Ethics in Artificial Intelligence (Module 1)

Last update: 08 March 2025

Academic Year 2024 – 2025
Alma Mater Studiorum · University of Bologna

Contents

1 Trustworthy AI

The European Commission's vision for artificial intelligence is based on three pillars:

1. Increase public and private investments,
2. Prepare for socio-economic changes (e.g., protect who gets substituted with AI),
3. Ensure a proper ethical and legal framework to strengthen European values.

To achieve this, in 2018 the Commission established the **High-Level Expert Group on Artificial Intelligence (AI HLEG)**: an independent group tasked to draft:

High-Level Expert Group on Artificial Intelligence (AI HLEG)

- Guidelines for AI ethics,
- Policy and investments recommendations.

1.1 AI HLEG's AI Ethics Guidelines

Voluntary framework addressed to all AI stakeholders (from designers to end-users) that bases AI trustworthiness on three components:

Lawful AI must adhere to laws and regulations. The main legal sources are: Lawful

1. EU primary law (i.e., EU Treaties and Fundamental Rights).
2. EU secondary law (e.g., GDPR, ...).
3. International treaties (e.g., UN Human Rights treaties, Council of Europe conventions, ...).
4. Member State laws.
5. Domain-specific laws (e.g., regulations for medical data, ...)

> **Remark.** The guidelines do not provide legal guidance. Therefore, this component is not explicitly covered in the document.

Ethical AI must be in line with ethical principles and values (i.e., moral AI) for which laws might be lacking or unsuited for the purpose. Ethical

Robust AI must be technically and socially robust in order to minimize intentional or unintentional harm. Robust

> **Remark.** Each individual component is necessary but not sufficient. Ideally, they should all be respected. If in practice there are tensions between them, it is responsibility of the society to align them.

> **Remark** (Law vs ethics)**.**
>
> **Law** Norms adopted and enforced by institutional entities. Law
>
> **Ethics** Norms that guide what should be done (instead of what can be done). It is rooted in shared societal values. Ethics

Example (Ethical washing)**.** To pursue their interests, some entities push to avoid regulations (which must be enforced) and state to adhere to ethical values (which are not explicitly enforced).

Example (Brussels effect)**.** Extension of EU regulations to other countries due to economic reasons (e.g., it is economically more convenient to have a single system respecting the EU's GDPR instead of having two separate ones).

The document itself is composed of three chapters:

Foundations of trustworthy AI Describes the ethical principles an AI should respect.

Realization of trustworthy AI Describes the requirements to achieve trustworthiness.

Assessment of trustworthy AI Describes trustworthiness assessment methods.

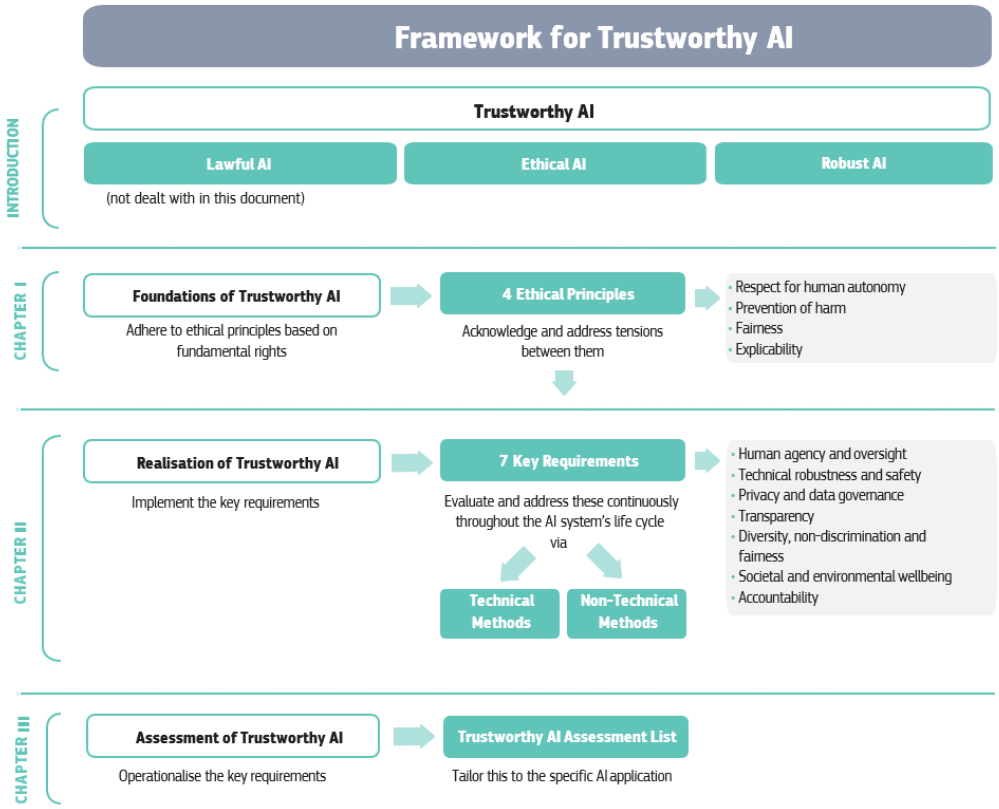

Figure 1.1: General overview of the document

1.1.1 Chapter I: Foundations of trustworthy AI

The concept of AI ethics presented in the framework is rooted to the fundamental rights described in the EU Treaties, EU Charter, and international human rights laws.

Remark (Fundamental rights)**.**

- Respect human dignity as moral subjects rather than objects in the pipeline of the system. AI systems should protect humans' physical and mental integrity, personal and cultural identity, and essential needs.
- Guarantee individual's freedom such as freedom of business, of the arts and science,

of expression, of assembly, and the right of privacy. AI systems should be mitigated for coercion, threats, surveillance, deception, ...

- Guarantee equality, non-discrimination, and solidarity. The output of an AI system should not be biased. Vulnerable groups that risk exclusion should be respected.
- Respect for democracy and citizen's rights. AI systems should not undermine democratic processes or citizen's rights such as the right to vote, to access public documents, to petition, ...

Remark. Seen as legally enforceable rights, fundamental rights can be considered as part of the LAWFUL AI component. Seen as the rights of everyone, from a moral status, they fall within the ETHICAL AI component.

This chapter describes four ethical principle for trustworthy AI based on fundamental rights:

Principle of respect for human autonomy AI users should keep full self-determination. AI systems should be human-centric leaving room for human choices and they should not manipulate them. Principle of respect for human autonomy

Principle of prevention of harm AI systems should operate in technically robust and safe environments. Attention must be paid to groups vulnerable to exclusion and to those subject to power asymmetries (e.g., employer-employee). Principle of prevention of harm

Principle of fairness The concept of fairness is described in a substantive and procedural dimension. The substantive dimension implies unbiased outputs and an equal distribution between benefits and costs. The procedural dimension involves the ability to contest and correct decisions made by AI systems and by humans using them. Principle of fairness

Principle of explicability AI systems need to be transparent, their capabilities and purpose should be communicated, and their decisions should be as explainable as possible. For black box algorithms, alternative explicability measures might be needed (e.g., traceability, auditability, and communication of capabilities). Also, the degree of explicability that is required is dependent on the context and the use case. Principle of explicability

Remark. There might be tensions between these principles (e.g., between prevention of harm and human autonomy in predictive policing) and methods to deal with them have to be established. Overall, the benefits of AI systems should exceed the risks. Practitioners should study these trade-offs in a reasoned and evidence-based way and not solely based on intuition.

1.1.2 Chapter II: Realization of trustworthy AI

This chapter defines concrete requirements from the principles of Chapter I. Stakeholders that these requirements involve are:

Developers Who research, design, and develop AI systems. They should concretely apply these requirements. Developers

Deployers Who use AI systems in their business processes and offer products or services to others. They should ensure that the systems they use meet the requirements. Deployers

End-users Who use the final AI system. They should be informed of these requirements and can request that they are respected. End-users

The main requirements the framework defines are:

Human agency and oversight AI systems should enhance human autonomy and decision-making (principle of respect for human autonomy): Human agency and oversight

- If there is the risk of violating fundamental rights, a study of the impacts should be conducted to justify it. External feedback should also be considered.
- Users should be provided with the necessary knowledge and tools to comprehend and interact with AI systems.
- Users have the right to not be subject to only automatic decisions if this significantly affects them.
- There should be oversight mechanisms (of varying degrees depending on the risk) to prevent AI systems from undermining human autonomy:
  - Human-in-the-loop (HITL): human intervention in every decision.
  - Human-on-the-loop (HOTL): human intervention in the design cycle and monitoring of the system's operation.
  - Human-in-command (HIC): human to decide if, when, and how to use an AI system in any particular situation.

  Public enforcers should also have the ability to exercise oversight with proper authorizations.

Technical robustness and safety There should be preventative measures to minimize unintentional harm (principle of prevention of harm): Technical robustness and safety

- AI systems should be protected against vulnerabilities and attacks that target the data (data poisoning), the model (model leakage), or the infrastructure.
- Possible unintended uses or abuse of the system should be taken into account and mitigated.
- There should be fallback plans in case of problems (e.g., switching from a statistical to a rule-based algorithm, asking a human, ...).
- There should be an explicit evaluation process to assess the accuracy of the AI system and determine its error rate.
- The output of an AI system should be reliable (robust to a wide range of inputs) and reproducible.

Privacy and data governance Quality and security of the data should be guaranteed through the lifecycle of the AI system (principle of prevention of harm): Privacy and data governance

- Data provided by the user and derived from it should be protected and not used unlawfully or unfairly.
- Datasets should be cleared from biases, inaccuracies, and errors before training.
- The integrity of the datasets must be ensured to prevent malicious attacks.
- Processes and datasets should be tested and documented.

Transparency There should be transparency in all the elements of an AI system (principle of explicability): Transparency

- The construction process of the dataset and the processes that lead to the AI system's decision should be documented.

- Decisions made by an AI system should be understandable and traceable by a human.
- The reason to use an AI system and the degree to which it influences decision-making and design choices should be stated.
- AI systems should not present themselves as humans and users have the right to be informed if they are interacting with an AI system. Depending on the use case, there should be the option to interact with a human.
- Capabilities and limitations of an AI system should be communicated to practitioners or end-users.

Diversity, non-discrimination, and fairness Inclusion and diversity should be considered in the entire lifecycle of an AI system (principle of fairness):

Diversity, non-discrimination, and fairness

- Biases should be removed from the data during the collection phase. Oversight processes should be put in place.
- AI systems should be user-centric and designed to be accessible by all people, regardless of disabilities.
- Stakeholders who might be affected by the AI system should be consulted.

Societal and environmental well-being The impact of AI systems should also consider society in general and the environment (principles of fairness and prevention of harm):

Societal and environmental well-being

- The environmental impact of the lifecycle of an AI system should be assessed.
- The effects of AI systems on people's physical and mental well-being, as well as institutions, democracy, and society should be assessed and monitored.

Accountability Clear responsibilities should be defined for decisions made by AI systems (principle of fairness):

Accountability

- Internal or external auditors should assess algorithms, data, and design processes.
- Potential negative impacts of AI systems should be identified, assessed, documented, and minimized.
- When there is tension between some of these requirements, trade-offs should be studied methodologically.
- There should be a redress mechanism for unjust decisions made by AI systems.

The chapter also describes some technical and non-technical methods to ensure trustworthy AI:

Technical methods

Technical methods

Architecture for trustworthy AI Embed trustworthiness requirements into the AI system as procedures or constraints.

Ethics and rule of law by design Methods to provide some properties by design.

Explanation methods Use techniques to understand the underlying mechanisms.

Testing and validating Define tests and validate the system in its entire lifecycle.

Quality of service indicators Use indicators to set the baseline for a trustworthy AI.

Non-technical methods Non-technical methods

Regulation Revise, adapt, or introduce regulations.

Codes of conduct Describe how the organization intends to use AI systems.

Standardization Define standards for a trustworthy system.

Certification Create organizations to attest that an AI system is trustworthy.

Accountability via governance frameworks Organizations should appoint a person or a board for decisions regarding ethics.

Education and awareness Educate, and train involved stakeholders.

Stakeholder participation and social dialogue Ensure open discussions between stakeholders and involve the general public.

Diversity and inclusive design teams The team working on an AI system should reflect the diversity of users and society.

1.1.3 Chapter III Assessment of trustworthy AI

This chapter defines a generic assessment list to implement the requirements of Chapter II. The list has been devised by first taking feedback from a small selection of companies, organizations, and institutions that implemented it. Then, it was extended to all stakeholders and another round of feedback was taken.

Assessment list Steps to concretely assess the trustworthiness of an AI system. The main considerations to take into account are that: Assessment list

- It should be tailored based on the specific use case.
- It can be integrated into existing governance mechanisms.
- It is continuously improved.

> **Remark.** In its pilot version, the list is composed of a series of questions for each requirement described in Chapter II.

2 Introduction to ethics

2.1 Morality

Morality There is no widely agreed definition of morality. On a high level, it refers to norms to determine which actions are right and wrong. Morality

2.1.1 Conventional and critical morality

Positive (conventional) morality Rules and principles created by humans that are widely accepted in a culture/society. Positive morality

> **Remark.** Conventional morality can change between different societies.
>
> **Remark.** In principle, the starting moral values of an individual are those of the society it was born into.

Critical morality Moral standards that are independent of conventional morality and are correct in general. It does not come from social agreements or beliefs. Critical morality

> **Remark.** Popular moral views are not necessarily true. Critical morality can be used as a ground-truth to determine whether conventional morality is correct.

2.1.2 Branches of moral philosophy

Value theory Field that studies values (i.e., source of goodness and badness). Some research questions are: "what is the good life?", "what is happiness?", ... Value theory

Normative ethics Field that studies what is morally required and how one should act. Some research questions are: "what makes right actions right?", "do the ends justify the means?", ... Normative ethics

Non-normative ethics

Descriptive ethics Field that uses scientific techniques to study how people reason and act. Descriptive ethics

Meta-ethics Field that analysis the language, concepts, and methods of reasoning in normative ethics. Some research questions are "can ethical judgments be true or false?", "does morality correspond to facts in the world?", ... Meta-ethics

2.1.3 Morality vs other normative systems

Laws Laws are not necessarily in line with what should be morally correct. Some legal actions are immoral (e.g., cheating) and some illegal actions are moral (e.g., criticizing a dictator). Morality vs laws

Etiquette Standards of etiquette and good manners are not necessarily moral (e.g., forks should be put to the left of the plate, but it is not immoral to put them to the right). Morality vs etiquette

Self-interest Sometimes, morality requires to sacrifice our well-being and, vice versa, immoral acts can improve our life. Morality vs self-interest

Tradition Practices that have been consolidated through time are not necessarily moral. Morality vs tradition

Religion Religions are based on the fact that there is a higher authority that created the set of moral norms. However, it is not clear whether God commands something because it considers them moral or things are moral because God commanded them. Morality vs religion

2.1.4 Absolutism and relativism

Absolutism There is a single true ethics. Absolutism

Relativism Judgment is relative to particular frameworks and attitudes. Relativism

2.2 Consequentialism

Optimific action Action that produces the best overall results. Optimific action

Consequentialism Family of theories that consider an action morally required if and only if it is optimific. Consequentialism

Remark. Consequentialism sees morality as an optimization problem.

2.2.1 Act utilitarianism

Act utilitarianism Instance of consequentialism that considers well-being the only thing that is intrinsically valuable. Act utilitarianism

Principle of utility Moral standard of act utilitarianism that considers an action morally required if and only if it is, among all the possibilities, the one that improves the overall world's well-being the most. Principle of utility

Remark. Strengths of act utilitarianism are:

- It is egalitarian and impartial as everyone's utility counts the same. Every person and every animal is part of the moral community.
- It justifies some basic moral intuitions (e.g., slavery is, in general, not optimific).
- It has by design a way of dealing with moral conflicts (i.e., choose the action that maximizes well-being).
- It is flexible as actions violating some moral rules can be performed if they increase the overall well-being.

Remark. Problems of act utilitarianism are:

- It is too demanding on the individuals as it requires constant self-sacrifice.
- It does not provide a decision procedure or a way to assess decisions.
- It has no room for impartiality (i.e., a family member is as important as a stranger).
- If the majority of society is against a minority group, unjust actions against the minority increases the overall world's well-being.

2.2.2 Rule utilitarianism

Optimific social rule Rule based on the idea that in the hypothetical case (nearly) everyone in a society were to accept it, the results would be optimific. Optimific social rule

Rule utilitarianism An action is morally right if it is required by an optimific social rule. Rule utilitarianism

Remark. Rule utilitarianism allows some degree of partiality. Moreover, an optimific action itself is not necessarily morally required if it is against an optimific social rule (e.g., it might be an optimific action to torture a prisoner, however, torture is forbidden by optimific social rules as it is more beneficial in the long term).

2.3 Deontology

Deontology Ethical theory which states that actions are good or bad independently of their consequences. An action is morally right if it conforms with a moral norm. Deontology

2.3.1 Kantian ethics

Kantian ethics Ethical theory based on fairness and consistency.

Remark. Two popular morality tests that however fail to assess fairness and consistency are:

- "What if everyone did that?", which can be interpreted as "if disastrous results would occur if everyone did X, then X is immoral". This test is sensitive to how the action is described as the same action can be morally right and wrong depending on its description.
- "How would you like it if I did that to you?", which is the golden rule. This test makes morality depend on a person's desire, which fails when the principles of a person are wrong (e.g., some nazis would accept to be killed if they discovered Jewish ancestors).

Maxim Subjective principle that one gives to itself when performing an action. In other words, it states what one is about to do and why. Maxim

Remark. Maxims are related to the individual's intentions.

Remark. Differently from consequentialism, morality in Kantian ethics does not depend on the results but on the reason of the actions. Two people doing the same action with identical results might be guided by different maxims.

Remark. By focusing on maxims, morality only depends on what is within our control. Result-oriented approaches are instead not always predictable and therefore should not be used.

Universalizable maxim A maxim is universalizable if it passes the following test: Universalizable maxim

1. Formulate my maxim clearly.
2. Hypothesize a world where everyone supports and acts according to that maxim.
3. Ask whether the goal of my action can be achieved in such a world.

Remark. Differently from consequentialism, the question asked aims at determining if our goal can be reached instead of determining whether the world would be better.

Remark. If a maxim is universalizable, it would mean that everyone could support it and our actions are not making an unfair exception for ourselves.

Principle of universalizability Basis of Kantian ethics. An action is acceptable if and only if its maxim is universalizable. Principle of universalizability

Acting on non-universalized maxims makes us inconsistent. Immoral actions can be therefore considered irrational.

Remark (Amoralist's challenge)**.** Amoralists are those that believe in right and wrong but act disregarding morality nevertheless. They challenge the fact that immoral actions are irrational as follows:

1. People have a reason to do something if it will give them what they want.
2. Moral duty sometimes fails to give people what they want.
3. Therefore, people sometimes do not have a reason to act morally.
4. If there is no reason to act morally, violating moral duties is rational.
5. Therefore, it is rational to violate moral duties.

Hypothetical imperatives Imperatives that require us to do what is needed to reach our goal. They are dependent on the individual's needs and disregarding them makes one irrational. Hypothetical imperatives

Remark. Hypothetical imperatives change if one's desires change.

Remark. Hypothetical imperatives are the way Kantian ethics deals with the amoralist's challenge.

Categorical imperatives Imperatives that do not depend on a single individual but are applicable to every rational beings. Categorical imperatives command to do things that one might want or not want to do. Disregarding them makes one irrational. Categorical imperatives

Remark. According to Kant's *argument for the irrationality of immorality*, moral duties are categorical imperatives:

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, you act morally.
5. Therefore, if you act immorally, you are irrational.

Principle of humanity Alternative interpretation of categorical imperatives. It states that one should treat humanity as an end and never only as means. In other words, one should never treat people without considering their dignity (i.e., intended as one's reason and autonomy). Principle of humanity

Remark. Kantian ethics revolves around integrity, which requires living in harmony with the principle one believes in. However, it does not capture the fact that if these principles are flawed, it would be morally more correct to have less integrity.

2.3.2 David Ross's prima facie duties

Prima facie duties Ethics theory which states that everyone has obligations that are however defeasible in case of tensions and conflicts. These obligations are: Prima facie duties

Fidelity Keep promises, be honest and truthful.

Reparation Make amends when we have wronged someone else.

Gratitude Be grateful to others whose actions benefit us. Try to return the favor.

Non-maleficence Refrain from harming others.

Beneficence Be kind and improve others.

Self-improvement Improve our own health, wisdom, security, happiness, ...

Justice Be fair and distribute benefits and burdens equably.

2.3.3 Nietzsche's critique of ethics

Superior human Who is beyond the morality of the common people. It only has duties toward equals and can treat beings of lower rank as it wishes. Superior human

2.4 Proceduralism

Proceduralism Approach to ethics that does not start by make assumptions on any basic moral views but rather follows a procedure to show that they are morally right. Proceduralism

> **Remark.** The golden rule, rule consequentialism, Kant's principle of universalizability are all instances of proceduralism.

> **Remark.** These theories are still originated from Kantian ethics.

2.4.1 Contractarianism

Contractarianism (political) Political theory which states that laws are just if and only if they would be accepted by free, equal, and rational people. Contractarianism (political)

> **Remark.** Rationality implies that everyone will cooperate limiting self-interest. In this way, everyone will give up a luxurious life but also avoid a terrible one.

Contractarianism (moral) Ethical theory which states that actions are morally right if and only if they would be accepted by free, equal, and rational people, on the condition that everyone obey to these rules. Contractarianism (moral)

Prisoner's dilemma Situation where the best outcome would be obtained if everyone stops pursuing their self-interest. Prisoner's dilemma

Table 2.1: Scenario that the dilemma takes inspiration from: two prisoners are interrogated separately, they can either stay silent (cooperate) or snitch the other (betray). The numbers are the years in prison each of them would get.

	Cooperate	Betray
Cooperate	2, 2	6, 0
Betray	0, 6	4, 4

State of nature Situation where there is no government, central authority, or any group than enforces its will on others. In such a situation, everyone acts to maximize its own self-interest. However, the effect is that everyone will be in worse conditions. State of nature

To escape from the state of nature, two things are needed:

- Beneficial rules that require cooperation and punish betrayal.
- An enforcer that ensures the rules are obeyed.

Contractarianism characteristics

- Morality is a social phenomenon: moral rules are basically rules of cooperation. There are no self-regarding moral duties, so any action that do not have bearing on others is morally right.
- Basic moral rules are justified.

  Veil of ignorance A test where rational people choose social rules solely based on their basic human needs (without other factors such as religion, ethnicity, sex, ...). In this scenario, it is expected that choices are made based on two principles: Veil of ignorance

  1. Each person will prioritize basic liberties, which will match those of everyone.
  2. Social and economic inequalities are allowed if everyone has equal access to those positions and the benefits should be aimed to the least advantaged members of society.

  Overall, what will be selected is going match the basic moral rules.
- There is a procedure to determine if an action is right or wrong: ask whether free, equal, and rational people would agree to rules that allow that action.
- Contractarianism justifies the origin of morality as originated from the same society we live in, but in a more rational and free version.
- Moral rules can be violated when people stop cooperating (i.e., when there is a state of nature).
- Contractarianism justifies the basic moral duty to obey the law, as otherwise there would be a state of nature. For the same reason, it justifies legal punishment and gives the state the authority in criminal law.
- By definition, contractarianism justifies breaking the law through non-violent civil disobedience when the law itself fails to set fair cooperation conditions.

2.4.2 Habermas' discourse ethics

Discourse ethics Ethical theory which states that an action is justified if and only if all those affected could accept it in a reasonable discourse (i.e., everyone involved is considered equal and free). Discourse ethics

> **Remark.** It is assumed that people are able to engage in a discourse and converge to a common choice.

2.5 Virtue ethics

Virtue ethics Family of theories that considers an action morally right if and only if a virtuous person (i.e., an ideal character, a role model) would do. Virtue ethics

Remark. Virtue ethics rejects the idea of having a simple test to determine what is morally right.

Remark. A virtuous person is different from one that habitually do the right thing. The former compared to the latter is motivated in doing so and has a greater sensitivity.

Moral wisdom Know-how that tells us what is morally right. It is developed through experience starting from the moral rules one have been taught as a child. Moral wisdom

Remark (Emotions in moral understanding)**.** Emotions can provide a signal to determine the rightness of actions:

- Emotions suggest what is morally relevant in a given situation (e.g., compassion, sympathy, kindness, *etc.* are virtues a moral person should have).
- Emotions tell us what is morally right and wrong (e.g., anxiety when doing something immoral).
- Emotions help to motivate us in doing the right thing.

Virtue ethics problems

- Virtue ethics does not provide rules to deal with conflicts between virtues.
- Virtue ethics might be too demanding depending on the standard set by the virtuous people.
- There is no absolute rule to select which virtuous people to consider. Depending on many factors, different people might choose different role models.
- There is no way to deal with disagreeing virtuous people.
- Virtue ethics considers actions right if done by virtuous people (and not the contrary, i.e., people are virtuous if they perform right actions). This boils down to the same problem of religion described in Section 2.1.3.

2.6 Principlism

Principlism Ethical theory originated from bioethics. It divides morality into two categories: Principlism

Common morality Moral norms shared by all individuals. It supports human rights and moral ideals such as charity and generosity. Common morality

Remark. Common morality follows absolutism and derives from human experience.

Particular morality Specific and content-rich norms that should not violate common morality. Particular morality

Professional morality Moral norms specific to a profession.

Public policy Regulations and guidelines promulgated by institutions.

Principles General guidelines for the formulation of rules. The moral principles are: (i) respect for autonomy, (ii) non-maleficence, (iii) beneficence, and (iv) justice. Principles

Rules Instantiation of principles that are more specific in content and scope. Rules

Substantive rules Rules of truth telling, confidentiality, informed consent, ... Substantive rules

Authority rules Rules that establish who can make decisions and perform actions. Authority rules

Procedural rules Rules that establish a procedure to follow. They are the last resort if substantive and authority rules are not suited. Procedural rules