Answer Key

1 Main Part

50.0 points · 19 questions

19/19 Grading Scheme

a Which of the following are real articles from the Universal Declaration of Human Rights (UDHR)? Select all that apply.

2.0 points · Multiple choice · 4 alternatives

Everyone has the right to freedom of movement and residence within the borders of each state.	1.0 point
Everyone has the right to work, to free choice of employment, to just and favourable conditions of work and to protection against unemployment.	1.0 point
Everyone has the right to access free healthcare and medical services provided by the state.	-1.0 points
Everyone has the right to a clean and healthy environment, free from pollution and harmful substances.	-1.0 points

b Which of the following are among the **four principles** of ethics, developed on the basis of the Belmont Report? Select all that apply.

2.0 points · Multiple choice · 4 alternatives

Explicability	-1.0 points
Justice	1.0 point
Respect for persons	1.0 point
Responsibility	-1.0 points

c Suppose we want to avoid coercing or manipulating other people to do what is good for them. What kinds of **paternalism** are there that get people to do what is good for them, without coercion or manipulation? You can either label or describe them briefly.

2.0 points · Open · 1/5 Page

2.0 points · Open · 1/5 Page

+1 point
Libertarian paternalism, soft paternalism, weak paternalism

+1 point
Libertarian paternalism, soft paternalism, weak paternalism

d In their article "Artificial Ethics Assistants", O'Neill et al. focus on ways in which AI could be used to help humans become more ethical. Which specific ways of helping humans be more ethical do they focus on?

2.0 points · Multiple choice · 4 alternatives

Avoidance of decisions based on poor information or biases	1.0 point
Improvement of one's understanding of one's core value commitments	1.0 point
Improvement of one's understanding of human virtue	-1.0 points
Improvement of one's ability to influence other people to do what is right	-1.0 points

e In the chapter "Logic and Conversation", H.P. Grice describes the phenomenon of "conversational implicature" and contrasts it with "conventional implicature". He defines conversational implicature as a situation in which the speaker says p but implicates q, and where the speaker "is presumed to be observing the conversational maxims, or at least the Cooperative Principle" and the supposition that q "is required in order to make his saying p consistent with this presumption."

Is it more difficult to get a large language model to reply appropriately to conversational implicature than it is to get it to reply appropriately to conventional implicature? Why or why not?

4.0 points · Open · 3/5 Page

e In the chapter "Logic and Conversation", H.P. Grice describes the phenomenon of "conversational implicature" and contrasts it with "conventional implicature". He defines conversational implicature as a situation in which the speaker says p but implicates q, and where the speaker "is presumed to be observing the conversational maxims, or at least the Cooperative Principle" and the supposition that q "is required in order to make his saying p consistent with this presumption."

Is it more difficult to get a large language model to reply appropriately to conversational implicature than it is to get it to reply appropriately to conventional implicature? Why or why not?

4.0 points · Open · 3/5 Page

+1 point
Def of conversational implicature requires awareness of context to recover speaker's intended meaning through logical inference. (Also acceptable: argument in terms of maxims)

+1 point
Def. of conventional implicature does not require awareness of context or use of logic, it is a matter of conventional (though not literal) meaning.

+1 point
LLMs cannot learn context from their source data, and they cannot make appropriate logical inference to respond without the context.

+1 point
LLMs can learn conventional associations between p & q because these will be represented in source data texts.

f According to Goodall's argument, why is it necessary to program ethics into automated vehicles? Select all that apply.

2.0 points · Multiple choice · 4 alternatives

There is no way to encode human morals into software.

Automated vehicles will be able to avoid all crashes with perfect sensors and algorithms.

Automated vehicles will almost certainly crash, even in ideal conditions.

Automated vehicles can predict various crash trajectory alternatives and select a path with the lowest damage or likelihood of collision.

All students awarded 2 points

g Write a short statement reflecting on whether Awad et al.'s recommended approach to programming moral machines, as discussed in their article "Crowdsourcing Moral Machines," is consistent with Goodall's plan for introducing moral machines. In your response, consider the methodologies and principles each author advocates for. Use the following quote from Goodall's article to help frame your reflection:
"Although artificial intelligence approaches allow computers to learn human ethics without the need for humans to perform the difficult task of articulating ethics as code, they produce actions that cannot be justified or explained in an understandable way."

4.0 points · Open · 1/2 Page

+1 point
Awad et al basic methodology (role of social science in collecting public views specific to a context)

+1 point
Goodall basic methodology (three stages: logic, ML/AI, XAI)

+1 point
Sophisticated understanding of Awal and or Goodall. Goodall: "The neural network could be trained on a combination of simulation and recordings of crashes and near crashes, with human feedback on the ethical response." Awal: "We very much agree that regulations of ethical trade-offs should be left to policy experts, rather than resolved by referendum. But we also believe that policy experts will best serve the public interest when they are well informed about citizens' preferences, regardless of whether they ultimately decide to accommodate these preferences

+1 point
Answers question by slotting Awal into Goodall's view somewhere, or (with correct details) explaining why impossible.

h Please fill in the blanks in the statements below using one selected word or phrase per blank, from the options given.

4.0 points · Free formatted question

+1 point
system complexity

+1 point
prior knowledge

+1 point
abilities

+1 point
socio-technical

i Within the theory of Meaningful Human Control, what example can be used to show that guidance control is sufficient for moral responsibility in some cases, even if there is no regulatory control? You can simply name or briefly describe the example, you do not need an extensive explanation.

2.0 points · Open · 1/5 Page

+2 points
Dual control vehicle with a trainer and a learner

j What are the two main elements required for Meaningful Human Control within Santoni di Sio & van den Hoven's account?

2.0 points · Multiple choice · 4 alternatives

Tracking condition	1.0 point
Knowledge condition	-1.0 points
Responsibility condition	-1.0 points
Tracing condition	1.0 point

k Within the theory of Human-Centered AI, in which kinds of technical systems should there be full automation? You can simply name the main factor or a clear example, you do not need to give an explanation.

2.0 points · Open · 1/5 Page

+2 points
Systems that require fast response, such as an airbag or a pacemaker

l According to Nickel, why does Human-Centered AI (HCAI) fail to guarantee trustworthiness of AI systems?

2.0 points · Multiple choice · 4 alternatives

HCAI guarantees responsibility, but this reduces trustworthiness.	0.0 points
HCAI is about control rather than authority, and trustworthiness depends on having authority.	2.0 points
HCAI undermines responsibility because it makes it unclear whether the user or the system is in control.	0.0 points
HCAI depends on a notion of human mastery that is vague and subjective.	0.0 points

m Which of the following statements best expresses how Vallor's use of the term "deskilling" in her article "Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character" differs from how Sambivasan & Veeraraghavan use the term in their article "The Deskilling of Domain Expertise in AI development"

2.0 points · Multiple choice · 4 alternatives

Vallor uses "deskilling" to describe the loss of moral and practical skills due to automation, while Sambasivan & Veeraraghavan use it to describe the reduction of domain experts to mere data collectors in AI development.	2.0 points
Vallor uses "deskilling" to highlight the economic devaluation of traditional skills, whereas Sambasivan & Veeraraghavan use it to emphasize the lack of technical skills needed in AI development.	0.0 points
Vallor uses "deskilling" to argue for the complete elimination of human skills by machines, while Sambasivan & Veeraraghavan use it to discuss the reduced need for expensive, expert employees in automated data collection practices.	0.0 points
Vallor uses "deskilling" to focus on the enhancement of moral virtues brought about through technological advancements, whereas Sambasivan & Veeraraghavan use it to criticize the ethical implications of AI development.	0.0 points

n Consider the following incomplete argument:

1) Skill displacement will occur if the technological autonomy ideal is pursued.
2) _____[MISSING PREMISE]_____.
3) We should avoid significant loss of value within lives, unless it is offset by comparative benefits.
4) We can achieve similar benefits by pursuing human-centered AI as by pursuing the technological autonomy ideal, but without skill displacement.
5) Therefore, we should not pursue the technological autonomy ideal.

What proposition should premise (2) specify in order to make this argument deductively valid? (A single sentence is a sufficient answer.)

2.0 points · Open · 1/5 Page

+2 points
Skill displacement involves significant loss of value within lives.

+1 point
Skill is a value

o According to McCormack, which of the following features of commercially available generative AI software for artistic practices decreases the possibility of **technical mastery**? Choose all that apply.

2.0 points · Multiple choice · 4 alternatives

These systems are undemocratic and difficult to access for artists, making it hard to master them.	-1.0 points
GenAI systems often change, preventing the user from gaining mastery over the tools they use.	1.0 point
The companies behind them emphasize efficiency and cost-effectiveness over technical mastery.	1.0 point
These systems emphasize literal descriptions of images over form and performance, which are required for mastery.	0.0 points

p Give a concrete example of a contemporary artistic practice that uses Generative AI as an object of critical inquiry? Briefly explain what makes it an example.

3.0 points · Open · 1/2 Page

+1.5 points
Example, e.g., ImageNet, description

+1.5 points
Artist uses GenAI as a tool, but also to challenge and reflect on aspects of it that are problematic and bring them to public attention.

q What is Chalmers' main argument for the likelihood of a singularity? Write it as a four-step, valid argument.

3.0 points · Open · 1/2 Page

+1 point
(Soon) there will be AI

+1 point
If there is AI, there will be AI+

+1 point
If there is AI+, there will be AI++ (singularity)

r Which three assumptions are together supposed to create a catastrophic risk related to the singularity, according to authors such as Chalmers and Bostrom? The first assumption is superintelligence. What are the other two assumptions?

2.0 points · Open · 1/5 Page

+1 point
Orthogonality

+1 point
Instrumental convergence

s Answer either RESEARCH ETHICS CASE or PRACTICE CHATBOT here.

6.0 points · Open · 4/5 Page

Model answer
diagnoses problem correctly (what's wrong with socially desirable answers or oversimplicity/ giveaways for the target context)
prompt engineering or alternate (named) approach
details from own assignment