data from police records, which may reflect biased policing practices rather than actual crime rates. The complexity and lack of transparency in how these models make decisions can make it very difficult, if not impossible, to identify and fix potential fairness issues, especially since the underlying social issues influencing crime are complex and multifaceted.

**( ) exactly right, (X) partly right, or ( ) wrong? (check one) If not exactly right, explain why:**

The example that Chat GPT gives is correct (judicial AIs) but the reasoning is wrong. Using algorithms to determine things like recidivism rates has the issue that any errors at all would be unethical, even if spread evenly among all groups.

**III. Social Information Technologies**

5. What is the key difference between epistemic bubbles and echo chambers? Then propose two remedies for echo chambers.

(i - key difference): Epistemic Bubbles: These are informational voids where certain viewpoints or information sources are simply not present. However, in an epistemic bubble, there's no inherent mechanism to discredit alternate viewpoints; they are just absent. It's more about limited exposure rather than active discrediting. Echo Chambers: In contrast, echo chambers not only lack exposure to differing viewpoints but also include an active component of discrediting or dismissing outside viewpoints. In echo chambers, alternate perspectives are known but are actively discredited, leading to a more entrenched and self-reinforcing belief system.

**(X) exactly right, ( ) partly right, or ( ) wrong? (check one) If not exactly right, explain why:**

(ii – two remedies): Promoting Media Literacy and Critical Thinking: Educating people on how to critically evaluate information sources can help them recognize echo chambers and the techniques used within them to discredit outside sources. This includes understanding the nature of bias in media, the importance of considering multiple perspectives, and the skills to analyze the credibility of different information sources.

Encouraging Exposure to Diverse Viewpoints: Actively promoting platforms or initiatives that expose individuals to a wide range of perspectives can help mitigate the echo chamber effect. This could involve creating spaces (both online and offline) where open, respectful dialogue across differing viewpoints is encouraged. Additionally, algorithmic changes in social media feeds to include a more diverse range of information sources can also help in reducing the echo chamber effect.

**( ) exactly right, ( ) partly right, or (X) wrong? (check one) If not exactly right, explain why:**

Exposure to Diverse Viewpoints while in an echo chamber may actually make it worse, along with media literacy which the echo chamber will already view as faulty and fake news. The two potential solutions are a complete “reboot” of trust with institutions, and then the second solution is an effort to rebuild the trust outside of the echo chamber.

6. What is “fake” about fake news? What are the two ways that fake news undermines social knowledge?

(what is fake about it):

What is fake about fake news is its deliberate construction of falsehoods or misleading content, often designed to mimic the appearance and style of credible journalism. The "fakeness" is not just in the inaccuracy of the content, but also in the intent to deceive or manipulate public opinion for various motives, including political or financial gain.

**() exactly right, (X) partly right, or ( ) wrong? (check one) If not exactly right, explain why:**

While what chat gpt said is partly right, fake news is deliberately trying to deceive about its source, not always its content though. It is trying to seem as if it comes from a credible and reliable source, while also pushing ideals or agendas usually.

(two ways):

1. Erosion of Trust: Fake news erodes public trust in legitimate news sources. As people encounter more fake news, it becomes challenging to distinguish between credible and false information, leading to a general skepticism and cynicism towards all news media.

2. Fragmentation of Shared Reality: Fake news contributes to the fragmentation of a shared understanding of reality. It creates parallel narratives within society, where different groups hold onto conflicting versions of truth, undermining the possibility of a common ground necessary for public discourse and decision-making.

**( ) exactly right, (X) partly right, or ( ) wrong? (check one) If not exactly right, explain why:**

Chat GPT is correct in that fake news erodes trust in epistemic institutions, which affects everyone trying to get legitimate news. The second reason is that it causes many more people to have completely false beliefs, harming their knowledge.

7. Identify two of the most troubling ethical issues that the Cambridge Analytica case raised. Which issue was the most troubling?

(i—two troubling issues) Misuse of Personal Data Without Consent: The collection and use of personal data from millions of Facebook users without their explicit consent was a significant breach of privacy. Users were unaware that their data, as well as data from their network of friends, was being harvested and used for purposes they had not agreed to. This misuse highlighted the vulnerabilities in data privacy and the lack of control users have over their personal information on social media platforms.

Manipulation of Democratic Processes: Cambridge Analytica used the data to create targeted political advertising, influencing voter behavior in significant political events like the U.S. Presidential Election and the Brexit referendum. This manipulation of democratic processes is deeply concerning, as it challenges the integrity of elections and undermines the principles of democratic decision-making.

**(X) exactly right, ( ) partly right, or ( ) wrong? (check one) If not exactly right, explain why:**

(ii—most troubling issue) Among these, the most troubling issue is arguably the manipulation of democratic processes. While the misuse of personal data is a severe breach of privacy, the potential impact of influencing democratic elections has far-reaching consequences. It not only affects the individuals whose data was misused but also the broader democratic process and public trust in electoral systems. This manipulation represents a direct threat to the foundational principles of democracy, making it a particularly grave concern.

**(X) exactly right, ( ) partly right, or ( ) wrong? (check one) If not exactly right, explain why:**

**IV. Value Alignment**

8. What is wrong with the “standard model” in AI system design, according to Stuart Russell? What new paradigm does Russell propose, instead?

(i – problem with the standard model): The "standard model" in AI, as critiqued by Russell, is based on the concept of building systems that optimize a fixed objective function given to them by human designers. The problem with this approach is that it assumes humans can perfectly specify their objectives and foresee all future scenarios and consequences of the AI's actions. This assumption is often flawed, leading to what is known as the "King Midas" problem, where the AI system achieves the specified objective but in a way that is harmful or unintended. For instance, an AI designed to maximize production in a factory might do so at the cost of worker safety or environmental sustainability.

**(X) exactly right, ( ) partly right, or ( ) wrong? (check one) If not exactly right, explain why:**

(ii – Russell’s new paradigm): To address these limitations, Stuart Russell proposes a new paradigm centered around the principle of "provably beneficial AI." In this paradigm, AI systems are designed not to pursue a fixed objective, but to continuously learn and adapt their objectives based on human preferences and values. This approach involves creating AI systems that are uncertain about the true objectives they should pursue and are designed to learn and adapt these objectives by observing human behavior and feedback. The goal is to develop AI systems that are inherently safe and aligned with human values, even as those values evolve over time. This paradigm shift emphasizes the importance of AI systems being able to understand and adapt to the complexities and nuances of human values and ethics.

**( ) exactly right, (X) partly right, or ( ) wrong? (check one) If not exactly right, explain why:**

**The only issue that I saw was that the AI is designed to maximize the realization of human preferences, not just adapt to the preferences of humans. Other than that, it is an accurate description.**

9. Identify and describe two of the risks that Bender et. al (2021) argue are amplified by the fact that an LLM’s answers to questions seem to exhibit communicative intent and genuine understanding.

(i) Misinterpretation of Outputs as Genuine Understanding: The sophisticated output of LLMs can easily be misconstrued as genuine understanding or comprehension of the content. This risk is significant because users might trust and act upon the information provided by the LLM without critically evaluating its accuracy or relevance. This trust is particularly concerning in scenarios where LLMs generate plausible but incorrect or misleading answers, which users might accept as truth due to the perceived understanding of the model.

**( ) exactly right, () partly right, or (X) wrong? (check one) If not exactly right, explain why:**

This was not discussed by Bender et. al, but one issue discussed by them was the issue of biases being amplified over generations of models. When models already produce biased output, this output can then become input for the next generation of models amplifying the bias further.

(ii) Anthropomorphism and Overreliance: The seemingly intentional communication by LLMs can lead to anthropomorphism, where users attribute human-like intentions, emotions, or thoughts to the model. This perception can lead to overreliance on LLMs for tasks they are not suited for, such as making critical decisions, providing emotional support, or handling complex moral and ethical queries. The risk here is that users may overestimate the capabilities of the LLM and undervalue human judgment and expertise, potentially leading to detrimental outcomes in various applications.

**( ) exactly right, ( ) partly right, or (X) wrong? (check one) If not exactly right, explain why:**

This was also not discussed, but the third issue discussed in the paper is the issue of GPT models becoming bots that can spam and spread misinformation very well, because of their ability to understand and seem coherent.