# A Computer Scientist's View of Ethics

## Concepts of Ethics

> *"Confusion of goals and perfection of means seems, in my opinion, to characterize our age."*
> (Albert Einstein)

Historically it may be interesting to notice that Aristotle (384–322 B.C.), the founder of formal logic, has apparently also coined the term "ethics", and he wrote several works about it.

Ethics is the part of philosophy dealing with right and wrong conduct, that is, it revolves around the question "What should I do?" (one of Kant's famous Four Questions).

But how do we know what is right or wrong, good or bad? In mathematics we prove theorems, in the natural sciences we do experiments, but on which grounds do we decide on ethical issues? This principal question is one subject of **meta-ethics**, whereas **applied ethics** is about right actions and practical decisions in particular situations or domains, in our case CS. Here it is certainly not the aim to teach what the right conduct is. Rather, we survey some conceptual frameworks and use them as a basis for discussing specific examples of ethical issues in CS.

In every professional activity one has to make **decisions**. There are always several options. (Even in a work environment where one only needs to follow instructions one can ask oneself whether one should stay in this system.) And decisions have reasons. By analyzing these reasons, pros and cons, one does not only prepare decisions being on the agenda. One also gets aware of the own value system. It is logically impossible to have no ethical principles at all (in the same sense as it is logically impossible not to communicate), they only remain below the level of consciousness if one does not reflect upon them. Finally, ethical decisions have tangible results: They influence early design decisions for products which are hard to revise later, they can avoid disasters, promote the reputation of a company, and thus even have economical effects.

There are apparently some popular misunderstandings about ethics:

- Ethics is not a show-stopper and does not *only* deal with threats and undesirable consequences of, e.g., new technologies. It does not say that technology only adds new problems. The question "What should I do?" is unbiased. Ethics of technology is not technophobic. It only requires to think proactively about the right goals, possible consequences of developments, and appropriate precautions, rather than being completely occupied by the process of solving "given" technical problems.

- The name "applied ethics" might be misleading. One cannot just straightforwardly apply some ethical theories to specific cases. Rather, they are only used as a guide and conceptual framework, but the actual "ethics of something" must be built within the disciplines. This is also the reason why scientists and engineers should analyze the ethical questions of their own fields. Who else if not them? Only specialists in the field have a solid background, deep technical knowledge and domain competence needed for a serious analysis. They can formulate and ask the right questions.

- In particular, ethical dilemmas cannot be "solved" by applying some philosophical theories (in the same way as a computational problem is solved once and for all by an algorithm). A dilemma is a dilemma because it has no really satisfactory solution. Therefore, do not expect ultimate solutions, but only analysis of problems and options, and temporary (yet well motivated) conclusions.

- Ethics and morality are not exactly the same. The verdict "immoral" feels stronger than "unethical", and ethics refers more to the right conduct in a certain role or context.

## On Ethical Theories

**Consequentialism** judges actions based on their consequences. A right action is one with a good outcome. Still one needs to clarify what a good outcome is, and for whom, and how this is measured. Some directions of consequentialism are specific about that. For instance, **utilitarianism**

considers a good action one that increases a positive effect for a maximum number of people.

**Deontological ethics** considers a rule or principle of behaviour as good or bad in itself. What counts is the general intention rather than the actual consequences.

An extreme consequentialist position is to say "the ends justify the means". An extreme deontological position can be questionable, too. (For instance, while it is an inherently good principle not to lie, there can be circumstances where telling the truth is harmful.) One can also wonder what makes a rule good in itself, if not its consequences.

We do not have to decide upon one of these positions, and they do not exclude each other. But we should keep this pair of concepts in mind, and ask ourselves on which grounds we make our judgements in a specific case.

### Some Fun: What Justifies Principles?

Asking about the **principles** of ethical behaviour leads to a more general philosophical problem known as the Münchhausen trilemma. (It also appears in many other domains.) The name comes from the fictional Baron Münchhausen, a character in a book by Gottfried August Bürger. In one of these highly implausible stories, the Baron managed to pull himself out of the swamp by his own hair.

The trilemma is the following (this reasoning should be appealing to computer scientists and mathematicians): We ask why some principle $P_1$ is good. We justify it by some more general principle $P_2$ which implies that $P_1$ is good. But one could challenge $P_2$ as well, and we may respond with some even more fundamental principle $P_3$ that would imply $P_2$. And so on. Now there exist three options, therefore the name trilemma: (1) We reach some axiom $P_n$ that does not need further justification, and the backwards chain ends there. (2) Some $P_n$ is identical to $P_1$, that is, we run into a cycle. (3) The chain is infinite (called an infinite regress). None of these options looks satisfactory. In (1), what should stop us asking what justifies $P_n$? In (2), nothing seems to be really explained, the principles just "support each other". And (3) is obviously impractical, we can never see the entire chain.

A consequence is the impossibility to justify any theory by itself. Figuratively speaking, no theorist can "pull himself out of the swamp by his own hair". In practice it seems that we have only option (1), that is, we must accept principles that we find evident or sufficiently supported, and confess that we cannot "explain" it further. This is then the position from which we judge specific questions, e.g., in ethics.

## Ethical Values

There do exist some widely accepted, very general ethical principles; we use **principle** and **value** as synonyms from now on. These values include: causing no harm on other people, respecting their dignity and autonomy, treating people equally, being honest, being cooperative, giving credits (material or immaterial) to people who gave something, increasing the general welfare, caring about the environment, etc.

Why are these values accepted? A pragmatic justification is that a society massively violating them cannot function in the long run. This may be a good enough "$P_n$" that need not be challenged further. This approach justifies values by the negative consequences of not following them, therefore this position is called negative consequentialism. Similarly, we may attempt to "measure" the importance of values and sort (rank, grade) them based on the weight and importance of consequences.

In fact, values can be more or less important. Some are absolutely essential, others are only desirable. If a conflict of values arises, more important values can override less important ones. Formally one could think of a **value system** or **value hierarchy** as a (partial) order of values. When taking a concrete decision we would first try and respect all our values. If, in a specific case, strictly adhering to all values leads to unwanted consequences, it could be acceptable to sacrifice lower values and stick to the higher ones, as far as possible.

Accordingly, values are often divided in two categories: An **instrumental value** is desirable and is a means towards achieving some other value, whereas an **intrinsic value** is considered an end-in-itself. These two types do not exclude each other, moreover, "being instrumental" can express a relation rather than an attributes: a value can be *instrumental for* another one, and it is then lower in the value hierarchy. Here are a few examples, among them one with more discussion.

*Knowing science* can be considered an intrinsic value (understanding the world, curiosity, intellectual pleasure) but is obviously also an instrumental value (basis of technologies, making life esier).

*Transparency* seems to be an instrumental value only. It depends on the context whether it is desirable or not, feasible or not: Transparency of personal data records is usually not desired. An algorithm or system with transparent structure is easier to understand or to maintain; here transparancy is good. Transparency can also mean to reveal the details of a search engine or of an evaluation algorithm. Here it is less clear whether this should be required: Note that public knowledge of such details gives opportunities to manipulate the results.

*Mobility* (in a wide sense) seems to be only an instrumental value as well; think about it.

## A More Detailed Example

*Efficiency* of products (being small, fast, user-friendly, have low energy consumption, etc.) is certainly a value to aim for. Higher efficiency improves the usability and decreases running costs and consumption of energy. (Let us focus on energy, for simplicity.) All this sounds only good. Are there downsides, too? In the following discussion we neither claim that negative effects *will* appear nor can we give a comprehensive analysis of a complex matter. We only make the point that there *might* be unwanted consequences that are *not obvious* and deserve analysis, and efficiency is therefore more an instrumental rather than an intrinsic value.

An innovation that improves efficiency prompts many customers to buy the new version instantly. Isn't that good? They will save energy from now on. But before that, the production of the new devices costs additional energy. (Let alone material, development costs, etc.) In total, energy will be saved only if the customers use the new version long enough, such that this initial extra amount pays off. Furthermore, when innovations are frequent, many customers may always want the latest version, thus they exchange the product more often, the production increases, and more rather than less resources are used up in total. One can object that most customers think twice and do keep their recently bought products. The question is how rational they usually are, what influences their decisions, and what the suppliers should conclude from that. In fact, there is a good reason to buy a more efficient new product immediately: If one waits, then the running costs will (unnecessarily!) be higher, until the new product is bought anyway, thus it is better to buy immediately. The only rational reason to wait is to speculate on an even better version appearing in the near future. Then it is cheaper to skip one version. Now, how do people make such decisions without knowing the future? There is a whole, mature field of algorithmics (!) dealing with decisions under uncertainty and so called online problems. (Here, the term "online" refers to problems where the input is revealed step by step, but one must continuously take decisions without knowing the future part of the input.) This research direction yields good strategies for many cost minimization problems, but this does not mean that people intuitively apply such strategies. One can assume that the supplier has more competence also on the economical aspects, and hence responsibility for marketing decisions. Advertising and providing information on coming improvements influence customer decisions. Pricing is, of course, another major parameter. One observation is that, to some extent, higher prices both increase the profit and prevent many users from too many purchases.

Assume that decision makers in a company have a model of user behaviour that more or less reflects reality (which is difficult enough). Still it remains the question what goals govern the use of that model, and here ethical values come in: Do we formulate a profit maximization problem only, or a multi-criteria problem that also tries to minimize the total energy consumption (more generally, the social costs)?

In this context one can also deeply discuss "planned obsolescence", a design of products that deliberately limits their lifetimes although they could physically be working for longer periods. Such a policy can be ethical or unethical depending on the intentions.

Altogether, a seemingly simple situation turned out, after some thinking, to be a very complex system, involving nontrivial modelling questions, algorithmic problems, and ethical decisions ("What do we try to optimize?"). Part of the problem is that different agents (supplier, individual customers) and society as a whole may have divergent interests.

### Some Meta-Ethics Questions around Values and Consequences

Here we list some questions that can be debated. There are hardly final "correct" answers to them. But maybe you want to think about some.

- Do propositions about ethics express objective facts and have truth values (true or false), exactly like propositions in mathematics or natural sciences, or are they just opinions, or something in between?

- Are ethical values universal, or are they always relative and subject to change, depending on real conditions? In other words: Are value systems static or dynamic? – At least, one can observe that new technologies often force us to think about new ethical questions that were simply not relevant before. On the practical side, unanswered ethical questions leave a "policy vacuum". Also old ethical questions may get new shape and need new discussion as a result of new technical possibilities.

- Regarding the different importance of values: Has minimizing bad consequences ("pain") priority over maximizing good consequence ("pleasure")? Can good consequences outweigh other, bad ones, that we therefore find acceptable? This type of questions can be practically relevant when we model a decision as an optimization problem: How do we form our optimization goal?

- Not all consequences are foreseen. Actual results may differ from the intentions. Nobody should be blamed for a bad outcome that was really unpredictable, even if this person's action provably generated this bad outcome in a causal chain of events. (Npte that a deontological position is applied here.) But how strong is the obligation to investigate all possible circumstances and consequences before acting?
- Is there a principal difference between actions and omissions, or is a deliberate decision not to act a special case of an action? This might appear as a purely academic question, but we can formulate it more practically as well: Is there an ethical obligation to do some (or even every?) good thing that could possibly be done, or is this too demanding, and only explicitly bad actions are ethically prohibited? Apparently there must be a cut-off point – but how is it determined?
- Ethical responsibility is, in general, larger than just accountability. As an example: The person that runs a system is in charge for avoiding possible harm caused by that system. But already the designer of the system could avoid predictable harm.
- Short-term and long-term consequences may differ a lot. (And optimal short-term decisions need not yield an overall optimal result even in simple scenarios – see greedy algorithms!) What is the right time horizon to consider in decision making?

### Formalizing Ethics?

Ethical principles should at least be coherent, that is, not be against logic. Moreover, they should be general and applicable to many cases, not only to those that gave rise to their formulation.

> *"Act only according to that maxim whereby you can at the same time will that it should become a universal law without contradiction."*
> *"Act in such a way that you treat humanity, whether in your own person or in the person of any other, never merely as a means to an end, but always at the same time as an end."*

> (Immanuel Kant. These quotes are known as the Categorical Imperative.)

An interesting question is whether formal systems of ethics can be created. More specifically and practically asked: Can we (in principle) “teach” robots and other complex autonomoous AI systems the right behaviour in a wealth of situations? Complete formalization is clearly far too ambitious, but some more modest goals are realistic: Design systems and implement rules in such a way that at least certain predictable types of unintended actions are excluded. There are activities towards forming such a robot ethics (see, e.g., the conference series WeRobot.)

Trying and formalizing limited parts of ethics and studying the logical implications also helps reveal inconsistencies and gaps in ethical theories.

## Aggregating Utilities?

This matter is discussed in some ethics literature. It has some generality, a high level of abstraction, and an optimization flavour, so it should also be fun for CS people (without forgetting the serious background).

Suppose we have the choice between several actions and can quantify the utility of every possible action for every individual in a population, in other words, we can exactly model their “happiness”. This assumption is, of course, already problematic, but it can be a reasonable approximation of truth in many cases. Let us work with the assumption. Then, every action has a vector of $n$ utilities, where $n$ is the number of persons. What action is the best to choose?

This is quite clear if some utility vector dominates all others, i.e., is component-wise the largest. Then we should choose this action. But what is the right choice in all other cases?

One standard approach would be to maximize the sum (equivalently: average) of utilities. But there could be solutions where some persons get a huge utility and all others are treated badly. The happy ones are sometimes called “utility monsters”. Intuitively this is considered very unfair, because a utility monster is not representative for the average happiness in a population.

An obvious idea to avoid such disfavour is to maximize the minimum of all $n$ utilities. But now there are examples where already a slight increase of the lowest utility is possible only at cost of making many other utilities much worse. Intuitively such a solution would be considered unreasonable, too.

On can try more refined optimization goals, for instance, maximize the median utility. Note that median is not the same as average, and the median utility represents the typical utility in a population better than the average, even if distribution of utilties is skewed. But similarly as above, the increase of the median can also disfavour many other individuals.

Let us consider a natural special case of our abstract setting: A fixed total amount of some goods can be divided and given to the individuals. Each individual has a utility function $u(x)$ which denotes the utility when $x$ units of the good are received. If the $u(x)$ are linear functions (utility is proportional to the amount) then the optimal solution is, in fact, to give everything to the utility monster with the highest ratio $u(x)/x$. Luckily, utility functions typically seem to be concave in reality. In this case the optimum is an equilibrium state where the derivatives $\frac{d\,u(x)}{dx}$ are equal for all individuals, which feels more acceptable. (Mathematical details are simple, think about them.)

The question which population is happier, based on vectors of utility values only, leads to several paradoxical conclusions discussed under the name “mere addition paradox”. (See, for instance: N. Hassoun: Another mere addition paradox? Some reflections on variable population poverty measurement. UNU-WIDER, UN University, working paper 2010/120.) It seems that every optimization goal can be fooled, in the sense that it gives counterintuitive or undesirable results in certain cases.

Now one might object that the whole discussion is artificial and purely academic, and it only origins from an attempt to compare incomparable objects: Utility vectors form only a partial order, and somehow we want to define a “better-than” relation that is a total order. But the point is that this type of decisions does appear in reality, whether we like it or not. Moreover, if “optimal” decisions are to be made by computer programs, we are the people who define the optimization goals.

In the beginning we had distinguished between technical and ethical questions. Once an optimization problem is defined, solving it is a purely technical matter. But there are no mathematical criteria for choosing the “right problem to solve”. Defining an optimization goal is a modelling step that has an ethical component, especially if it directly affects people. (More concrete examples follow later.) **Modelling** is an important keyword here.

*“All models are wrong, but some are useful.”* (George Box)

Assuming that a utility is a single number and that the total benefit for $n$ individuals is merely a function of these $n$ numbers is already a strong abstraction and simplification, and this point is easy to forget. Still the assumption is justified in many cases and makes decision problems manageable. We only have to keep in mind that *it is* an abstraction. The implications are: When adopting a model, always ask yourself what has been "abstracted away". Consider the model only as a guide, do not automatically follow the final optimization result, check it against reality, perhaps refine the model, etc.

A more fundamental criticism is that human individuals should never be considered part of some gross utility function. But this criticism seems to be over target. Such considerations cannot be avoided, and using them as models is fine, as long as humans are not really considered to be instrumental values, and their moral rights are respected.

## Examples of Ethical Issues in (Computer) Science

Nowadays ethical questions in CS are even a frequent topic in the media, and a curriculum of ethics in CS is taking shape. The following lists of examples are by no means systematic or complete. These examples are mostly described without sources, and they are mainly intended as inspiration. Suggestions of further examples are welcome, too. You will notice that some of the examples are instantiations of the "aggregating utilities" issue.

### Examples in Science and Engineering in General

- Are scientists and engineers responsible for their discoveries and inventions at all? Can we separate "pure " science and development from the (actual or potential) implications of these activities, or do we have to think about ethical issues right from the beginning, as an integral part of research and development? Are researchers even responsible for unintended future uses of their work? And can an innovation as such already be ethical or unethical?

- An instance of the actions-versus-omissions question: Is it unethical not to develop or improve some technology, although this would be manageable and the benefits for many people would be obvious?

- If some resources are scarce, where do we put them, which people do we help, what do we prioritize – and by what criteria? (Think of expensive development of new medical treatments. One cannot do all possible projects simultaneously.)

- How do we protect the rights and interests of people being involved in scientific (e.g., medical, psychological, sociological) studies? Some issues are: avoiding any kind of harm, protecting anonymity and privacy of data.

- Risk versus cost: There can be a trade-off between costs and reliability of a product or service, including avoidance of disasters. What are criteria for the right balance?

- Engineers and scientists take decisions *for other people.* Who has the benefits, who takes the risks, who bears the costs, and who decides? And are these the same groups or different groups of people? In the latter case, what are the ethical implications? (For instance, is it ethical to expose other people to risks without asking?) Often these issues are not so visible, as engineers do not have strong client relations (compared to, for instance, physicians or lawyers). A widely accepted policy in *medical* ethics is "informed consent": patients are informed about benefits and risks and then approve or deny the treatment. But in big technological projects affecting many people, a similar procedure is not even feasible. What rules should be followed instead in the political decision processes?

- How should scientists communicate possible dangers they have detected (for instance, prediction of earthquakes)? One of the subtle points is how to communicate the intrinsic uncertainty of the prediction in an understandable way. The question has numerous facets: modelling issues, cost-risk trade-off, imposing risks on others, legal aspects, etc.

- Technology solves existing problems and frees people from stupid or strenuous work. But once a technology is established, it also creates new desires and demands. People are expected to do more complex work more efficiently, and (thanks to communicaton devices) to be available permanently. Pressure and work density can increase rather than decrease. Employees can get a feeling that they never finish their work in a satisfactory status. – It would be naive to blame technology as such for all these issues. But what is the right approach to these problems?

- Technologies are accepted, as it is simply convenient to use them. (By the way, is convenience an intrinsic value?) But they also create dependency, loss of control, and unexpected bahaviour, as detailed in the following points.

- Skill degradation (or deskilling) is the loss of human skills that are not sufficiently trained any more but could be desperately needed should the automatic systems fail. In emergency situations Users can even be tempted to trust machine signals more than physical evidence.
- Complex systems tend to be more vulnerable to both failure and attacks. Many complex systems cannot simply be switched off in the case of unexpected events.
- A complex system can even develop chaotic dynamics by itself, without malicious attacks. For instance, it is suspected that power networks can become unstable due to fine-grained steering by smart control units: A slightly decreasing currency price suddenly creates a huge demand, as many washing machines etc. are automatically started at the same time.

### Examples in Computer Science

These examples are more specific to CS. However one should also notice that CS is becoming an increasing part of any science, due to powerful and ubiquitous hardware, big data, etc. Hence these questions are of central importance, not only for our special field.

- One can hold that computer programs must never take final decisions, and this must remain in the hands of humans. Computers can at most be a helpful tool to prepare decisions. (Apparently this was first pointed out by Joseph Weizenbaum in his book "Computer Power and Human Reason" in 1976.) Think, for instance, of support systems for medical decisions. One good reason for not delegating final decisions to machines is that the internal algorithms are based on models that overlook circumstances that a person can still recognize in a specific case. – However, this strict view is not always realistic: In real-time applications no human can react fast enough. Even in "slow" decision processes, programs may take more objective and fair decisions than biased humans. In any case one cannot circumvent the principal question: What are the consequences of delegating decisions to machines?

- Given that final decisions by humans are not always possible (or not even desirable?), apparently a weaker principle is always applicable: It should be transparent whether a decision was made by a system or by a person. It should also be transparent whether a user “talks to” a system or to a person.

- Computers should not completely replace people in fields that require human feelings like empathy (as was also pointed out by Weizenbaum). For instance, when are caretaker robots a good thing to have? They can do some really hard work, but they should not be applied *only* because this makes health care cheaper. (We see again that efficiency is more an instrumental value rather than an intrinsic value.) There is also a risk that developers are fascinated by the technical challenges (e.g., making automatic decisions, modelling emotions) but do not ask enough what the clients really need and want. Another meaningful principle is that caretaker robots should not be used to collect sensitive private data during their work.

- Before an algorithm is applied in the real world, the algorithm has been developed by someone, based on assumptions made by someone, then it has been implemented by someone as part of a program, and someone has decided to use it for just this application. Where in this chain are the responsibilities for the real outcomes, and what are these responsibilities? This question is particularly difficult when systems make autonomous decisions that nobody has explicitly programmed.

- As a tangible example of the previous point: The maneuvers of a driverless car are controlled by algorithms. Who is liable when it causes an accident? Furthermore, already a human driver can get into dilemma situations. (For instance: make way to avoid a collision but then endanger someone else standing there...) It will be a delicate matter to program ethical (!) rules for such situations in the software for driverless cars. Many other points must be taken into account: both driverless and conventional cars are in traffic simultaneously, and human drivers react as well, and they should not get confused by the autonomous maneuvers; software can be vulnerable to spoofing attacks; programs can be unexpectedly slow in critical situations, etc.

- Still related to the previous point: Complex ethical decisions are made by humans on a case-to-case basis and intuitively. The new twist in applications like driverless cars is that *general* rules for such decisions must be programmed *beforehand.* It is hard to imagine that this will be up to every single developer or programmer. One can expect a discussion in society, to reach some consensus about ethically acceptable solutions. But what computer scientists, together with philosophers, can contribute to this process is to recognize and analyze the dilemma situations and the options, and formulate clear questions requiring yes/no answers.

- Users may trust the results of complex programs without further scrutinizing them, forgetting that they rely on model assumptions, let alone possible bugs. Especially Machine Learning algorithms crucially depend on their built-in model assumptions, called their "inductive bias". Once things have been quantified, they "look more objective", which can be an illusion. One can easily give small examples of pretty standard algorithmic problems (e.g., facility location problems) where different optimization goals, each of them appearing plausible as such, lead to different solutions that disfavour some agents in various ways.

- Compilers must be correct, this is an absolute strict demand. But they "only" translate programming languages into other programming languages, which is a well-defined and manageable task. Natural language translation is intrinsically more complicated, and the question is: Where is the responsibility for possible consequences of wrong or misleading automatic translations?

- Good structure and transparency of algorithms and programs is not only a technical matter, but it also supports maintenance and updates. However, the situation is different for some machine learning algorithms: For principal reasons it is not transparent how they generate their results. (If this were transparent and we knew the decision criteria, we would not need a learning algorithm but could directly implement these criteria instead.)

- Programs that easily provide solutions to complex design tasks may reduce the application of human creativity and instead favour standard solutions. This can be an issue in, for instance, software packages that help architects.

- CS provides powerful tools for market research. Both vendors and customers can benefit from the results, but also ethical issues come up. One example is stereotyping: Algorithms group people according to the values of some variables. Thus individuals may end up in groups where they actually are untypical members, nevertheless they are treated as members of those groups. Technology can support discrimination, for instance, selling products to certain customers at worse conditions, total exclusion of certain customers, and several unethical pricing policies. And note that "products" also include things like health insurance ...

- Information technology enables surveillance and massive violation of privacy. Is privacy a value at all? One could argue: "Surveillance only makes society safer, and if I have nothing to hide, so why bother?" However, what can companies, authorities, and others do with all the personal data? Furthermore, already the consciousness that one *could* possibly be watched may change behaviour, for instance, people become more cautious and conformistic – this is known as the "chilling effect".

- Internet users also enter many personal data on purpose, for instance, in order to register for services. In other applications they allow the collection of detailed health data, and so on. They may not realize that these data could be sold, combined with other data, analyzed and later used in unexpected ways that harm the users. Search terms entered in search engines may be analyzed, too, and possibly lead to totally misleading conclusions. A complex question to discuss is: Is it only the users that are responsible for the selection of data they give away?

- Is it correct to provide software or services that users may easily apply unintendedly to their own disadvantage? Is it enough with warnings? (This question similarly applies to every industrial product.)

- What is an appropriate privacy policy of a social network, regarding the use (storage, dissemination, security, etc.) of sensitive data of participants, data of friends, etc.?

- An example of a cost versus safety trade-off in the field of privacy and data anonymization: Prior to the release of a data set for analysis one can remove unique identifiers from the records containing personal data. But still combinations of values may allow unique identifications, with some computational efforts. The more information is removed, the better is the protection of privacy, but the more useful aggregate information for analysis purposes may be lost, too.

- Issues like surveillance, violation of privacy, dependency, skill degradation, loss of control, etc., come to a head by developments like the Internet of Things and the Smart Home. Similarly, robots may get autonomy to such an extent that they may follow own goals, with the potential of fatal consequences. (See also the visions in the works of Nick Bostrom.)

- Already in ancient times the technology of writing(!) has been criticized, claiming that written information lowers peoples' abilities to keep information in memory. (This is an early example of skill degradation.) Maybe this concern was exaggerated, but nowadays the problem appears on a larger scale: Information technology can produce information overload. More information does not automatically imply more knowledge and wisdom. It becomes harder to filter relevant information, to reflect upon it, and to put it in context. One can loose orientation even in very limited domains of knowledge. On the other hand, CS provides many tools for extracting, indexing, and compressing information.

- Personalized news is great, because an individual gets to see what (s)he is really interested in. But this also narrows one's scope if one is *only* exposed to preselected information.

- Search engine results and electronic maps can ruin small companies if they are not displayed and therefore remain invisible for many potential customers.

- What are the pros and cons of open software? (Here one should not forget the economical aspects like investments and payments.)

- Perhaps more than in any other technology, information technology is prone to "vendor lock-in" where customers are completely dependent on products from a single supplier and can change systems only completely and at high costs.

- Some human IT work can be split through the internet into many small and simple tasks (microtasks, microwork), done by many individuals at their own computers. This can be a great way to accomplish big tasks that cannot be done algorithmically, and to create jobs. But the downside is (possibly) low payment and poor social conditions. Moreover, workers are isolated, they can hardly fight for their interests, the microtasks can be stupid work with little room for personal identification with the work and own development. In extreme cases workers may not even know the complete task they contribute to.

- Huge public buildings like shopping malls and major railway stations have to have evacuation plans. Because evacuation can hardly be practiced (not at all, or only rarely), also computer simulations of emergency situations are applied sometimes. This is still quite an effort, but it would be irresponsible to save these costs, arguing that "a catastrophy will hardly ever happen". Simulations can spot weaknesses of the emergency plan and lead to changes of the plan or even of the building. One of the main ethical issues is now the trade-off between the complexity (and cost) of the simulations and their reliability. At first glance, evacuation could be modelled as a variant of a network flow problem. (But already here one must set an ethically correct optimization goal, such as: Minimize the time to get *all* people out of the building.) Then, in the event of an alarm, instructions would be given such that people move to the exits according to the precomputed optimal solution. But this would be a very naive approach. Among other issues, such an approach disregards the patterns of human behaviour and crowd dynamics: People do not behave rationally, especially in exceptional situations, they may run to the nearest exit even if they were told to run to a farther but less crowded exit; groups (friends, families etx.) stay together; people follow the "herd" or some ad-hoc leaders. On top of that, some exits may be blocked unexpectedly. A naive model can still serve as a benchmark to compute the theoretically possible evacuation time, but then a serious model for simulations should take as much as possible of the reality into account, as well as eventualities. A special point is that we cannot know how many people will be in the building. There may be unusually many, and the simulation results may suggest more expensive measures (rebuilding), in order to be prepared for this case. Shall one go for a cheap solution assuming an average number of people, or a costly but safer solution based on an unlikely assumption?

- Predictive analysis of crime with CS methods may be able to actually prevent crime. But is it right to precautionarily suspect people that have never committed a crime (and probably will never), based on statistical data and predictions made by algorithms? What are the priciples of a correct use of such methods?

### Explicitly Ethical Computer Science?

We emphaised earlier that ethics is not technophbic. From general principles such as increasing social benefit one can even derive an imperative to develop technology, and this is part of the specific role of a scientist or engineer. Moreover, CS contributes directly to ethical goals in many non-trivial ways. Here is a list of a few, arbitrarily picked examples.

- Despite several disadvantages, automation in general increases safety.
- Optimization as such leads to saving of resources of all kinds.
- Methods for information extraction, compression, and retrieval save time for manual search.
- Data mining methods can detect fraud via untypical patterns in data sets, which would escape the notice of human observers because these patters are deeply hidden in the data. Software can detect plagiarism of text or music, provided that it is used in an informed way and false positives are erased manually.
- Formal proof methods verify the correctness of systems, which can be safety-critical.
- Predictive analytics using big data can recognize early signs of an epidemic outbreak.
- Privacy policies (for social networks, etc.) can be formalized on a fine-grained level.
- Methods for data anonymization are algorithmically challenging.
- The field of *mechanism design* studies, in an algorithmic style, how a fair allocation of resources can be achieved despite built-in selfish behaviour of the involved agents. (For those who have a deeper interest, good keywords to start further studies are: congestion games, price of anarchy.)
- Optimal trade-offs between conflicting goals can be quantified.

**Recommended Further Reading**

`www.nickbostrom.com` has links to selected papers like
"The Ethics of Artificial Intelligence"
"Ethical Issues In Advanced Artificial Intelligence"

F. Kraemer, K. van Overveld, M. Peterson:
Is There an Ethics of Algorithms?
Ethics Inf. Technol. 13, 251–260 (2011)
(open access article on springerlink.com)

Here the authors argue that many algorithms in real applications make value judgments, and the underlying assumptions should be transparent. An interesting approach is to leave the choice of those parameters that involve ethical decisions to the users. They give an example from medical image analysis, where the ethical dilemma is in the trade-off between false positives and false negatives. The sheer complexity of classification of borderline cases makes it difficult even to give criteria for good program design choices.

Remarkably, in 2015 there has been a conference about the general topic:
`https://cihr.eu/the-ethics-of-algorithms/`

Slides about "Responsible innovation and value sensitive design" by Ibo van de Poel, TU Delft (Netherlands) are available on the web.

Jaron Lanier: "Who Owns the Future?", Simon and Schuster (2013), and related material on the web (interviews, magazine articles)

A big source of material (for those who have time and a deeper interest) is a whole course on Professional Ethics held by Gordana Dodig-Crnkovic:
`http://www.idt.mdh.se/kurser/cd5590/`

## Research and Publication Ethics

Some issues have been already discussed in the context of scientific writing, but there are more. Note in following that “research” does not only refer to academic research, but most points apply to research in industry, too.

- First and foremost, cheating in research, in whatever form, is unacceptable. Cheating hurts science in many ways: wrong results are produced and other researchers build on them in good faith, the wrong people get reputation and funding, quality decreases, general trust in science is damaged, and so on.

- Peer reviewing is the process of evaluating scientific manuscripts for possible publication, or evaluating project proposals for possible funding. It is done by anonymous peers (other scientists in the same field). This is the main tool for ensuring quality of research. The idea is that only experts can give informed and motivated judgements. But peer reviewing can also have unwanted effects: several kinds od bias, conflict of interest, lack of competence, sloppiness, unfairness.

- Who should be the authors of an article? Widely accepted rules are: Authors should be all persons that have substantially contributed to the work. All authors must have approved the publication, and all authors are responsible for the whole content. Persons that had only minor or auxiliary roles should be mentioned in an acknowledgement section. Authorship for any other than scientific reasons is not appropriate.

- Publication is a business and a “currency” for scientists. The careers of scientists crucially depend on their publication records. This pressure can lead to various temptations and forms of misconduct, in the worst case even criminal activities.

- A publication may not be really objective: A subject, a solution, a method or a conclusion may be pushed for other than scientific reasons. Authors may deliberately promise too much (“overselling”), perhaps in order to get attention or attract more funding.

- Research must be free, in order to facilitate open discussion and production of new ideas. Researchers themselves are the most competent persons to decide what research subjects are worth considering. On the other hand, research is costly, and society has an interest in an effective usage of these investments. Some research topics are clearly more urgent than others because of societal needs. What is the right extent of external steering of research? What is a good balance between basic and applied research? Freedom of research also entails some responsibility of researchers to find the most promising research subjects that will be important in the future. – One should be aware that the greatest technological breakthroughs are based on fundamental results in the natural sciences rather than on activities that were genuinely applied research from the beginning. The most striking example is quantum physics. It is the basis of virtually all modern technology but started as an endeavor to understand the inner structure of matter, without having any applications in mind. A myopic way of supporting only immediately useful research would miss such great opportunities.

## Fallacies of Statistics

Wrong use of statistics means wrong "knowledge" and wrong conclusions which can even be dangerous, especially in medicine and justice. Usually it does not origin from evil motives. Apparently our brains are not well suited for working with probabilities, and even mathematically educated professionals may sometimes have a poor understanding of probability.

Computer scientists should be aware of the most common fallacies of statistics, for several reasons: Computer programs for data analysis and inference are based on statistical reasoning after all. Statistical methods that we implement should at least be proper methods. Mathematicians, computer scientists, etc., are also those people who could be consultants for others when it comes to statistical inference. Furthermore, many studies of the quality and performance of IT products include user surveys that should be properly performed.

### Interpretation Mistakes

We list some common mistakes, some of them being instantiations of the same underlying misconception. These notes shall not replace a course in statistics, they are merely reminders of possible statistics issues in scientific writing.

- To start with a very elementary but widespread mistake: The median (the value in the middle) and the mean (average, expectation) are often confused or wrongly communicated. Note that the median is a "typical" value, whereas the mean can be totally different when the distribution of values is skewed.

- Pretended accuracy: If a study claims that 71.43% of persons were satisfied with some product, one should be suspicious: This number is very close to 5/7. Looking into the study one may find that in fact only 7 people have been asked! Besides the numbers of cases one must give a confidence interval, and for a very small sample this interval is broad, such that no reliable conclusion can be drawn at all, even if the sample was representative and homogeneous.

- A hypothesis is accepted based on a small sample that gives no evidence against the hypothesis, thereby forgetting that the sample is also too small to support the hypothesis. In the simplest case of this mistake, a general claim is made, such as "all objects $O$ have property $P$", and no counterexample is found among the few considered examples. But this is not enough to conclude that the hypothesis is true. (By the way, this mistake is also made if one believes that some algorithm is correct because it worked well on a few test instances, and believes that a laborious correctness proof is not needed ...)

- Data are oversimplified. Records consist of only a few key numbers, ignoring the complexity and multidimensionality of the real-world subject, nevertheless they form the basis of an analysis.

- Conclusions are oversimplified. For instance, we may find that the expected loss (of property or even lives) in a certain type of disaster is rather low, and conclude that this is an acceptable low risk. But an expected value is only an average! It could be the product of a very small probability of the event and an unbearable high loss *if* the event occurs. Shouldn't one invest in prevention, although the probability of disaster is indeed very low?

- Persons are grouped according to some known data, and predictions $P$ are made for the memberso of each group (This $P$ could be: will not be able to pay back a loan, will commit a crime within the next year, etc.). Suppose that the figures are correct, data are meaningful, and the prediction is based on serious theories. Still it would be problematic to say: "A person $X$ from the grpup $G$ will do $P$ with a probability $q$." (Plug in a number for $q$.) This would do the individual $X$ unjustice.

The method allows at most to conclude: “If we randomly pick a person from group $G$, this person will do $P$ with probability $q$.” Note the difference between these two statements. Interpretation can be crucial if further conclusions are drawn about actions and precautions.

- This is another misconception regarding low probabilities: An event may have a low probability for each individual, but in a large enough population it will happen to *some* individuals, for no special reason. (Think of a lottery.) From low probability alone we cannot conclude that “this cannot have happened by chance, there must be a reason behind it”. In fact, this type of fallacy has led to very bad cases of miscarriage of justice.

- Cherry picking: Unfavorable data are omitted from a single data set, a whole study with undesired results remains unpublished, etc. Here is a toy example: 10,000 people are asked to predict the results of 10 coin tosses. The probability to be always right is about 0.001. That means that still about 10 people are always right. We publish only their results and declare them prophets. We also explain that, “of course, prophecy is rare, but these few people have clearly demonstrated their prophetic skills”. – Here the nonsense is obvious, but there are more subtle scenarios prone to this type of error. An interesting proposal to avoid suppression of unwanted results is that the publication of *all* statistical studies, e.g., in medicine, should be mandatory, and the scientific quality of studies should be evaluated based on their methodology rather than on their positive findings.

- A related fallacy is data dredging, also known as HARK (“hypothesis after knowing results”): Patterns are seen in data, but without a previously defined hypothesis to be tested. This is of no value because any large enough amount of data will exhibit some spurious patterns by chance. (Some patterns even appear with necessity, for combinatorial reasons. For instance, Ramsey’s Theorem says that for every $k$, every large enough graph contains a clique or co-clique of size $k$.) If many combinations of parameters are considered on the same data, one will always find surprising correlations between some. (In a pseudo-study deliberatly made to show the issue, it was “demonstrated” that consumption of chocolate supports a diet for weight loss.) Any such statistical finding must be verified on a second, independently chosen data set. However, just deriving a *hypothesis* from data is a correct procedure.

- A statistical result is carried over to another or a more general population the sample is not representative of, or the sample is biased towards a subgroup. (Figuratively speaking, by catching fish with a net with meshes of a certain width one will only find fish of at least that size, and one should not conclude that all fish species have this minimum size.) A sample can be biased because gathering data can be more costly for some subgroups, or because people can decide by themselves whether they want to participate in a study.

- Correlations are taken for causal dependencies without thinking. Two events may show correlation fpr other reasons, for instance, because both are caused by a common third factor. Although this may be widely known, in complex big data analysis scenarios it is easy to forget this fallacy. On the positive side, such relationships can be accurately modelled and analyzed by probabilistic networks: graphical models, Bayesian networks, etc., which is an important subject in Machine Learning.

- In medical studies, false causal relations can be ruled out by comparing the effect of a treatment with a control group. Then the statistical fallacy is avoided, but this can raise an ethical dilemma instead: Can we give people in a control group treatment, or prevent them from necessary treatment, knowing that this can harm those people, but also knowing that the scientific results will be important for many more other people? What should be the ethical values and principles steering such decisions?

- A more subtle phenomenon is Simpson's paradox: Conclusions from data may depend on the grouping of individuals! There exist many variations of the paradox, for instance: The same trend appears within each group, but the whole population exhibits the opposite trend. A method X performs better than a method Y in each group, but in the whole population the result is the other way round. (Small numerical examples can demonstrate the sources of this, at first glance, incredible effect.) Now it can be tempting to choose a partitioning into groups that "supports" a desired result. Also, the analyst can be honest but unaware of the fallacy. Obvious questions arise: What is the correct grouping of a population? What is the correct interpretation? Here we cannot discuss the matter in depth; the point was mainly to draw attention to this non-obvious paradox.

**Made-up example of Simpson's paradox**: Two methods M1 and M2 for data analyis, prediction or whatever, are applied to two small datasets D1 and D2. We record how often each method was successful on each dataset, according to some criterion. The table shows the results. Every table entry means "successful applications / total number of applications". For each of the datasets, M2 has the higher success rate, but for both datasets together, surprisingly, M1 performs better. Now start thinking: How can that be?

| | $D1$ | $D2$ | $D1 + D2$ |
|---|---|---|---|
| $M1$ | 0/1 | 3/4 | 3/5 |
| $M2$ | 1/3 | 1/1 | 2/4 |

## Misleading Graphs

We can gather much information instantly through our eyes. In the era of big data the importance of data visualization has only increased. Visualization is a nontrivial field in its own right, including psychological aspects but also algorithmic questions. (How can we draw a graph in the best way, according to some optimality criteria?)

But graphics can also be misleading, either on purpose or unintendedly, even if they are based on true data. A few "methods" to construct misleading graphs are:

- Comparable quantities get different scaling, to make some of them appear larger than they are. Distorted perspective or additional dimensions can have similar effects.

- An axis starts at some positive value rather than at zero, thus differences and changes appear larger than they are. Truncation is justified when small differences need to be displayed, but then the truncation should be clearly pointed out. In scatterplots, truncation can also cut away outliers without clearly stating this fact.