Ethical frameworks for cybersecurity

Michele Loi - University of Zurich

Reasonable expectations about moral theory

Moral theory is systematic reflection on ethical claims and the reasons for them

1. Moral theory will not give you a formula
2. Moral theory helps you to recognize ‘the shape of’ an ethical problem
3. It will never substitute your moral conscience
4. It will never replace the advice of domain experts

How will moral theory help you?

1. You will familiarize with common ’patterns’
2. You will familiarize with established modes of reasoning
3. You will become aware of the limits
4. You can better organize the ‘moral data’

Some of the most widely used approaches are:

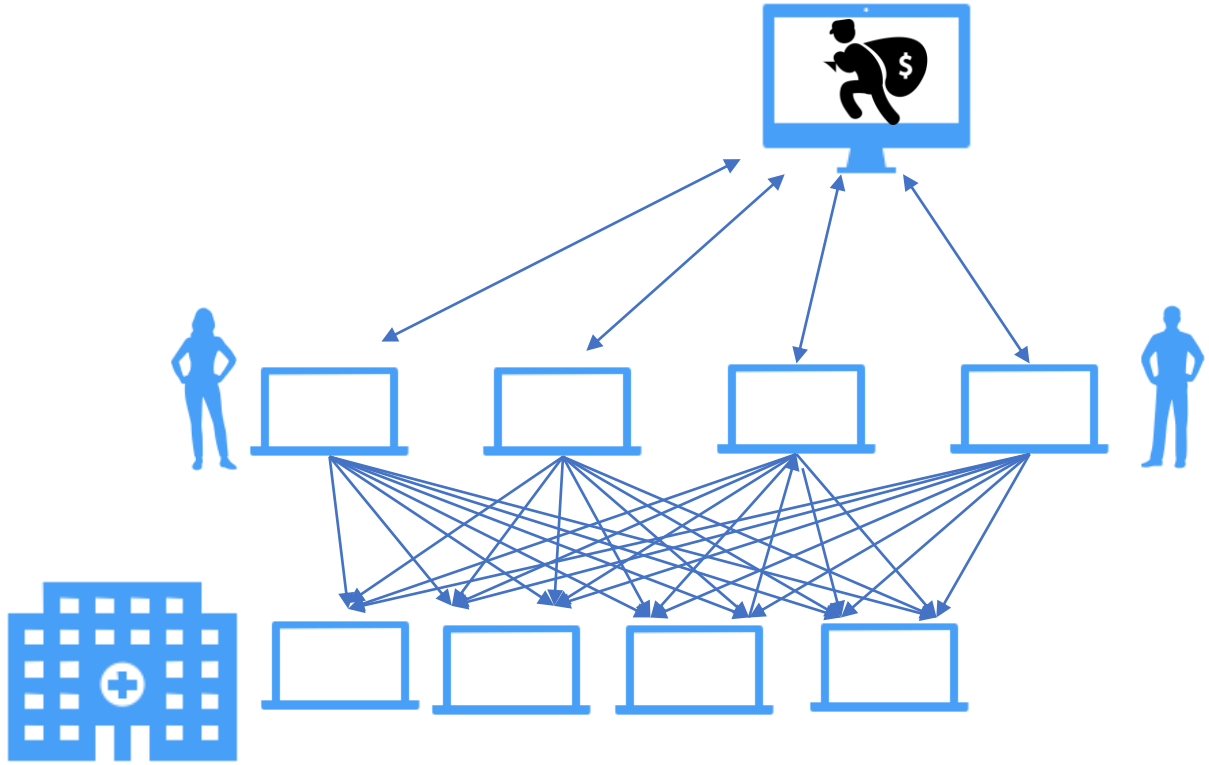
- **Principlism**
  - 4 principle version (3 principle version in Belmont and Menlo report: Benevolence, Respect for persons and Justice.)
- **Human rights**
- **Utilitarianism**
- **Contractualism**

.

[GET MORE: **MOOC 'introducing the terminology'**]

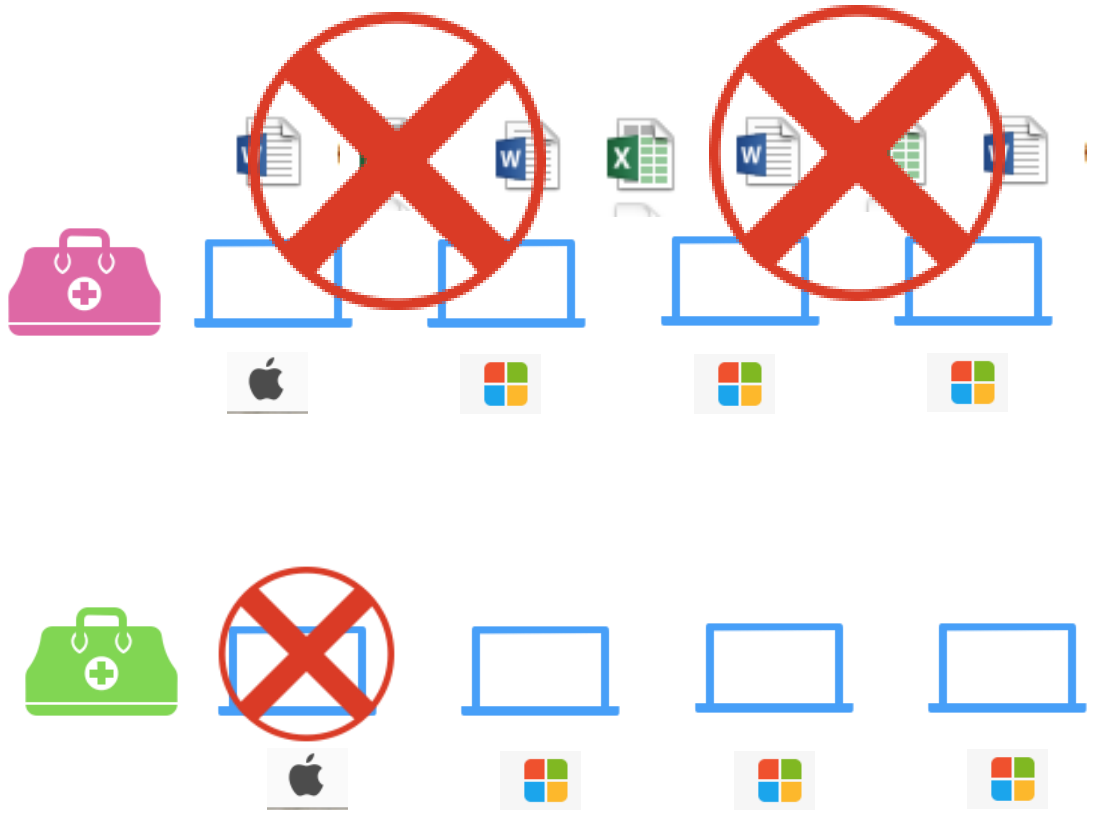
A choice of anti-malware:

You are dealing with **malware** that turns the affected computers into nodes in a botnet performing a distributed **denial-of-service attack** against servers in an important **hospita**l, which risks placing the health of its patients at risk. The **malware is designed to retaliate by wiping out the entire hard disk, as soon as it is disconnected from the malicious server**. A preliminary study of the malware shows that it could be fought with **two different software approaches**. Each of them fails in specific ways to limit the damage. Due to time and resource constraints, **you can develop only one** of them before the malware spreads causing harm to computers from which the health of patients depend. Which one do you develop?



The dilemma:

- A. it protects all computers but deletes all excel and word files during installation.
- B. it only works on non-Apple operating systems
  - Apple systems will have to be quarantined and <u>will lose all data</u>.
  - 5% of the computers in the botnet are Apple ones.



Guided exercise

Principlism

- Beneficence:
  - A. & B. you promote the health of the patients connected to the system
- Non maleficence:
  - A. Damages all computer owners significantly
  - B. Damages a small proportions of owners, more severely
- Autonomy: no difference
- Justice:
  - A. distributes the costs equally
  - B. sacrifices a minority (Mac users) to generate a greater benefit

Human rights

- Human right to health of the patients
- No other human right is in place, able to differentiate between A and B
- Provides little guidance in the choice between A and B

Utilitarian

Option A: major damage to 100% of computers

Option B: even greater damage to 5% of computers, no damage at all to 95% of computers

Utilitarian

The total sum of damage is minimized in option B

That makes option B preferable

Contractualist

- Contractualism requires parties to agree on a fair rule.
- One established contractualist approach involves identifying *the strongest individual complaint* against each solution, to exclude that solution and to adopt the other solution:
  - A: deletes word and spreadsheet files
    - Strongest individual complaint: partial deletion of content
  - B: wipes out everything from Macs
    - Strongest individual complaint: total deletion of content

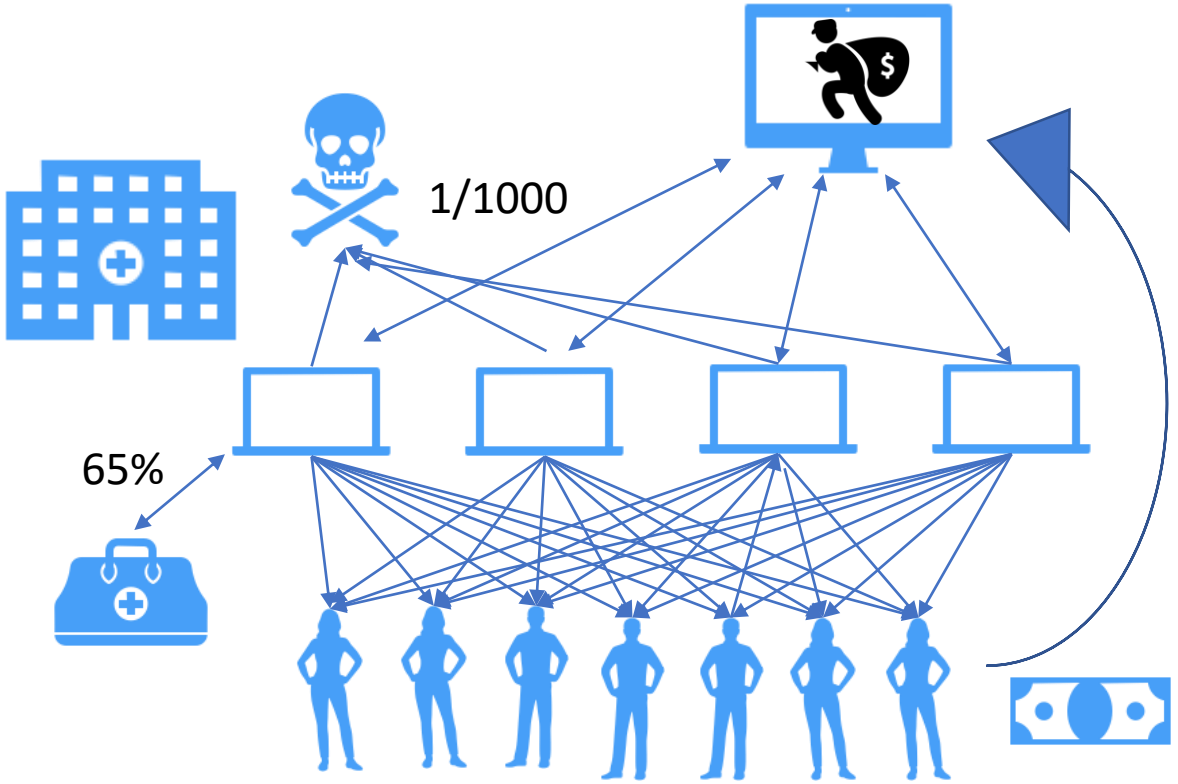
The strongest *individual* complaint is the complaint of Mac users against approach B

Example of a difficult cybersecurity case

- Some cybersecurity can involve hard-to-predict risks and potential benefits
- Even when the risks are known, assessing risk from an ethical perspective can be difficult

Responding to ransomware:

You are the **leader of a CERT team** and you have identified **ransomware** (a software virus that encrypts the data in the computers infected and directs the victims to a payment service where, after paying **1000€, the victims can obtain the decryption key**). A partner software company has already begun developing a decrypt tool; you estimate that the company **has a 65% chance of success within one** month (and close to no chance of succeeding later). At the moment, **1000 computers are affected, all belonging to the network of an important hospital**. Unfortunately, it is impossible to reconstruct what data was saved in each computer and the date of the latest backup. It is known that **each computer could be critical for the life of a patient, but it is not known which computer is critical for which patient**. There are **1000** patients and the probability that an alteration or deletion of data in a single computer will cause the death of the patient connected to it **is 1/1000** for each device (assume these are independent events).



Icon made by Freepik from www.flaticon.com

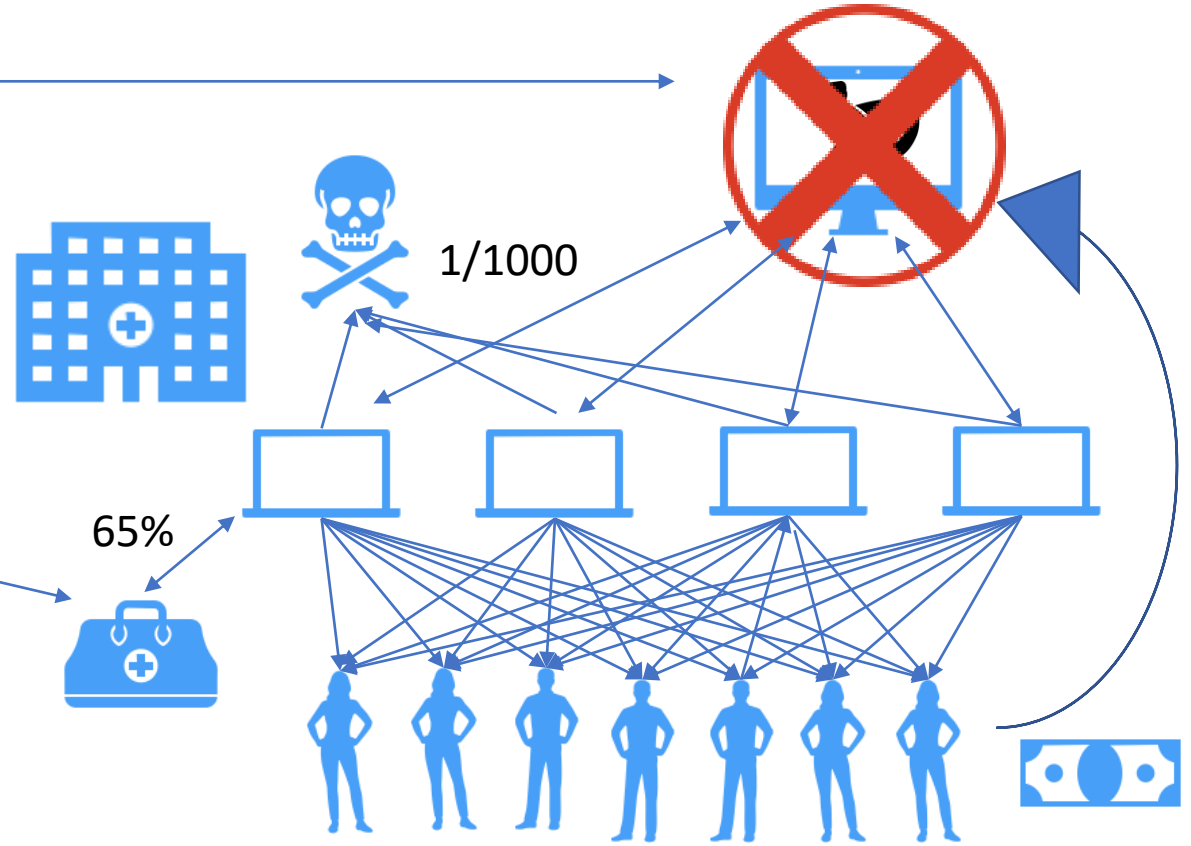
Your options

- *Policy A:* shoot down payment server

you **quarantine** all the affected computers and **shoot down the payment servers**. These measures will prevent **the spread** of the infection and reduce the **incentives** for attackers to involve other computers in similar attacks in the near future. However, the **malware** is designed to detect your response and retaliate to it. It will **irreversibly introduce random changes in the data, in ways that are extremely hard to detect, or simply delete it**. It is not possible to identify **what patients are affected** in a reasonable amount of time.

- *Policy B:* wait 1 month

you **do not isolate the affected system** and do not bring down the payment server; after one month, either you have obtained the **decrypting tool with no losses**; or you have not, in which case the infection will have spread to **other 1,000,000 computers** overall in the following months, with an expected aggregate economic loss for your society of **400,000,000 €**, with a maximum individual harm of **1000€** for each person affected.



Icon made by Freepik from www.flaticon.com

Principlism

- Beneficence:
- Non-maleficence:
  - Policy A: a 0.63 chance (1-(999/1000)^{1000}) that at least one person will die
  - Policy B: 0.35 chance of failure of the decryption tool; possibility of a large aggregate loss, 500€ average damage, max damage of 1000€.
- Autonomy / respect for persons:
  - Policy A: you *impose* a risk on patient, against their consent;
  - Policy B: you *allow* internet users to be exploited by hackers, *against their consent*
    - Which is worse? Arguably A.
- Justice:
  - Are you giving equal weights to the claims of patients and users (in A or B)?

  It is difficult to identify a 'common measure' -> see utilitarianism VS contractualism.

Human rights

- Human right to life; B>A
- Human right to property (art. 17 Universal Declaration of Human Rights) A>B
- The human right to life is both a more established right and generally one with a higher priority

B = wait one month

Utilitarianism

- **Expected utility utilitarianism**. This is standard cost-benefit analysis where we calculate the expected value of each option
- Expected value of option = value of the actually occurring outcome of that option * probability of the outcome.
- What makes this case problematic is the nature of the comparison of costs and benefits.
  - **option A harm =** 0.001*1000*value of a patient's life in €
  - **option B harm =** 0.35*400,000,000 €

  Which one is greater?

Contractualism

Let us determine *the strongest individual complaint* against each solution:

- A: quarantine immediately
  - Strongest individual complaint: being exposed to a 1/1000 chance of losing life
- B: wait one month
  - Strongest individual complaint: being exposed to a 35% chance of losing 1000 euro

Which individual complaint is stronger overall? If you think the risk of losing a life gives one a stronger complaint, the fair rule prefers B.

Challenges of all these approaches:

1. Probabilities difficult to assess
   - In the example: magnitude of harm * probability both known
   - In reality: neither known (or very vaguely approximated)
2. Comparison of different goods
   - Value of human life vs. economic harm
   - How much is a human life valuable in €?
3. Moral assessment of risk:
   - What is worse, a 1/1000 risk of death OR a 35% risk of losing 1000€?
   - Are these risk assessments subjective? Or can they be generalized?

Utilitarianism: the value of human life

Legal approach (in settling for damages)

Value of life = e.g. 3.4 million US dollars for the son of two university graduates

Metrics needed = statistical expectations of future salaries (in society *as it is*).

If so:

Value of male fetus>value of 6 year old girl

Value of white person > value of latino person (in USA)

Ethically controversial!

The Washington Post

Wonkblog

In one corner of the law, minorities and women are often valued less

By Kim Soffen

Oct. 25, 2016

Source: https://www.washingtonpost.com/graphics/business/wonk/settlements/?tid=ss_fb

Summary table

Theory	Choice	Justification	Difficulties
Principlism	B	allow a wrong > impose a wrong	-Assigning weights for non-maleficence - Doing/allowing
Human rights	B	life > property	
Utilitarianism	A	B IFF x >140,000,000 € x = value of a human life in €	Value of a human life? Aggregated small harm > value of an individual life
Contractualism	B	0.35 risk of a 1000 € loss >0.0001 risk of life	Moral importance of different vulnerabilities: subjective or objective?

Questions:

- Which approach is the most convincing?
- What makes it so?
- Which approach is the least convincing?
- What makes it so?

Conclusion

Hopefully:

1. You are now more familiar with four moral theories
2. You are aware of typical problems in difficult cases
3. You are more convinced about the importance of moral reasoning, than you were before

To know more, see the chapter *Ethical Frameworks for Cybersecurity* in our forthcoming book *The Ethics of Cybersecurity*