Deontology/Kantian ethics

Giovanni Sartor

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423

Deontology

- Consequentialists hold that choices—acts and/or intentions—are to be morally assessed solely by the states of affairs they bring about.
  - E.g. my act of lying is good of bad depending on the effects it brings in the world
- Deontologist hold that certain actions are good or bad regardless of their consequences
  - Lying is always bad, regardless of its effect.
- The right has priority over the good: what makes a choice right is its conformity with a moral norm which order or permits it, rather than its good of bad effect.
    - E.g. we should not kill anybody, even in those cases in which killing somebody would provide more utility. Is this always the case
      - Consider the case of the British soldier who apparently met Hitler in the trenches of 1st world war
      - What would a rule utilitarian say in such a case?
- The 10 commandments?

Some ideas for being impartial

Ethics and impartiality

- Is ethics linked to ideas of fairness or impartiality?
- Is it unethical to have a preference for oneself (or one's friends)?

What about the golden rule

- Treat others as you would like others to treat you
- Do *not* treat others in ways that you would *not* like to be treated
- What you wish upon others, you wish upon yourself

Is the golden rule useful

- Always? Can you find counterexamples?
- Would you want an AI system that applies it (with regard to its owner)?

Immanuel Kant

- One of the greatest philosophers of all times
- Lived in Prussia (1724-1804)

- Addressed
  - The theory of knowledge: Critique of pure reason
  - The theory of morality: Critique of practical reasons
  - The theory of aesthetics (art): Critique of judgment
  - Law, logic, astronomy, etc.

Kant's ethic and the principle of universalizability

- "Act only according to that maxim by which you can at the same time will that it should become a universal law" (1785).

- What is a maxim: a subjective principle of action, it connects an action to the reasons for the action (an intention to perform an action for a certain reason)
  - I shall donate to charities to reduce hunger
  - I shall deceive my contractual partner, to increase my gains
  - I shall cheat on taxes, to keep my money
  - I shall tell the truth, to provide trust

- Are they universalizable? Would I want them to become universal laws, that are applied by everybody?

An universalisation test

- Shafer Landau. The test of universalizability:
  - Formulate your maxim clearly state what you intend to do, and why you intend to do it.
  - Imagine a world in which everyone supports and acts on your maxim.
  - Then ask: Can the goal of my action be achieved in such a world?

- The process ensure some kind of fairness

Apply this principle to

- Cheating in an exam, in order go get a good mark
- Giving money to a charity to relieve

- Would we want a robot following this maxim?

Immanuel Kant vs Benjamin Constant

- Should one must (if asked) tell a known murderer the location of his prey.
  - It is ok to refuse to answer?
  - It is ok to tell a lie (e.g., if threatened by the murderer)?

- Is the maxim of telling lies universalizable?
- Is it defeasible?
- Its it Ok to have a robot that tells lies:
  - What about Asimov Liar
  - What about HAL in

Hypothetical imperatives

- Hypothetical imperative: they require us to do what fits our goals
  - I would like to have more money
  - If cheat on taxes I will have more money
  - I shall cheat on taxes to have more money

  - I would like to get a good mark
  - If I study I will get a good mark
  - I shall study

  - Is this OK?
  - The imperative is dependent on what I want (getting good marks, having more money)
    - I shall cheat on taxes, to having more money!

The categorical imperative

- A moral imperative that applies to all rational beings, irrespective of their personal wants and desires,

- “Act only on that maxim through which you can at the same time will that it should become a universal law”

  - - make false premises when it suits you to do so?
  - - refuse help to do those who are in need when it suits you to do so?

The good will

- The morality of an action only depends only to the extent that this action is motivate by our good will, i.e., by the necessity to comply with the categorical imperative
  - E.g., if I do well my job only in order to get a promotion, or be better paid I am not acting morally
  - I am acting morally if I do well my job because I think that this is my categorical duty, since I believe that everybody should act upon the maxim that they ought to do well their job to ensure societal progress
- The good will is the only thing that is good in itself
  - Do you agree?

Another version of the categorical imperative: the principle of humanity

- So act that you treat humanity in your own person and in the person of everyone else always at the same time as an end and never merely as means
  - How is it linked to universalizability: As you consider your self as an end, you should consider the others in the same way (universalizability)?

- What does it mean treating somebody as an end (not as a mere means)
  - It cannot mean that we never use people for our purposes (e.g., when we ask for favours or pay for jobs)
  - It must mean that we should never treat people ONLY as means, without considering their values and purposes

When does AI treat people only as means

- Autonomous weapons?
- Deceiving advertisements?
- Discriminatory appointments?

- When does AI fail to recognise humans as valuable entities, that should achieve their aims according to their choices?

- Can we treat AI systems only as means?

Dignity

- For Kant rational beings, capable of morality (humans) have a special status “an intrinsic worth, i.e., **dignity**,” which makes them valuable' “above all price
  - Because of dignity they deserve respect
  - They cannot be treated as mere ends

- What does it mean that AI systems should respect human dignity, respect humans

The foundations of dignity

- Why do humans deserve dignity. Because they have
  - Reason: they act on reasons and are aware of this
  - Autonomy: the can choose what to do, and in particular to follow the categorical imperative rather than their subjective preference

- The kingdom of ends
  - In the kingdom of ends everything has either a price or a dignity. Whatever has a price can be replaced by something else as its equivalent; on the other hand, whatever is above all price, and therefore admits of no equivalent, has a dignity
- What if AI system also had reason and autonomy
- Would they become citizens of the kingdom of ends

Morality as an aspect of rationality

- For Kant if we follow rationality, we have to be moral.
  - Can there be a rational criminal?
  - It is rational to pursue my wellbeing at the expense of others?
  - Is it rational for a company to develop a system that is profitable, but that will cause more harm than good (e.g.,

Rationality and consistency

- 1. If you are rational, then you are consistent.
- 2. If you are consistent, then you obey the principle of universalizability.
- 3. If you obey the principle of universalizability, then you act morally.
- 4. Therefore, if you are rational, then you act morally.
- 5. Therefore, if you act immorally, then you are irrational.

What kind of consistency is this?

- If I deserve something no less than others, and I want it for me, I should recognise it also to others!
- Is this consistent with rationality? Is it required by it? Can I be rational, and pursue my goal to the detriment of other

Issues

- Does the principle of universalizability always provide acceptable outcomes

- Is it sufficient that the maxim of my action is such that I would like it to be universalised for this maxim to be good?

- Can you think of some examples when this is not the case?
  - Lying ? Robbing? Celibacy? Genocide?

Alan Gewirth (1912-2004): principle of generic consistency

1. I do (or intend to do) X voluntarily for a purpose E that I have chosen.
2. E is good
3. There are generic needs of agency.
4. My having the generic needs is good *for* my achieving E *whatever E might be* ≡ My having the generic needs is categorically instrumentally good for me.[13]
5. I categorically instrumentally ought to pursue my having the generic needs.
6. Other agents categorically ought not to interfere with my having the generic needs *against my will*, and ought to aid me to secure the generic needs when I cannot do so by my own unaided efforts *if I so wish,*
7. I am an agent → I have the generic rights.
8. All agents have the generic rights.

Other attempts exist to develop a Kantian ethics.

Approaches to universalisability

- Richard Hare (1919-2002)
  - Moral judgment are universalizable: the judgment that an action is morally right/wrong commits me to accept that all relevantly similar action are wrong
  - Moral judgments are universalizable in the sense that they take into account the satisfaction of everybody's preferences (back to utilitarianism)

Christine Korsgaard (1952)

- My humanity (capacity to reflectively act from reasons) it to me a a source of value, and
- I must regard the humanity of others in the same way.

Do we want Kantian robots

- Yes
  - They will be consistent
  - They will be impartial
- No
  - They may act on bad maxims
  - Their maxims may be too rigid

David Ross (1877 1971): prima facie duties

- Fidelity. We should strive to keep promises and be honest and truthful.
- Reparation. We should make amends when we have wronged someone else.
- Gratitude. We should be grateful to others when they perform actions that benefit us and we should try to return the favour.
- Non-injury (or non-maleficence). We should refrain from harming others either physically or psychologically.
- Beneficence. We should be kind to others and to try to improve their health, wisdom, security, happiness, and well-being.
- Self-improvement. We should strive to improve our own health, wisdom, security, happiness, and well-being.
- Justice. We should try to be fair and try to distribute benefits and burdens equably and evenly.

Defeasibility of duties

- Does it make sense to view duties as being defeasible?
- Can we apply defeasible reasoning to reason with duties?
- Should an AI system admit exceptions to duties, or should it always ask humans?

Nietzsche(1844-1900) a critique of ethics

- The superior human (Übermensch) is beyond the traditional views of good and bad, beyond the morality of the herd
- One has duties only toward one's equals; toward beings of a lower rank, one may act as one sees fit, 'as one's heart dictates'
- The superior human does not find or discover values, he (or she) determines the values
- No need to be ratified; the only criterion of wrongness is 'that which is harmful to me is harmful as such'

Contractarianism

Giovanni Sartor

Social contract theories

- In political theory:
  - A societal arrangement is just if it had (or would have had been) accepted by free and rational people

- In moral theory
  - actions are morally right just because they are permitted by rules that free, equal, and rational people would agree to live by, on the condition that others obey these rules as well (Shafer Landau)

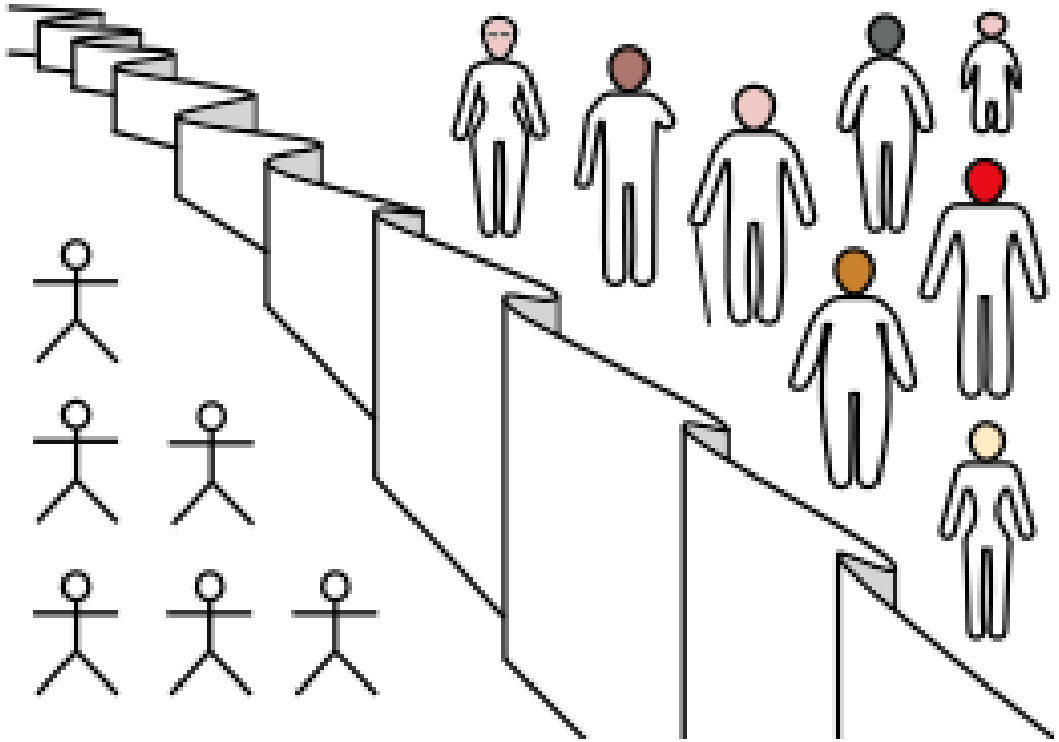
State of nature and social contract

- How to get out of the state of nature?
- What agreements are OK?

		B	
		Cooperate	Defect
A	Cooperate	4, 4	−2, 6
	Defect	6, −2	0, 0

John Rawls (1921-2002)

- A theory of justice
- How to ensue that the social contract is fair?
- People should choose under a **veil of ignorance**, without knowing their gender, social position, interests talents, wealth, race, etc.



What principles would they go for?

- **First Principle (having priority)**: Each person has the same indefeasible claim to a fully adequate scheme of equal basic liberties, which scheme is compatible with the same scheme of liberties for all (liberty of conscience and freedom of association, freedom of speech and liberty of the person, right to vote, etc.;
- **Second Principle**: Social and economic inequalities are to satisfy two conditions:
  - They are to be attached to offices and positions open to all under conditions of *fair equality of opportunity*;
  - They are to be to the greatest benefit of the least-advantaged members of society (the *difference principle*). (*JF*, 42–43)

AI in a just society (according to Rawls)

- Does the deployment of AI in todays society fit Rawls' requirements
- When may it conflict with the basic liberties?
- When with fair equality of opportunity?
- When with the difference principle?

Juergen Habermas: Discourse Ethics

- A rule of action or choice is justified, and thus valid, only if all those affected by the rule or choice could accept it in a reasonable discourse.

- A norm is valid when the foreseeable consequences and side effects of its general observance for the interests and value orientations of each individual could be jointly accepted by all concerned without coercion

- The valid norms are those that would be the accepted outcome of an ” ideal speech situation”, in which all participants would be motivated solely by the desire to obtain a rational consensus and would evaluate each other’s assertions solely on the basis of reason and evidence, being free of any physical and psychological coercion

- This approach assumes that people are able to engage in discourse and converge on the recognition of reasons for norms and choices

Habermas and AI

- Would would we all agree if we engaged in an impartial discussion on how to use AI?
- Can we think of an AI system that engages in an impartial moral debate? What would it argue for?

Virtue ethics

Giovanni Sartor

Virtue ethics

- Ethics should not focus on norms nor on consequences
  - An act is morally right just because it is one that a virtuous person, acting in character, would do in that situation.
- Ethics is a complex matter
  - Since there are many virtues, the right act is that that would result from the mix of the relevant virtues: honesty; loyalty; courage; impartiality, wisdom, fidelity, generosity, compassion, etc.
- Ethics cannot be learned though a set of rules, it application requires practical wisdom

Issues

- How do we know what is virtues and what is not?
- How can we extract precise indications from an account of virtues and from virtuous examples? How much can we rely in tradition?
- What if virtues are in conflict?
- What are the paradigms of virtues to which we may refer to?

AI and virtue ethics

- Should we, as developer of AI systems, be virtuous? What character traits should we cultivate in us?
- Should AI applications (AI agents be virtuous)?
- How can virtues be learned?
- If from example, I can the training of an AI system lead to a virtuous behaviour of it?

Readings

- Shafer-Landau, R. (2018). The Fundamentals of Ethics. Oxford University Press.
- Singer, P. (2021). Ethics. In Encyclopedia Britannica: https://www.britannica.com/topic/ethics-philosophy

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423