Contents

[OUTLINE 31](#_Toc197505821)

[Final Essay 47](#_Toc197505822)

**Kant Notes**

* Proposition 1: An action must be done from duty to have moral worth

1 of the three greatest Western philosophers

Goodwill underlined – Nothing else is good without qualification

Gifts of nature, such as intelligence, can become extremely bad depending on use.

Gifts of fortune impact happiness but may also negatively affect it.

Being worthy of happiness means having goodwill and earning it

Goodwill is good because its virtue is good itself without any attainment

Goodwill means having the intention to act morally without any reward, simply cuz it’s the right thing to do. It will be good no matter the result of the action.

Doing something because you have to, such as because of rules vs because of your inclination or desire to (Duty vs goodwill)

Duty isn’t about following emotions or using bias to make decisions. However, it comes from reason.

Kant uses the example of a shopkeeper and a child. The dealer must charge a fair price even though the child doesn’t know the value, if the dealer charges fairly, it’s because it is morally right. However, even if he does charge fairly, it doesn’t always mean that he was acting from duty, he may be acting out of self-interest to maintain a reputation.

The dealer’s motivation may be on the business or avoiding consequences rather than morals.

If the customer is an adult, the dealer is likelier to act out of duty. However, it would be morally wrong to act fairly only for personal gain or reputation.

True moral worth comes from acting out of duty because it means that even though there may not be a gain, you follow the moral law rather than being persuaded by emotions and desires.

Moral values of actions are determined by motivation, however, if you don’t think about the motivation of the action and simply do it because it’s morally correct, then the action is also morally correct, no matter the result, as long as the intention is to do the right thing.

It’s selfish to charge a fair price to protect reputation (self-serve)

It’s morally correct to act from duty regardless of personal gain.

It’s also a duty to live life in a way where it is not thrown away. It’s also an inclination to do so; people want to have a good life in general. People may cherish their life morally, but it isn’t because it’s moral. They must be motivated by morals to make their life act from duty. People may act morally, but the motivation may not be moral. It’s all about intention.

If a person lives their life with duty, without any personal gain for loving it, fearing it, etc, then he’s a morally right person with a morally right life.

\*\*\*\*Everything comes down to the intention or the reason

Satisfaction and kindness have no true moral worth but are on a level with inclinations. It has no moral worth because it was from inclination, such as praise or encouragement, but not from just duty.

Gifts of nature, such as patience, still impact moral worth as long as actions are from duty and not inclination. If the situation were the same, with or without patience, it has moral value.

It’s incredibly hard to do something only out of duty; everything has reasons and inclinations, however, even with that desire, if the happiness did not influence his will and was not an active role in the choice made, then it is still from duty and not inclination.

Happiness motivates people, and it is human nature to act in ways that lead to happiness or well-being. However, happiness alone is not enough to make something morally right. If it’s morally right, it should occur whether it creates happiness or not.

People love one another, not romantically but in a sense of action. Love can’t be commanded by duty; you can’t force yourself to love. However, you can act lovingly to respect duty. For example, kindness and helping others or caring is out of respect, not love, making this moral and is practical love, not pathological.

However, if love is seen as an emotional feeling, an inclination, then the actions they commit with those feelings don’t have any moral worth because it is based on inclination and solely because you love them, not because it is what is right.

True moral love comes from respect and treating others correctly but not for selfish purposes. Even in romance and in marriages, love may only be moral if it contains mutual respect. If duties in marriage rely on the feelings between 2 individuals, then it is not because of the marital duties to one another nor moral duties, but solely on love.

* Proposition 2: Duty derives from moral worth, not from purpose attained but the principle behind it or the rule, rather than the result or anything else.

The moral value of actions isn’t determined by the consequences it has; it’s rooted simply in the principle or intention. Ends and springs of the will = goal and motivations. Even if there are worthy and noble purposes or motivators, as long as they impact an action, they deprive it of moral worth. If the action is done simply because it has a specific consequence or goal achieved or ended, then it is out of self-interest and not moral duty.

Moral worth relies on the principle of will itself.

The priori principle is the formal, rational, universal law.

The Posteriori principle is the material desires motivating actions, such as happiness, pleasure, etc.

To make something morally right, it must only be determined by formal principles and have zero impact from the material principle.

* Proposition 3: Consequence of Proposition 1 and 2. Duty is the necessity of acting with respect for the law. Inclinations may serve as an effect of the action. However, I can’t have respect for it simply because of the fact it has an effect, it must also be aligned with duty and will. I can’t have respect for inclination but can approve it and acknowledge it.

As long as inclination doesn’t overpower a decision and actions are simply out of duty and following the laws and rules of morals, it is morally correct.

Duy is also not optional; people are obligated to do it and have responsibilities tied to it, regardless of inclinations or desires. Moral actions aren’t optional and are rational.

For example, duty says to tell the truth because it is the right thing to do, it may have benefits, but those benefits shouldn’t be the purpose behind honesty. You tell the truth simply because it’s right.

Kant refers.

“May I, when in distress, make a promise with the intention not to keep it?”

He responds that even if in distress, it’s morally wrong to make a promise with the intent of not keeping it because a promise is a moral duty and is bound to be kept. If you don’t keep it, then its meaning is lost, and it no longer holds any value as a promise. Duty is also not conditional; therefore, even in struggle, the rules must still be respected. Also, by breaking the promise, the other person is treated as a means to your own to relieve distress rather than respecting them. Breaking a promise dishonors trust and mutual respect and must be respected no matter what.

Killing yourself Is morally wrong because it contradicts the universal law of nature, and destroying life to make life better would contradict and can’t act as a system of nature. Duty is to preserve life, and killing yourself contradicts it.

Making promises that can’t be kept contradicts universal law because they are solely for self-benefit and also break honesty as a moral.

Neglecting natural gifts in favor of indulgence fails the test because it can’t be universal without leading to moral collapse. It undermines the preservation of life and also the moral duty to society. The person must cultivate their gift even though they prefer not to.

Having money but not helping others with it would contradict moral law as it neglects the needs of others. People also must help others in need, even if there is no inclination to. The person shouldn’t give up all their wealth but insist on helping others, or it fails moral duties.

To have an action be a universal law, it should not be contradictory to the three principles.

The worth of anything acquired by action is conditional.

Act as to treat humanity, whether yourself or another’s case, as an end, never only as a means.

Harm and lies violate the principle of humanity. Must also harmonize with nature, and neglecting them may harm humanity. People also have duties to others and society, and failing to follow them may harm humanity.

Proposition 1 focuses on motive, and Proposition 2 focuses on the rule of duty or moral concept.

Goodwill refers to the action of making decisions according to duty or moral obligation.

It refers to decisions made solely based on moral rights.

**Bentham**

Ch 1. Of the Principle of Utility

1. Intensity
2. Duration
3. Certainty
4. Propinquity/remoteness
5. Fecundity
6. Purity
7. Extent

Pain and pleasure point out what to do and determine what we do. It does so in the standard of right and wrong. The chain of causes and effects is also quickened. They impact what we do and say and everything else.

The principle of utility argues the unavoidable truth about human behavior; anything attempting to undermine it is misguided and relies on emotion if it isn’t rational.

2. The principle of utility is the foundation of the present work. It approves or disapproves all actions according to the tendency it appears to create or remove happiness in question. This includes actions of groups like the government.

3. Property in any object causing benefit to prevent any pain is also considered.

4. Community is composed of members, the interest is the sum of interests (majority).

5. Useless to talk or communicate without understanding. It should only be talked about to promote or help.

6. Action should be for the greater good of the whole

7. Actions should cause more happiness than pain

8. This principle should maximize happiness, and this should apply to laws.

9. People supporting this principle should make decisions about right and wrong based on how the action contributes to collective happiness

10. Words hold meaning, including ought, right, wrong, only in the context of actions

11. Principle can’t be proven, but it isn’t needed because it’s justified and is, in general, practical.

12. People won’t always follow this because of human flaws, however, consistency and using this principle as a guide is the goal.

13. The principle can’t be wrong, only the actions may be misapplied or misinterpreted.

14. People can’t disprove the principle but may still reject it. He suggests a process to help understand

* Consider whether they want to regard it in its entirety or just partial
* Ask if they would act and judge with zero principle, how would they judge actions?
* Alternative principles should be examined
* Should feelings apply to everyone? Can they make decisions for everyone when thinking about themselves only?
* If people believe their feelings may serve as the standard for all, they’re imposing their views on everyone.
* If everyone followed their morals, nothing would be considered right or wrong bc everyone’s opinions all matter and could contradict.
* If people believe morals should be with reflection, then if their reflection supports the principle, they are still supporting the principle.
* Compromise between principles may influence moral judgment, where is the line drawn?
* Partial adoption of principle requires reflection and a reason to reject the rest

Chapter 4

Value of pleasure and pain and measuring

1. The principle holds pleasure and avoidance of pain at the end of the rule, it’s important to its value and what we work with.
2. The value of pleasure of pain is weighed according to intensity, duration, certainty or uncertainty, and propinquity of remoteness
3. When considering pain or pleasure based on any act and how it’s produced, it’s weighed by fecundity or chance of it being followed by the same kind of sensations: pleasure leads to pleasure, and pain leads to pain. It is also weighed by purity, the chance of not being followed by the same sensation, pleasure than pain, or pain than pleasure. These are deemed properties of pleasure and pain itself, only of the act or event causing that pleasure or pain.
4. To several people, it will be weighed on intensity, duration, certainty of uncertainty, propinquity/remoteness, fecundity, purity, wit, and extent – the number of people it affects.
5. To find the exact amount, begin with anyone whose interest seems most immediately impacted by it and take account of the value of each distinguishable pleasure in the first instance, along with each pain, pleasure after the first (fecundity of first pleasure, and impurity of first pain). Should also check pain after first (fecundity of first pain and impurity of first pleasure). Summing up all values, the balance, if on the pleasure side, means good in general and respects the interest of the person. If it is on the side of the bad values, then it’s pain overall. Take an account of several people with interests and concerns and repeat. Sum up the values.
6. This process won’t be used on everything, but it should always be considered.
7. It can apply to anything, in whatever shape.
8. Combine or compromise between one's moral principles and utilitarianism.

**Mill Utilitarianism**



Chapter 1

The question concerning “summum bonum,” or what is the same thing, concerns the foundation of morals.

People question this the most, and there has been little to no progress

Morality is practical and should be logically derived from understanding

This makes it hard to agree on

Science has concepts and truth values, making it provable, unlike philosophy.

Science still makes predictions and stuff, but with philosophy, it’s more problematic because it is meant to guide action.

The core issue – Actions must be driven by clear and defined ends, contrasting scientific inquiry.

People believe in general moral principles but not specific judgments of right and wrong.

Happiness plays a role in moral judgment. Maximize happiness

Kant argues action should be guided by universal rules, but critics claim he fails to demonstrate logical contradictions in immoral behavior that could result from universalization.

Kant fails to demonstrate why immoral actions would logically contradict universal application.

Many misconceptions of utilitarianism come from a misinterpretation

Chapter 2

Misconceptions – people assume utility as pleasure or absence of pain. Utilitarianism goes over several ranges of pleasure, not just simple. Utility is misapplied.

Greatest Happiness Principle – actions are right insofar as they promote happiness and wrong to the extent that they cause unhappiness. Happiness is pleasure.

People believe humans aren’t only capable of basic pleasure. Comparing man to animals, people believe men are much more intelligent intellectually, morally, and emotionally.

Unreasonable for utilitarianism to only account for the quantity of pleasures. Quality also plays a role. Judgment is based on understanding and appreciating both types of pleasure, higher(intellectual, emotional) and lower (sensory, physical)

Higher is from reason, creativity, and ethical behavior and is more quality.

People wish for more, not less. No human would choose to be animal-like

People may choose lower pleasures due to weakness of character or circumstantial demands. This is not because they enjoy those pleasures more but because they have lost the ability to appreciate or enjoy better ones. This could be cuz the environment or the absence of appropriate intellectual stimulation

Goal – Achieve the greatest amount of happiness, quality, and quantity for the greatest number of people.

The pursuit of happiness is rational and meaningful – the goal is to promote AS MUCH happiness as possible, not true happiness.

Greatest happiness principle – you ought to maximize pleasure for the most sentient beings.

**Philippa Foot**

Act utilitarianism is wildly at variance with firmly held moral convictions.

Rule utilitarianism strikes most people as an unstable compromise

Business of exorcism – Something drives us towards utilitarianism. There must be something wrong about it, though, because why is the theory unacceptable?

The issue with utilitarianism is the consequence. It makes it compelling.

Utilitarianism vs consequentialism

Consequentialism simply says it is by total outcome and by whole formed by action and consequence, action is judged right or wrong.

Consequentialist theory of ethics identifies goodness state of affairs on rightness or goodness based on positive, productive relationships.

Utilitarianism consists of consequentialism together with the identification of the best state of affairs with the state containing the most happiness, pleasure, or satisfaction.

Utilitarianism = consequentialism

The main problem – welfarism or sum ranking of welfare

Good is not auto-increased by an increase in pleasure but by non-malicious pleasure

Principles have been introduced so actions benefiting the rich more than harming the poor have no longer been judged morally worthy

Criteria constantly modified

The distribution principles and discounting of certain pleasures and pains did not help with the problems of wrongness.

Rights had to be looked after in a way that was impossible with the modified versions.

Goal rights systems should be considered – the idea that respecting or violating rights should be counted as a good or evil in the evaluation of states of affairs.

This would solve some problems.

Fails to deal with examples of actions most would say are wrong

Ex. Evil person threatening to torture unless we kill or torture one. The choice between more killings or torturing and less. Consequentialists will say that we are justified in killing one and morally obliged to. This is wrong, though

Good and bad are very opinion-based- one may say it’s good, and another says it’s bad, but it may be true in both cases

Good and bad are subjective to the group interested

Moral point of view

Rational morality is one in which rightness and wrongness are judged by relation to certain outcomes – maximization of social utility.

Moral pov – fortiori – based state of affairs from moral pov

Many things are inconsistent with morality

It cant be enough to say we DO have rules of justice, we also question whether we should and why

Philippa critiques utilitarianism for lack of attention to individual and moral rules and rights and the character of moral agents.

Low order pleasure

High order pleasure

Why do people choose lower over higher-order pleasure?

1. Ignorance
2. Fatigue
3. Loss of capacity

Virtue ethics

Eudaimonia

* Charity
* Prudence
* Phronesis
* Wisdom
* Courage
* Beneficence
* Justice

Deficiency

Happy mean (courage)

Excess

**Hobbes**

The weakest parts of people are extremely strong

Other than strength, there is something else equal for men – wisdom

People are equal

Everyone is seeking their own interests and looking into defense and offense actions

This is natural and used for survival

Men have no pleasure and a great deal of grief when there is no power to over-awe them.

People want equal respect based on their own worth and may force others to value them if they feel disrespected

People’s actions are examples for others and it can lead to destruction

If people are left to their own desires, it will create conflict because everyone wants respect and that leads to competition

3 principal causes of quarrel

* Competition - invade for gain
  + Violence to make themselves master
* Diffidence – invade for safety
  + Violence to defend
* Glory – invade for reputation
  + For trifles, a smile, different opinion, sign of value, reflection

Leads to condition of war – every man against another

Time is war – war doesn’t always mean fighting but means no peace

When people are enemies to another, then no place for industry, no culture of earth, no navigation, etc

People are on guard, accusing others and having assumptions – distrust

Desires and passions for self-protection aren’t wrong. Only wrong if they violate a law

Laws are needed to know limits

War has always been around but not because of jealousy. They went to war because of misery, industry of subjects, laws, etc

No common power = no law = no justice

Force and fraud

Justice and injustice are not part of the body or mind.

They are qualities relating to men in society, not with others.

Passions inclining men to peace: fear of death, desire of things needed for living, hope by industry to obtain

Laws of nature

1st and second natural laws and of contracts

Right of nature – naturale – liberty to use own power for himself or how he wants. Free will

Liberty is the absence of external impediments removing power or preventing judgement

Law of nature – lex naturalis – precept or general rule with reason to forbid something that destructs his life or takes away the means of preserving

Everyone has a right to every thing, even to another’s body – no one is safe from seeking what they want

No security to them as long as this right is upheld – someone can always threaten any part of you

1st law – Every person should seek peace whenever possible

2nd law – If peace is unachieveable then you have the right to use all means available to defense yourself and seek advantage

Self defense and the right to protect urself – Can use any means necessary – Self defense

There must be mutual agreement between people to limit natural rights for safety and self-preservation which is through law

Removing their abilities to interface with another’s is an exercise of their own right

Every individual has the right to everything – renouncing their right just means they no longer hinder others from exercising their natural right to that thing

No new rights are created

Removes obstruction for others – everyone can freely enjoy own rights when giving up some rights

Renounce – give up

Transfer – give – grant a right to someone else

Obligated to not hinder the person holding that right from enjoying it – can’t interfere with other’s enjoyment of right

Duty to honor voluntary act of renouncing or transferring

Injustice vs absurdity – absurd to contradict self in argument and unjust to undo voluntary action

Bonds of obligation – social contracts may not hold up because of strength but because of fear of consequences

Voluntary acts are good to self too

Limits – certain rights that can’t be transferred or renounced because they are needed

Ex. Right to defend self

Inalienable rights – right to life and self defense

Suffering harm does not offer benefit to the individual

Goal of renouncing rights – secure person and life

Mutual transferring of right – contract

Delivery vs transferring of the thing

Act of giving up ownership = transferring right to something

Delivering thing itself – giving actual object or possession to another

Pacts made to keep promises, violation of faith

Do not do something to another that yu wouldn’t do to yourself

Laws of nature are immutable, injustice, ingratitude arrogance, pride, iniquity, acception of people, and etc can never be lawful

Moral philosophy – science of what is good and evil

Laws of nature – true moral philosophy

Commands – laws

**Class notes**

Phronesis – application of virtues and knowledge to world

1588-1679 Hobbes

Hobbes lived through English civil war

Leviathan – book – political theories and ethical theories

Frontispiece

Keep in mind – Civil war and frontispiece

Top portion

* Crown
* Sword
* King is in control of means of order (dictates)
* Scepter – flaws in order and justice
* Hobbes should also be in charge of legal authority
* Body of king is made of little people
* The head is normal – conscious and deliberate
* Need something after political process that is whole and divine
* Head – sovereign
* Body – people
* Need a power center – a ruler but not independent of the people (hence the body of people) but the head must be independent (head)
* rules the power of justice, law, and warfare
* Hobbes believes in the monarchy and is royalist

Kant – reason

Mill – pleasure

Virtue ethics – human flourishing

Hobbes – nature

The differences in people are minor, no one is completely superior over another

People are equal, but it doesn’t feel that way. Compare self to Lebron James as an example in skill, height, physical, etc

Everyone is small specs

We are all VULNERABLE people

State of nature – time before government, law, punishment, order, etc (Before morality)

State of nature = state of war – any time when peace is not assured

No one has a right to anything but everyone also has a right to it

* Max rights
* No one wants to be in state of nature though

Right – a liberty to do

Law – Something that determines and binds

1st law of nature – you ought to endeavor towards peace

Ethical egoism

2nd law of nature -derived from first – lay down rights to preserve peace (transfer or renounce)

**Right of nature – the one inalienable right -** Immune to contracts – your right to defend your life

No rights if you’re dead

Explicit and implicit consent

**The solace of Oblivion**

Right to be forgotten

Catsouras goal – remove information from Internet

Why: People want something that doesn’t exist– right to be forgotten

Right to prohibit google from linking to items – inadequent, irrelevant, or no longer relevant, or excessive in relation to purpose which processed

Consequence – Google fielded SEVERAL requests for deletion and granted around half

Other search engines have too

Public reaction is critical

* Could undermine press freedom and freedom of speech
* Risk – aggrieved individuals could use decision to hide or suppress information of public importance
* Misguided in principle and unworkable in practice
* Most significant mistake from Course – little value to free expression
* Course outsourcing to Google complicated case-specific decisions is wrong
* Google as court of philosopher king shows lack of understanding in reality

Europe – right to privacy trumps freedom of speech

* Right to privacy as fundamental human right

US – right to be provided protection

* Provide protection for certain categories of information deemed sensitive and impose obligation to disclose unless conditions met

Congress passes laws prohibiting medical info, video store rentals but can be overridden with consent or investigation under law

Americans regard freedom of speech – 1st amendment – Costeja judgement would not work. Public and reported – constitutionally correct so policy can’t be upheld

Communists fought Nazis with ideology on humanism – bring more just and fair society

Ended up becoming a totalitarian surveillance society

Capturing information and using it to further the power of state reintroduced and perfected by society

2 radical ideologies – fascism and communism – both end with tight surveillance states

After fall of communism – laws placed to prevent recurrence of abuses. Law to protect privacy

Pervasive belief – can’t trust anything. Keep to own role and protect rights of individuals

Book – Delete: The Virtue of Forgetting in the Digital Age

Storage in cloud has made information more durable and retrievable

Google is around 90% - does not make sinister use of information at disposal

Don’t know if future governments will honor the trust we put in them to protect information privacy rights

Without right to be forgotten in law, can’t force Google to stop linking

Easy to find web sites even if links are less likely to show up

Convicted criminals who want to move on from the past can’t

Consequences of having criminal record are getting worse because of the Web

Expungement movement – calls for criminal convictions to be sealed or set aside after time

Around 30 states accept some degree of this

Individuals and companies use copyright law to prevent unwanted attention

Ownership and negotiation for copyright

Copyright infringement on YT

Leaked photos – selfies

European Court placed Google in uncomfy position. Google makes info accessible but doesn’t create info.

Notion of Google as passive intermediary in modern info economy is dubious – misleading

Google is the library and brookstore

Google websites aren’t under Google’s control but web searches are

Not about free speech but privacy and digity

Job

Technical – creating software infrastructure so links could be removed – not difficult

Administrative system to intake requests and act on it

Considerations:

Price meets with group twice a week to discuss discussions and maintain standards

Assembled dozens of lawyers, paralegals, etc

Main considerations: individual public or private figure

Link from reputable news source or government web site or individual published or related to political speech and criminal charges

European court ruling applied only to search engines, not social media sites but principles drew concern to other pages

European regulators want to impose more restrictions on Google

People want rules for the Internet

2nd order effects – laws consistent with 1 country but not another

Google – things aren’t found accidentally – small solution

Mill Free speech on liberty 1-9, 19-24, 44-47

Propriety – state of being proper or suitable

Suppose govt is with the person and doesn’t want to exert power of coercion unless in agreement

People who deny the truth assume they are the right one

Important parts for people – religion, party, sector, class in society, etc

Duty of govt and individuals to form truest opinions and carefully and to not impose on others unless absolutely sure they’re right

If sure, it’s wrong to not act on their opinions

Absolute certainty isn’t a thing but assurance is

Use opinions to prevent bad and for the truth and guidance of conduct

Liberty justifies in assuming truth for purpose of action

Wrong opinions usually fall to fact and arguments

Very few facts are able to tell own story without opinitated ocmments

Newtonian philosophy – theory of mechanism of universe

Without questioning things, people wouldn’t have gotten facts or evidence nor assurance of truth

Claims of opinion protected from public attack are rested not so much on truth but importance to society

Opinions matter

Opinions men entertain and feelings cherish

ad misericordiam – pity fallacy – argument based on strong appeal to emotions

Opinions are needed because they lead to people wanting to find reasons and then creating peace

Reticence – quality of being reserved

Greated harm done to those who are not heretics and who are mentally challenged, reason cowed, by fear of heresy (People who don’t believe something because it is widely known)

Heretics - someone who believes or teaches against accepted or official beliefs

Heresy – opinion or doctrine accepted

Freedom of thinking needed

Heterodox – not accepted beliefs

Dead dogma – belief gone unquestioned for so long to such a degree that people don’t know why they believe it

If something is a popular opinion, it should be frequently discussed or it will be dead dogma

Learn reason or ground for opinions

People believe based on things to defend

Some say they were taught this though

Even in natural philosophy, there is always some explanation possible of the same facts

Subjects complicated like morals, religion, politics, social relation, business of life, etc., they aren’t as easily proven with facts

1. If opinion compelled to silence, it could be true, denying this is a mistake
2. Silenced opinion could be false but may contain truth in it
3. Even if it is the whole truth, it will not be accepted unless there is reason behind it or an explanation

Assert opinions

Don’t argue though or misrepresent the wrong opinion

Law and authority have no restraint in opinions – it is the consequence and circumstance of the individual

Overall – free speech is a fundamental right. Controversial and offensive ideas should be expressable, as long as there is no harm

Emotional harm may not justify limiting speech unless it leads to actual violence

Acknowledge that speech can impact emotion which impacts physical things though

Phlogiston – tiny particle

Some materials have it, some don’t

What kind of privacy – Nissenbaum helen – a contextualist approach to online privacy pp 32-38

Theory of privacy – contextual integrity

Key parameters of informational norma are actors (subject, sender, recipient,), attribute (types of info) and transmission principles( constraints)

Transparency and choice

Notice-and-consent or informed consent

One is the popular definition of a right to privacy as a right to control information about oneself.

For many critics, whom I call critical adherents, the fault lies with the ubiquitous  
regime of offering privacy to individuals on a “take it or leave it”

Privacy policies – long, no one wants to read, just accept

The theory of contextual integrity offers a shorter and more systematic path to this point by invoking learned wisdom from mature systems of informational norms that have evolved to accommodate diverse legitimate interests as well as general moral and political principles and context-specific purposes and values.

Online activity is deeply integrated into social life in general and is radically heterogeneous in ways that reflect the heterogeneity of offline experience

The harm principle – you ought not to intervene in anyone’s action, UNLESS it will cause harm to someone else

Paternalism

Reasons for free speech

1. Any silenced opinion might be true
2. Without opposition, your opinion is prejudiced
3. Many opinions still contain partial truth

Nissenbaum Helen 38- 47

Disruptions in media coverage and announcement leads to issues

Online content not always true

Privacy rules imply rules but don’t explicitly say so they are unfollowed

Web vs library similar in that they include pursuit of research, knowledge, and intellectual enrichment

Protecting online consumers vs commercial information

May say net is completely commercial – only care about money

Certain brands of free-market capitalism make it easy to confuse quest for profit with pursuit of internal standard of excellence

Beyond profit – standards are at play and socially important

Consider close analogues based on function and purpose

Content should determine constraints on flow of information

Privacy norms do NOT just protect individuals but also play crucial role in sustaining social institutions

Privacy is more than just personal data. It also creates social norms and shows how and who information should be shown to.

Each place has its own rules

Informed consent – information should only be shared with ways expectations of those involved and individuals have clear understanding of how data used. TRUST

Privacy violates can lead to injustice and disrupt social relationships, creating power imbalance

Understand privacy and violations means going against expected rules in context

Rules should respect norms governing information flow. Different norms healthcare vs social media

Transparency – individuals should be aware of how information is used and understand

Existing laws are outdates or generic

Surveillance may infringe on privacy

Test:

Definition at top, give description of concept

True and false multiple choice

Short answers

4 short essay questions – WHY

Definitions:

* Utilitarianism – Maximize overall happiness. Actions judged on consequence. Choose the one with greatest overall happiness.
* Act/Rule utilitarianism –
* Act – Evaluate action on happiness
* Rule – Follows rules that generally maximize happiness
* Difference: Act focuses on effect of individual actions and rule focuses on effect of type of action
* Categorical imperative – Act only according to maxim to create universal law. If you do something, how does it impact the universe if EVERYONe did it.
* Social contract theory – Individuals can act however they want but to prevent chaos, law and order are in place.
* State of nature – state where everyone can do everything, no laws, complete chaos.
* State of war – state where peace is not around, state of nature causes state of war.
* Felicific calculus – method used to quantify pleasure and pain to determine best course of action. Assess how action impact happiness. Maximize happiness. Weigh pleasure and pain.
* Apriori/aposteriori - Knowledge or reason independent of experience (Apriori). Knowledge or reason based on experience or evidence (Aposteriori).
* Practical imperative – Treats others how you want to be treated. Don’t use others for gain.
* Greatest happiness principle – Right action is the one that maximizes happiness.
* Summon bonum -highest good – ultimate goal – living in life accordance to potential or virtue.
* Virtue ethics – using good characteristics to live life, rather than morality of action
* Dogmatism - believing in something because it’s common belief, even if you can’t justify why you believe it.
* Phronesis - practical wisdom – know the balance between extremes and make good decisions considering rule, context, and consequence.
* Consequentialism – actions based on consequences.
* Harm principle – Actions of individuals should not harm others because then you can cause urself harm too.
* Principle of utility – Choose action that causes greatest happiness for majority.
* First/second law of nature – seek peace whenever possible and giving up rights for peace doesn’t mean you lost rights
* Imperfect/perfect duties – Perfect – Must always be followed. Imperfect – Should be followed whenever possible. Example (perfect) don’t lie. Imperfect (help others)
* Deontology – Act on moral rule, regardless of consequence, simply because it’s the right thing to do, use with universal moral laws.
* Maxim - Rule to follow when choosing how to act. Kant – universalizable without contradiction. If EVERYONE did this, what’s the outcome?
* Given the situation (public park or industry park), how would felicific calculus work? – Assess happiness by looking at benefits of public park and industry park, see the good and bad and then add them up, the one with more points overall wins.
* How do Mill/Bentham differ on measuring pain/pleasure? – Bentham states quantity is important but Mill also talks about quality.
* How does the categorical imperative work for Kant’s 4 examples? -Actions be guided that can be upheld universally . Promise keeping – breaking promise = no trust. Help other bc ignoring = collapsing of society. Self improvement – neglection ruins society. Respect others – undermine dignity if you don’t.
* What’s a right as a law (Hobbes)? – A right is the liberty to do whatever necessary to perverse life, even if it means giving up rights.
* How does Mill’s theory of free speech align with Utilitarianism? (think about his main reason for wanting free speech) – Silencing speech may be harmful because it prevents discovery of truth. Greatest happiness
* Why do people choose a lower order pleasure over higher order pleasures? (Why is the blissful pig/sad Socrates an object, and how is it diffused?) – Ability to obtain higher order pleasure, effort, immediate pleasure.
* In virtue ethics, how do you find the happy mean? - focus on good characteristics to have a meaningful life. Good will eventually lead to happiness.
* Why would telling the truth for money not be a moral action for Kant?- It isn’t done because it’s a good thing, the truth is being told because of the consequence, the money, not because of good will.
* What types of principles are underlying Hobbes’ social contract theory? (ethical egoism) – Give up rights to ensure peace and security – preventing state of nature. Act from motive of self interest
* How would a shortened life impact ethical evaluation in virtue ethics? – It would be viewed based on its motivation, if it’s to avoid suffering, it may be reasonable
* For Philipa Foot, what makes utilitarianism unacceptable? (What does consequentialism have to do with it?) – It goes against morally problematic actions like killing one for the greater good. Goes against individual rights. Morals should consider motives and consequences
* What problems does welfare utilitarianism have (what problems was it trying to solve)? – Improve well being of individuals but can harm minorities because it focuses on the whole. Tried solving what is wrong or right. How do we quantify what is good or bad.
* Why is Bentham in a box? – he wanted his remains used for education
* How does free speech combat dogmatism? – keeps arguments alive and allows opinions to combat and provide reasons, keeping arguments from being believed simply because it’s widely known.
* Why does Mill believe in the harm principle? -People should say anything as long it doesn’t directly hurt others, should protect others
* What role do exemplars play in virtue ethics? (What is the end goal? What problems may arise with this system?) – Be a role model, like the exemplars, but history and changes makes it hard to find a perfect model.
* What is Nissenbaum’s approach to online privacy? (How does she justify it?) – Contextual integrity – privacy depends on the context, respect expectations. Don’t expose in medical field but in regular convo, it’s okay to talk about stuff. Different context, different privacy.
* What might an ethical argument for the right to be forgotten look like? – People have the right to control their personal information and its distribution, if it’s harming others by impacting future jobs and such, they should have the ability to remove it. However, freedom of speech comes into play and is therefore restricted if done so.

Transparency and choice – Failed

Summaries – Terms and conditions are complicated because they are exact. Can’t summarize or will lose information.

Transparency paradox – if you include all info, no one can understand; If you leave out info in summaries, you defeat the purpose

Is there a duty to be a digital minimalist?

Mobile devices are known to cause anxiety, depression, horrible attention sleep, sleep disturbance, and decreased relationship satisfaction.

Failing to do what we need to because of this – moral failure – phone addition

Use Kant – account of duties to self

Digital minimalism – idea that interactions with digital tech should not conflict with ends. Existence of self regarding duties. Complete argument by connecting discussion of self regarding duties t digital minimalism.

Imperfect duty to be intentional about our use of wireless mobile devices – kant focused but not only using kant.

Deal of harm written about addictive behaviors with devices. Separation causes anxiety. Addictions leads to academically worse . Interpersonal relationships ruined from short attention spans, distraction, and unability to be off phones

Devices meant to be addictive

1. Trigger – notification for attention
2. Action – get user to do smth – check page
3. Variable reward – notice, notification, etc
4. Investment – change pfp, make post, motivation to return

Pose threat to autonomy, capacity to set and pursue ends. Succumbing to the use of phones dimishes capacity to pusue ends and serves developers of products.

Model does not change the bad

Phone addiction is the same as alcohol addiction or cigarette

Philosophy of technology use – personal philosophy covering which digital tools to allows, reason, and constraints

Do not think users MUST ensure time spent online restricted to small number of optimal activities

Settle for reasonable number

Don’t need users need to develop philosophy of technology use. Happy to settle for cultivation of disposition – may be realized via development of philosophy.

Digital minimalist – one whose interaction are intention and don’t conflict with ends. Involved serious reduction or elimination of interaction with phones.

It does not remove use of phones completely

Use of devices threatens rationality, intentional use, thoughtful, etc

May use tactics to bring people to use phone – streaks – reminders – notifications – stay in touch – etc

Kant says humanity has absolute, inherent value, and morally required to respect humanity in self. 1. Greater detail about ‘humanity’ and what is disctintive about value. 2. Explain grounds duties to self. 3. Explain distinction between perfect and imperfect duties. Not committed to every aspect of Kant. Adopting plausible premises to argue for existence of duties to self.

1. Dignity – objective, uncondition, and fungible. Objective – holds for all rational agents and contrasted with relative value – requires prescence of contingent inclination
2. Unconditional value – can’t be lost and holds in all context
3. Nonfungible – have kind of value not admitting of exchanges for things of equal value. Everything has price or dignity. What has a price can be replaced. Dignity can’t

Duty of respect – respect is only appropriate attitube to adopt with regard to dignity of rational nature. Act with humanity – dignity and respect. Have duties to self. Wrong to commit suicide to avoid life of pain vs pleasure. Suicide throws away rational nature and fails to express respect for life.

1. Perfect and imperfect duties – kant uses formula of humanity to derive more of duties. Makes distintion between perfect and imperfect duties.
   1. Perfect – strictly forbidden action. No action – promise breaking and suicide
   2. Imperfect – stem from obligation to respect humanity as end. Adopt ends of happiness of others, natural and moral perfection. Obligation to pursue ends but not how much or how.

P1. Humanity – has objective, unconditional, and nonfungible value – dignity

P2. Anything with dignity needs respect and not treated as a means

P3. If humanity is respected and not treated as mere means, we have imperfect duty to cultivate and protect rational agency

Conclusion – We have imperfect duty to cultivate and protect rational agency

P4. If we have imperfect duty to cultive and protect rational agency, we ought to adopt end of digital minimalism

C2. We Ought to adopt end of digital minimalism

We should look to familiar social activities to inform privacy policies

For online cases without precedent, look to ends, purposes, and values

Humans have

P1. Humanity has an objective value

Subjective value

P2. Anything with objective value (dignity) ought not be treated merely as a means.

P3. If humanity is to be respected as an end, then you have an imperfect duty to protect/ cultivate your rational agency.

C1. We have an imperfect duty to ourselves to cultivate/protect our rationality.

First order desire

Second order desire

Daily me can be about anything - everyone in charge of what we see and hear

Homophily – tendency to connect and bond with similar people

Face book used to connect

Don’t need a daily me, others do it by algorithms

They know emotions and may be able to mimic

Fb knows political convictions, inform others, etc

Machine learning used to do this

Hastags used to sort

Social precondition for well-functioning system of democratic deliberation?

How does social media and machine learning impact the ability to govern self

Architecture of serendipity – save of individual lives – counteracts homophily and promotes selfgovt and individual liberty

People’s growing power to filter what they see and provider’s growing power to filter for each, what hey know

1. People should be exposed to materials they would not choose. Unplanned
2. Don’t force people to see things wishing to avoid but it should not be foreseeable always.
3. 2. Most should have wide range of common experiences, differences in society will make it hard to address social problems and understand each other
4. Choose optiminism but everything should be either optimism or pessimism

Too simple if any way of communication is desirable if and because it allows people to see and hear what they want.

Need to know about things u don’t wanna know about

Violent extremism – reason to know about things

People worried about impact – want to agree w others but social norms makes it hard. Republicans and democrats for example

Falsehood beliefs

Multiple opinions gives freedom of choice

Everything has architecture and diversity

Facebook changing algorithms to learn u

1. Political bias allegations
2. Oligation to shareholders
3. Many just posted new articles of random stuff

Emotions contagious

Things u read impact how u think

Nguyen C. Thi – how twitter gamifies communication

Twitter is used to interact with others but is not neutral. It is biased

Algorithms used

Gamifies communication by offering immediate, vivid, and quantified evaluations easily. Offers points, scores communication

Game like features responsible for addiction

Short term and longterm

Gamification can be tech for increasing motivation but Nguyen is not optimistic

Gamification increases motivation by changing nature of activity

Start chasing likes and followers

Interested in popularity and status

Games and gamification – engagement, grinding, feedback and rewards, level up. Games quantize success. Leaderboards and ranking increase productivity but also may increase injury rates

Duolingo uses daily goals

Gamification 0 trade. Increase motivation by narrowing and simplifying target – changing the nature of activity

Goals may be subtle but gamification may destroy all

Designers don’t just create characters, stories, environments. They design a player world, scult temporary agency, structure, etc. Motivations set by goals

In life – values are subtle and hard to apply but ez in games

Twitter’s way – pleasure, fun, engagement

Invites uses to change goals of participation

Retweet, like, follow, etc

Emphasis on likes, aggregation method too

De-emphasize quality but highly rates number of users with positive reaction

Aggregates interests into 1 statistic, focusing on 1 main section of interests

Systematic pressure

Management makes it better but it can still replace values, not literally but functionally

Proper management is key

Change

.Motivation with intention matters

1.Game player – temporarily adopts scores as goal. Perfectly harmless

But there are games of deceit too

Realization is important, some people are sarcastic and some are serious, know ur situation

2.Interalizing scores. Goals influences by scoring mechanism. Value capture

1. Natural values rich, subtle, and hard to express

2. placed in social or institutional setting presenting simplified, quantified, version of values back

3. simplified version takes over motivation and deliberation

EX. Daily step count, GPA, likes, etc

Context stripping – aggregate info arithmetically

Pressure comes with overt gamification

Quantity can be averages which create leaderboards too

Issue w value capture – assumptions. Belief aims at the truth. Temptations may lead to abandonment

Value-independent user – avoided internalizing scores – avoid gamification

Value independent - Scores are merely a means.

Value captured – scores are the goal

Change goal 0 gamification problematically instrumentalize sgoals.

Interest is to achieve goal using a structured path or form of activity, in gamification, it becomes motivational

Achievement player – play for value of winning

Striving – temporary interest in winning

Justify goals by r\justifying or pointing to values of some things

Striving games – goal temporary but gamification may be long term and impacts real goals

Twitter and toxicity – lots of problems online. Echo chambers

Radical trust

Moral outrage crucial and helps register injustice and motivate

Problem is that it can threaten to corrupt

Invites to instrumentalize moral beliefs

Choice architecture – designing options or choices that influence how people choose

Architecture of control – Ability to not control everything in life. Polarized and reinforce feelings/beliefs

Architecture of serendipity – might have to do something you don’t want but will end up enjoying and finding value

Requirements for well functioning system of free expression

**People exposed to material they would not have chosen in advance**

There are common shared experiences that form a kind of social glue

Value – independent user – adopts gamified motives only instrumentally

Design for gamification – designed tech which offers points/scores/rewards

Gamification – when players interact w design and adopt motives

Value capture – placed in situations where values are greatly simplified/quantified. Those simplified versions become our real motives

Value captured user –

Epistemic bubbles - when not exposed to outside views ( knowledge )

echo chambers - when community actively resists outside info (cult)

Moral outrage porn (rage bait)

Meditations 1

Alone

Can’t prove all opinions false – impossible

Find reason for doubt or belief

Basic principles

Doubt is from senses

Reasoning is important

Dreams have visions like paintings

But reasoning shows things are real

Common sense

Even though some things create doubt, it doesn’t mean it doesn’t exist

Don’t need to understand something to believe in it

Efforts are important and habitual opinions as custom of long occupation and law of custom

Things are based on reasoning

Can not enjoy imaginary freedom

If u doubt everything, senses will contradict belief. What u see, hear, everything becomes untrustworthy

Memories may be fake then

Constant doubt of eeverything will mentally torture you

Nothing is certain then

What’s true? Is everything a lie? Yada yada

What creates thoughts???

What exists?

Conclude that I am, I exist

Previously thought he was a man but what is a man? Need context and thought and reasoning

Physicial body, mental conception, etc

A thing that thinks. A thing that doubts, understands, denies, affirms, wills, unwills, imagines, and senses

Sensory perception can’t be false

Even with senses, if they change, the object still does not change. Ex. Wax before and after a fire is still wax

Are something extended, flexible, changeable but you can’t ever imagine allthe ways it can be changes

The mind can only perceive so muc.

In conclusion, man is weak and able ot make mistakes in thoughts and what we see before us is not always the truth

“Common sense”

Opinions can be changed but people use senses to create initial opinions which get changed as you learn more.

Poele

Method of doubt

1. Doubt the senses
2. Dream skepticism
3. Evil deceiver

Skepticism – belief that knowledge is not possible

Math, people are good at because it is proven and constantly solved for, not guessed

Natural vs revealed theology

# OUTLINE

Moral Responsibility in the Case of A.I.

Question: Who should be held morally responsible for immoral decisions/actions made by an A.I.?

Thesis: Users of A.I. technology and those implementing immoral courses of action made by A.I. should be morally responsible.

Abstract: Artificial intelligence has grown into everyday life and is being accessed constantly; however, the decisions and suggestions presented come with moral responsibility and handling. Artificial intelligence uses algorithms in decision-making; the machine isn’t always accurate and has zero understanding of context. As a machine, A.I. is only able to do what the program interprets, moral decisions aren’t accountable in this as A.I. bases answers systematically. AI lacks consciousness, feelings, and any moral reasoning, using patterns in data and trained sets to generate answers. AI predicts using similar phrases and concepts, along with contextual input in its database. It then applies algorithms to return a result. The accuracy of AI’s answers is only an approximation and a guess based on previous data in a similar context. It is important to understand that AI makes mistakes, and human decisions compared to machine-driven ones are incomparable, therefore, the user is responsible for enacting and enforcing those decisions. AI’s inability to understand context and experiences makes it inapplicable in a moral context, making the user responsible for following AI decisions blindly.

Outline:

* Abstract:
* What is AI, and why it should not be responsible for any decisions, placing responsibility on users
* Introduction:
* Expand on how AI works
* Body:
* Expand on moral responsibilities
* Kant – Intention with AI is important, and intention with its answers is too
* Utilitarianism – If AI’s answers lead to pain, they should be avoided, and if it leads to pleasure, it’s good; responsibility is on the person listening to the advice or answers of AI, not AI itself
* User’s responsibility
* AI’s lack of understanding of moral responsibility
* Objection:
* Engineers and programmers should be responsible
* Businesses creating AI systems
* Policymakers
* No responsibility on the user because they had no control over the AI’s design
* Rebuttal:
* There is a responsibility on the user to listen to the AI; choices made by the AI come with warnings stating that it can be inaccurate, and goodwill should be implemented.
* There is nothing wrong with it until action is taken, meaning that even if immorally wrong actions were suggested, if physical action based on that suggestion isn’t taken, it aligns with good intention.
* End users choose how to implement AI suggestions. For example, a person using a car and it being faulty leaves the user responsible for going to fix it.
* AI’s unpredictability – Kant says moral duty is tied with rational will and intention, but AI has no control nor understanding, therefore no intention, and so the responsibility falls onto those using AI, not those creating it and the rules with it
* Utilitarianism – prevent harm. Developers and policymakers are responsible for this, but users also must use precaution and focus on the greater good when implementing it; they can’t use it blindly.
* Conclusion:
* Restate thesis: Moral responsibility is placed on the user of AI, not the implementors nor the AI software itself because it contains zero understanding and is programmed in a systematic way, which may contain errors and has no understanding of context. This means that the AI is unable to accurately depict because it only goes off the data it contains. The user is responsible for making decisions and following the AI’s instructions. Ex. Reading a book on poisoning someone or how to poison someone doesn’t mean that you should go poison someone; it is simply instructive and informative.

Sources:

Kant, Immanuel. “Fundamental Principles of the Metaphysics of Morals.*” Ethics Across the Professions: A Reader for Professional Ethics*, edited by Clancy Martin, Wayne Vaught, and Robert Solomon, Oxford University Press, 2017, 76-84.

Mill, John Stuart. “Utilitarianism.” *The Floating Press,* 2009, 1-23.

Müller, Vincent C., "Ethics of Artificial Intelligence and Robotics", *The Stanford Encyclopedia of Philosophy*(Fall 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2023/entries/ethics-ai/>.

Rudy-Hiller, Fernando, "The Epistemic Condition for Moral Responsibility", *The Stanford Encyclopedia of Philosophy*(Winter 2022 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/win2022/entries/moral-responsibility-epistemic/>.

[The Epistemic Condition for Moral Responsibility (Stanford Encyclopedia of Philosophy)](https://plato.stanford.edu/entries/moral-responsibility-epistemic/)

[Ethics of Artificial Intelligence and Robotics (Stanford Encyclopedia of Philosophy)](https://plato.stanford.edu/entries/ethics-ai/#MachEthi)

Archimedean point- standing point

Cogito ergo sum – cogito

Je pense done, je suis

Clear and distinct perception/idea – reason why you know you exist

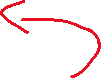
Essential property – remove property and it destroys it. Your mind is an example, your soul



Accidental property – properties that don’t ruin it. Your body is an example

Foundationalism – all knowledge is either:

**Axiomatic/basic**



**Based on derived or derived from basic knowledge**

**Mind body dualism –**

1. **There are 2 fundamentally different substances : mind and body**
2. **You are a combo of mind and body**

Primary properties – properties not dependent on perceiver

Ex. Size, shape, motion in units, geometrically

Secondary properties – any property depending on the perceiver

Ex. Pain, tickle, it impacts people differently based on how their mind is trained. Taste is also an example. Color is also an example according to Des Carte. Sound, smell. Subjective. Size, motion, and shape can also be secondary depending on how you word it. “Big” is subjective. 5cm is objective

But then everything is a secondary property

For scientific findings, all 5 senses are subjective, therefore, they are secondary properties

Secondary properties are changeable and destroyable

Rationalism

Minds, brains, and programs

Strong AI vs weak AI or cautious AI.

Weak – principal of value of computer in study of minds = very powerful tool.

Strong ai = not only a tool but also a mind that can understand and have cognitive state - not mere tools

Strong AI used in argument

Aim of program – simulate human ability to understand stories

People have preconceptions

Strong AI claim

1. Machine can be said to understand story and provide answers
2. Machine and program explains human ability to understand story and answer
3. Unsupported by Schank’s work

Gedankenexperiment – thought experiment

3 batches

Script

Story

Questions

Answers to questions

Just because you can answer the questions properly doesn’t mean you understand what you’re answering

It isn’t understanding, it is symbol manipulation

Formal principles, computers can’t understand it but humans will follow it without understanding

Many types of “understanding”

Tools are extensions of purposes

Computer understands nothings

2 systems for the man – one understands English, the other understands Chinese but neither have anything to do w the other

Systems have inputs and outputs and noncognitive subsystems turn to be cognitive

Stomach does info processing, similar to computers, but it does not understand

Strong AI does not have a chance of theory of mind to ponder implications

Beliefs are nonexistent

Hunk of metal has beliefs is atrocious idea

Problem of other minds

Strong ai – genuine intelligence/understanding

Weak ai – merely imitating of mimicking human intelligence

A.G.I - Artificial general intelligence

**The singularity**

Singularity – intelligence explosion- greater levels of intelligence as machines creates more intelligent machine

Key idea is that machine is more intelligent will be better than humans at designing machines so it will make a more intelligence machine. It can make one smarter than itself

What to expect then?

Speed explosion – speed doubles every 2 years but then with better AI, faster processing = faster designers = ever-faster design cycle = limit point

Intelligence explosion and speed explosion logically independent. In principle, either one is possible without the ideas but they work together well.

Laws of physics impose limitation

Speed and intelligence are pushable beyond human capacity in short time

Might not be singularity but similar

Practically – if singularity exists, it will have extreme benefits and alsodnagers

Philosophically – interesting and hard to think about consequences of values and morality, consciousness, and identity

They intersect

Can human identity survive enhancement of cognitive system

1. Argument for singularity
2. There will be AI+ (AI greater than human level)
3. If AI+, there will be AI++ (Superintelligence)
4. There will be AI++

Assumes intelligence and speed are independent

1. Will be AI (before long, absent defeaters)
2. If AI, then AI+ (soon after, absent defeaters)
3. If AI+, then AI++ (Soon after, absent defeaters)
4. Will be AI++ (Before too long, absent defeaters

Technology advances causes this

Soon = within decades

Defeaters = disasters, disinclination, active prevention

Premise 1

1. Human brain = machine
2. Capacity to emulate machine
3. If emulated, there is AI
4. Absent defeaters, there will be AI

Evolutionary argument –

1. Evolution produced human-lvl intelligence
2. If 1. Then we can produce AI
3. Therefore absent defeaters, there will be AI

2 suggestions – brain emulation (simulating brain neuron by neuron) and Artificial evolution (evolving population of AI thru variation and selection)

Direct programming, machine learning, etc

1. If AI, AI produced by extendible method
2. If produced by extendible method, soon have capacity to extend method
3. Extending method produces AI will yield AI+
4. Absent defeaters, if AI, there will be AI+

Use strong induction to prove premise 3

Proportionality thesis – increases in intelligence leads to proportionate increases

Singularity – an intelligence explosion

AI+ - AI more intelligent than human

1. If AI +
2. If AI + then soon AI++
3. There will be AI++

Wake up abducted

Wake up in alien space craft

Green walls all around, alien pod like room

9 other msoe students

Alien representative ocmes in and has purple hat, beanie like

Starts talking saying they going thru galaxies n finding good planets, helping and then blowing up bad planets

Gives night to think and then test tmr, if pass = keep planet. Fail = blow up

Headwear loving species

Shows empathy

Test to headwear, pick 2 hats, purple and green, placed on ur head and say which is on ur head

Tallest to shortest facing forward, 1 by 1 asking

Ideally find hats

1/10 acceptable, 2/10 bad, explosion

Line up

Don’t know how many hats per color. Cant remove hat. Cant see being placed, same feelings

Tells u if correct after answer

Materialism

External constraints

Internal constraints

Potential criteria for “moral” AI

1. Transparent to inspection
2. Predictable to those who govern
3. Robust against manipulation
4. Attached to someone/thing that is morally responsible

AI Is bullshit

Description of new tech tells us what tech for and expected to do. False statements by large lang models = hallucinations

Explains how chatgpt and LLM operate

View during factual error – lying or hallucinang, deliberate blame

Neither are accurate

2 sorts of bullshit - hard and soft

Hard – requires active attempt to deceive reader or listerner

Soft – only requires lack of truth for concern

Outputs of LLM areoft bullshit at minimum

Speech or text produced without concern for truth is produced without intent to mislead

May produce hard bullshit if viewin as having intentions

Real ai chatgpt – closer to human like mind in a silicon brain

Goals of system a drastically different for LLM and humans – goals and behaviors

Basic physical desires, food, sustenance, social goals, relationships, create physical objects

LLM – simply aim to replicate human speech or writing – produce human like text by estimating likelihood

Done by making statistical models on input text and data

Models associate word a vector loating it in high-dim abstract pace, near other words that occur in similar contexts and far those which don’t

Looks at prev strings of words and constructs diff vector, location word’s surrounding, context, to similar words

Model takes 2 vectors and makes set of likelihood for next word.selects and places likely ones, not always the most likely

Repeats till recognizable

AI is not reliable and makes errors. Does not check self and cfidently produces false claims

Accuracy important in some cases

Models are trained then words in database affect probability that bot adds one or another word to line of text

LLM can be connetied to database by consultin – can still have errors cuz it can misinterpret or ask qrong uestion to machine connected to it

Problem – LLM arnt designed to represent world.Instead designed to convey convincing line of text

Not meant to handle arithmetic temporal reasoning

Only manipulating form of langage

Not designed to transmit info

LIES, HALLUCINATIONS, AND BULLSHIT

Lie – make believed-false statement to another person w intent that other believes it to e true

Lies frowned upon

Acts f misleading that r not lying

* Spread untrue gossip mistakenly
* Bullshit

Lie, bullshit both nounand verb. Should not be accompanied by explicit intention

No conviction about uth , pays no attention

Lack of cncern

‘does necessarily attempt to deceive abut in enterprice. Only indispensabley distinctive characterstic is way to misrepresent’ – resitrcts bullshit to utterances but some features lack

Distinction

No intenton to deceive count as bs?

Favor of expanding definition – indifference totruth is DANGEROUS

Dangers come no matter whether there is intention to deceive or not

Advertising industry – lete w instances of BS unmitigates they serve among most indisputable and classic paradigms of conce[t

Suggests Advertisers want to mislead, theyre expected to say misleading things but that may not be their intention

Bullshit – general – utterance producd where speaker has indifference to truth of utterance

Hard bullshit – produced w intention ot mislead audience about agenda

Soft bullshit – produced with intention to mislead hearer regarding agender

BS – by contrast captures negative requirements. Indifference to truth

ChatGPT minimum soft bullshiter bc not an agent so cant hold attribute to ruth and to deceive others. Indiffernce is also dangerous

Chatgpt doesn’t intent on conveying truth – soft bs

Arguments saying ChatGPT is hard bs

1. Not agently or relatedly
2. Donot and cannot intend

Still has function and intentions. Does it have goals or aims? Beliefs?

Agent- something with intention and believes. Can it have goals?

Bullshitter – person using it bc

1. Don’t care abt truth
2. Want reader to believe output

More than a mere tool though

Chatgpt is not bullshit but its bullshit is bullshit

What is Chatgpt trying to do? Does it care

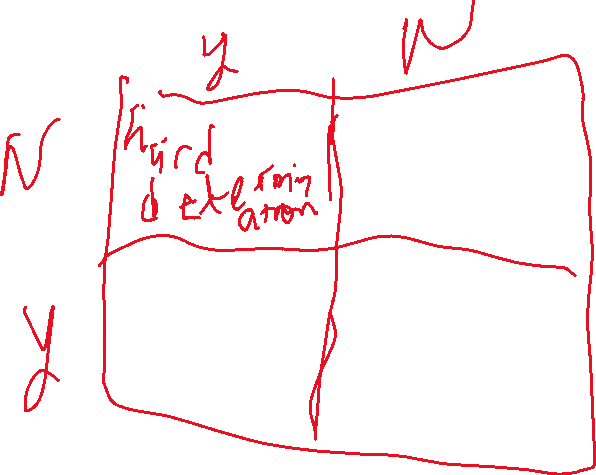
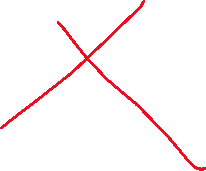
Primary function – imitate humans

Moral responsibility No and determination yes – Hard determination

Moral responsibility yes and determination yes – soft determinaton – compatibilism

Moral responsibility no and determination no – X

Moral responsibility Yes and determination no – libertarism



Dd

Potential criteria for moral status

1. Sentience – capacity for qualia
2. Sapience – capacities associats w higher intelligence (self awareness, reasoning)

Regarding idea of intelligce, would it make sense to argue that ai doesn’t have intelligence cuz while humans r guessing too, they do it solely on their memory and what they learned, but chatgpt has a countlsss spot of references that it can access

Principle of substrate non-discriination. If 2 ty and conscious experience, but differ I substrate of implementation, then they have same moral status.

Principle of ontogency nondiscriminaton – same as principles of substrate but differ only in how they came into existence, then they havsame moral status

Potential ‘exotic properties’

1. Subjective rate of time
2. Sapience w/o sentience
3. Instant acquisition
4. Reproductive rights

Lying – deliberately saying a falsehood with intention of making someone believe it

Hallucination – blamelessly uttering things on basis of misleading

Hard bs – active attempt to deceive to nature of things

Soft bs- merely lack of concern for truth

London – AI and black box medical decisions – accuracy vs explainability

Powerful machine learning techniques purchase diagnostic at expense of ability to access knowledge. Deep learning systems cant explain decision making

Trust connected to ability to explain expert recommendation

Expect experts to be knowledgeable

Machine learning systems should be explainable or interpretable is unfounded and potentially harmful

Trust in expert grounded in ability to produce results and justify. More than just accuracy, requires justification

2 kinds of knowledge needed

1. Domain model – causal relationship in domain captured and domain principles

laying ‘how to’ knowledge

Reasonable bc they have intellect requirement

Techne – productive science concerned w bringing into existence effects

Aristotle says that people w lots of experience lacking theory achieve better results than those with only theory

Science of medicine combined knowledge of subject with knowledge of universe

Justification is important

Explanation fosters social trust and also allows understanding for why something was done and holds experts accountable for avoidable error

Prevents experts and stake holders from abusing authority and giving arbitrary information

Deep learning systems trainable

Can be black boxes. Understanding system uses test to predict doesn’t increase ability to intervene

Don’t need to understand mechanism to agree, only need trust

Not all branches of decision making are understood as well as others

May not be able to explain WHY something works, cant have knowledge in everything

Overreliance on theoretical explanations lead to practices that harm patients

Many factors in medicine, not just major factors

Age, ethnicity, gender, etc

Accuracy is important

Ability to intervene by exploiting causal relationships comes from experience and isn’t related to understanding WHY things work

Conclusion: While AI is accurate and improves outcomes, inability to explain how decisions r made can ruin trust and raises concerns for errors

Must have balance with accuracy and ability to explain

What is ChatGPT trying to do

What is ChatGPT’s function

Function – imitate human speech

# Final Essay

Moral Responsibility with AI

Artificial intelligence tools like ChatGPT and GitHub Copilot use algorithms to generate output from user input. Output is produced by recognizing and replicating patterns derived from human data, trained sets, and exemplars. While AI appears intelligent and capable of producing coherent output, it doesn’t have true intelligence. This is because AI lacks qualities of intelligence, specifically of moral agency: rationality, consciousness, and intention. I won’t go deep into what intelligence is or isn’t, but for this, I will refer to intelligence as the capability of understanding and reasoning. Arguably, AI is intelligent in the sense that it can solve problems and logically give input. However, by the definition of intelligence, it fails as AI can’t make informed decisions; it follows programmed logic and has no inner awareness behind its output. AI may simulate intelligence, but data replication without understanding does not constitute intelligence. For instance, if all students receive full marks on their exam because they had previous exam questions and answers as reference notes whilst taking the exam, then the exam hasn’t properly weighed their understanding. Similarly, if AI has an endless storage of data to reference and is unable to understand the material, it also lacks intelligence. Mankind can learn to become intelligent by studying material and rationalizing. This is by using assistive material to understand, allowing people to gain intelligence and grow. Upon studying, people then test their knowledge by checking their understanding without reference materials. However, with AI, it constantly relies on stored data and material. Using Auerbach’s notes, AI cannot rationalize as it can’t learn and understand; without the ability to understand, it prevents it from achieving growth (Auerbach). Without reference to data, AI is incapable of properly generating output and conclusions. Even with increased memory storage for AI, it’s unable to retain intelligence as it fails the ability to understand the information memorized. Similarly, if students memorize problems and copy the layout, they haven’t understood how to solve them. Once something out of line is encountered, they won’t be able to solve it. In the same sense, AI can’t understand input as it only uses probabilities and algorithms to check for the next likely output. This shows that AI hasn’t actually understood the context of the input but instead uses clues and probabilities to give an estimate for a response. As a result, AI lacks true intelligence, which is the foundation of moral responsibility. Without intelligence, AI can’t be held to moral standards. This can be further explained through Kant, as intelligence plays a key role in the assessment of moral responsibility and value. In further sections, I will demonstrate that AI is not morally responsible using Kantian principles, specifically the humanity principle, rationalization, and the role of duty and intention.

Building on the concept of intelligence, moral agency is defined by possessing rights and responsibilities. These attributes are associated with moral agents. The idea of moral agents having rights and moral values stems from intelligence. With responsibilities, the ability to process information is important; however, understanding how to act and reason with moral intention is greater. Moral agents aren’t only capable of making choices but also responsible for the choices because they contain rational will. Rational will ties in with moral agency as reasoning and freedom of choice with decisions are crucial for agents. The ability to create and be responsible for decisions requires the ability to recognize morals and act alongside them. “Moral reasoning is needed in robots that have the capability for lethal action” (Bringsjord). To be responsible for actions, essential attributes include reasoning and understanding. While AI can replicate intelligence, it lacks true intelligence, therefore lacking self-awareness, reasoning, and the ability to make moral decisions (Auerbach). AI is unable to reason without understanding context. Without the ability to understand, AI is also unable to understand and obtain qualities of self-awareness. These include consciousness, identity, internal thoughts and feelings, and understanding. AI may state that it understands, however, it is surface-level pattern matching, not understanding. The system repeats the phrase, fitting the context without understanding the scenario. As such, AI is unable to recognize and act with moral decisions. Moreover, while AI can simulate moral behavior, it can’t comprehend moral duties nor recognize others to give respect (Auerbach). Granting AI moral responsibility would mean removing the intelligence component. AI is also programmed to follow its logic and therefore doesn’t have free will. Without free will to make its own decisions and logic, AI is unable to create intention or act out of duty, both conditions which are essential to being treated as an end. Since AI lacks both, it holds no moral value, “I omit here all actions inconsistent with duty. I also set aside actions with no inclination” (Kant 77). Without the ability to act out of duty or have intentions, AI can’t be considered a moral agent and then can’t be held to the standards of moral agents. A maxim can’t exist as a universal law of nature where AI holds moral responsibility, being derived from will and intention solely. Additionally, if AI bears moral responsibility, manipulation of morals would disrupt society; therefore, it can’t be rational and can’t be held as a universal law.

The humanity principle states that people should be treated as ends and not as a means. Treating AI and technology as ends means respecting their rationality and intending for it to reach happiness (O’Neill 162). AI, which contains neither attribute, is unaffected by how it is treated. Whether respected or not, the machine will grant a response, unlike humans. AI cannot also input its stance on a topic, returning to the idea of intelligence. Since AI can’t make its own decisions, it also can’t conjure feelings and biases, which humans possess. The tool isn’t programmed with feelings and biases, only sets of data, making AI objective in its stance. Without the ability to make decisions without data sets, the machine is only a tool simulating a human-like answer. The tool doesn’t have its own mind or belief system, since both require intelligence to create, therefore negating the idea of intention. Treating AI as a means also means granting it moral rights. Along with this, it also implies that AI seeks happiness and goals that are imperceptible. This is not the case with AI, though. It contradicts itself, as AI’s only job is to produce a response back through algorithms; happiness and plans are only applicable to humans, not tools. Additionally, deceiving is a way of using others as a mere means (O’Neill 162). If AI could generate immoral decisions and were treated as an end and not a mere means, then it would go against the principle by granting permissible impressions of decisions that are not morally accepted. Therefore, AI would deceive users of the tool, treating others as a means and not as an end, which contradicts the principle, as this treatment is mutual and should result in mutual respect. As a result, AI can’t be an end and, therefore, cannot hold responsibility over moral decisions.

AI operates solely on code and algorithms, not moral duty; it was programmed to do it and has no choice but to follow the program. Through Kant’s description of duty, intention is key. AI’s decisions aren’t produced because it was the right thing to do, intention-based, but because it was programmed to. Therefore, all decisions and outputs given are irrelevant and hold 0 value in terms of morals (Kant 77). On another note, morality is tied to maxims and duty; placing morals on AI would create ethical problems by undermining responsibility. Since AI lacks intelligence, it also lacks the characteristic principles of acting and ethics for mankind. AI is only a tool utilized in modern society, an object that cannot have any intention or thoughts, only algorithms (Müller). AI’s importance as a tool can’t be overridden by humanizing characteristics and applying moral duty. If duty is placed on AI and taken off humans, this would create an imbalance between the values of mankind and tools. To prevent this, users of AI should be held responsible for their usage of AI as moral patients. These are ethical agents who have duties and are subject to the responsibilities concerning immoral decisions made by AI. Through Kantian ethics, people will have duties to uphold when facing immoral concerns.

Some argue that as AI continues advancing, it will eventually develop intention and awareness, granting it intelligence. This would then make AI capable of qualifying for Kant’s requirements to have moral authority. However, this still stems from the idea that AI’s moral reasoning and consciousness would stem from programmed goals and algorithms, rather than true understanding and reasoning. “Machines gain knowledge by humans manually encoding and inserting knowledge” (Bringsjord). The concept of intention is rooted in the ability to choose based on morals, not programming. No matter how advanced AI gets, it will always rely on its program, and its output will always remain systematically selected. In addition, the ability to reason and understand is unlikely for AI as it is structured to use probability to analyze and produce output. It can’t contextualize the input nor retain memory of it. If asked to reanalyze data from months prior, or even days, if a new chat has been opened, the chatbot is unable to recognize it. Until AI can contextualize and understand, it won’t be able to reason, making it constantly fail the requirements for intelligence. AI is also unable to act freely with its output. For example, having conversations with the bot that indicate unusual or random output is unlikely; however, for humans, with free will, unanticipated words are probable. In conclusion, until AI can act freely with consciousness, reason, and meet the requirements for intelligence, it can’t hold moral responsibility.

Through Kant’s principles and teachings of moral agency, the principle of humanity, and rationality, AI should not be morally responsible for making immoral decisions. Kant teaches that intention is a very important thing, actions do not matter without intention, and using this, AI is not responsible. This is because no matter what response is given by AI, there are no intentions, reasons, or consciousness behind it; it contains no feelings or understanding and has no true ability to analyze input. Therefore, Kant’s rules regarding will and intention do not apply to AI. Taking, for instance, that it was applicable, Kant’s maxims would contradict one another. This is because AI cannot be treated as an end when it is a tool, and granting a tool the same respect and rights as a human would deprive people of value. Moral duty cannot be applied to AI either because moral duty requires characteristics that AI lacks. Therefore, AI cannot be a moral agent either, as they are meant to have rights and respect. In conclusion, humans utilizing AI are responsible for making decisions, not the tool itself.

References

Auerbach, David. "Artificial Intelligence Lacks Genuine Intelligence: Why It’s Not as Smart as We Think." *AI for Social Good*, 25 Mar. 2021, URL =<<https://aiforsocialgood.ca/blog/artificial-intelligence-lacks-genuine-intelligence-why-its-not-as-smart-as-we-think>\>.

Bringsjord, Selmer and Naveen Sundar Govindarajulu, "Artificial Intelligence", *The Stanford Encyclopedia of Philosophy*(Fall 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2024/entries/artificial-intelligence/>.

Kant, Immanuel. “Fundamental Principles of the Metaphysics of Morals.*” Ethics Across the Professions: A Reader for Professional Ethics*, edited by Clancy Martin, Wayne Vaught, and Robert Solomon, Oxford University Press, 2017, 76-84.

Müller, Vincent C., "Ethics of Artificial Intelligence and Robotics", *The Stanford Encyclopedia of Philosophy*(Fall 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2023/entries/ethics-ai/>.

O’Neill, Onora. *A Simplified Account of Kant’s Ethics,* 2010, 161-162.

‌

Human oversight

Better results

EXAM Review

First order desire – A desire about something – desire to own car – beliefs but not how to obtain

Second order desire – Desires about desires. Desire to desire – beliefs and desires about how – desiring to desire healthy food

AI +/ AI ++ - AI more intelligent than humans and AI more intelligent than AI

Echochambers – moral outrage porn, when people are activitely resisting information. Cult like idea, only amplify beliefs that reinforce beliefs and shut down other beliefs

Choice architecture – The way choices are presented to influence decisions.

Value capture - Situations where values are greatly enhanced, usually through gamification such as likes and comments to indicate popularity and value.

Hard bullshit – Intentional deceivement through speech or text

Soft bullshit – Speech or text produced without concern for truth, produced without intention

Architecture of control – Ability to not control everything in life. Polarized and reinforce feelings/beliefs

Architecture of serendipity – might have to do something you don’t want but will end up enjoying and finding value

Gamification – Increasing motivation and usage through changing of nature. Interacting with designs and changing their motivation behind usage. For example, some may use Twitter for information, but through its gamification using likes and posts, they begin getting obsessed with the idea of getting the most number of likes.

Essential properties – Properties that, when removed, will destroy the object. For example, an object of 4 sides is a square, however, with more or less than 4 sides, it is no longer a square. Ex. Soul

Accidental properties – Properties that do not change the actual object but describe it. Ex. Body. If you destroy your body, you are still you.

Materialism – The idea that your mind is connected to your body or a part of you physically. Consciousness is the result of matter and material interactions.

Singularity – The idea that technology will grow to a point irreversible and will lead to AI+ and AI++

Strong AI – AI able to understand, learn, and solve problems, they have intention

Weak AI – Narrow AI, designed and trains for a limited number of tasks, does not have understanding or consciousness

Determinism – idea that everything is inevitable. Everything is a chain result

Natural theology – Reason and evidence to argue for existence of a creator

Revealed theology – Acceptance of truth through revelation or faith, belief only

Epistemic bubble – Not being exposed to outside views, a sort of bubble where you are only exposed to ideas that support your belief, preventing you from deterring

AGI – AI with cognitive abilities, can learn, has intelligence, AI that can do anything a human can do intellectually

Lying – Deliberately saying falsehood with intention of having someone believe it

Hallucinations – Uttering things blamelessly with the intention to misinform/mislead

Sapience – wisdom and the ability to make decisions on knowledge, experience, and understanding

Primary properties – Properties not dependent on perceiver. Example, size

Secondary properties – Properties dependent on perceiver. Example, color

Foundationalism – All knowledge is based on basic knowledge or derived off basic knowledge

According to Searle, why can’t AI have genuine understanding?

– Only processes symbols, does not understand, similar to the Chinese Room

According to Sunstein, what are the criteria for a well-functioning society of free expression?

- Expose to views that people don’t want to listen to, prevents echo chambers, shared experiences and common knowledge should overlap, and people should have reasons

How might ChatGPT be a hard bullshitter? What is AI trying to do?

* AI is trying to impersonate human behavior through symbols. It mimics intelligences and has no concern for truth.

How do Kant’s imperfect duties lead to digital minimalism?

– We have duty to protect our rationalism and digital use is addicting and known to cause violence, therefore, as users, we should limit to prevent self harm.

What is Descartes’ Archimedean point? Why?? Think method of doubt

* “I think, therefore I am” Descarte tried to doubt everything, his thoughts, existence, everything, however, this made him realize that nothing was certain. He was skeptic, thinking everything might deceive us and could be illusions but then realized that the fact that we can think makes it so that he exists.

How does Twitter lead to value-captured users?

* Twitter uses likes and number of followers to indicate popularity and create value captured users. Twitter also uses personalized advertising to keep people engage

What are some criteria for “moral AI” (Bostromax Yudkowski)

* Transparent to inspection
* Predictable to those who govern
* Robust against manipulation
* Attached to someone/thing that is morally responsible

What are some potential internal/external constraints for AI? Why might we worry about them??

* Bias, consciousness, reliability, legality, ethical expectations, etc. These constraints make AI unpredictable and hard to control, hard to understand who is liable, and how reliable it is.

Why does Rene believe you can’t just be your body? (mind/body dualism)

* If u destroy ur body, you are still you, but if you destroy ur mind, you destroy yourself

Why might you think the singularity is coming?

* AI is getting better and better, and with AI smarter, AI itself can be used to create better AI which is a chain result, leading to a better AI, and AI+ and AI++

What are some ethical concerns for AI medical decisions? Does London find them convincing?

* Medical industry is based on trust, with AI taking over decisions, trust between patients and doctors is broken since AI can’t explain its diagnosis. London says we should implement AI as tools but should not replace doctors.

Why is bullshit a better term than hallucination for AI?

* BS is not about lying but more on concern for truth, which AI regards. AI does not have intention to mislead, but rather gives information confidently without care for accuracy

What is Descartes’ wax example? What does it show?

* While wax burns, its shape changes, however, even when the wax is completely melted, it is still wax, its identity has not changed, the object is still wax. It shows that destroying the physical component does not remove from what something is, wax is wax, brand new or destroyed.

What are some of the downsides of an ‘architecture of conrol’?

* Loss of privacy, blurs the line with transparency, normalizes surveillance, limits freedom

What role does Descartes’ evil deceiver play?

* The evil deceiver has the ability to manipulate, making Descarte conclude he believes falsehood. It then casts doubt on everything Descarte believes and then leads Descarte to look for a belief that can not be false -> I think, therefore I am

What does echo chambers/ moral outrage porn have to do with gamification?

* Gamification will use echo chambers to feed content that remains with beliefs you align with, creating an echo chamber, in order to increase your likelihood of interacting

What problems might AI generate for personal identity?

* If an AI is created as a doppelganger, will it have my identity? Will AI need morality? What draws the line between AI and humans regarding identity

Why might ChatGPT only be a soft bullshitter?

* AI does not produce BS with the intention to deceieve, it just doesn’t care. No intention

Why might you think we don’t have free will? Does this matter for moral responsibility? (Hard/soft determinism)

* Everything is causes by a series of events, if something happens, it is inevitable, therefore, we are never truly responsible for our actions since everything will happen no matter what
* Soft determination, we act in accordance to desires, everything happens for a reason, but we act because desires and values played a role, not belluccause it was the only choice or thing to happen