DAT315. 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 deals with right and wrong conduct, that is, it revolves around the philosophical question "What should I do?" (one of Kant's Four Questions). There are some differences between ethics and morality. The verdict "immoral" feels stronger than "unethical", and ethics refers more to the right conduct in a certain role or context.

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 how 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. The name is a bit misleading, as "applied ethics" does not just straightforwardly apply ethical theories to specific cases. Rather, theories are only used as a guideline 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 competence for a serious analysis; this task cannot simply be delegated to politics and administration.

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 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 the decisions being on the agenda. One also gets aware of the own value system. It is impossible to have no ethical principles at all (in the same sense as it is impossible not to communicate). They only remain below the level of consciousness if one does not reflect upon them. Last but not least, 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 they can even have economical effects.

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 a 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 is 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 the synonym **value** 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, caring about future generations, etc.

A pragmatic justification of these general ethical values 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. Still it is not easy to deduce a principle like "treat all people equally" from that. Anyway, this approach justifies values by the negative consequences of not following them; this position is called negative consequentialism. Similarly, we may "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.

Two types of values are often distinguished: 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 attribute: a value can be *instrumental for* another one ("... for the sake of ..."). The instrumental value is then a lower value in the partial order. Here are a few examples.

Knowing science can be considered an intrinsic value (understanding the world, curiosity, intellectual pleasure) but is obviously also an instrumental value. *Transparency* seems to be an instrumental value only. It depends on the context whether it is desirable or not, feasible or not: Compare transparency of a program, algorithm or system with transparency of records with private data.

A More Detailed Example

Efficiency of products (being small, fast, user-friendly, have low energy consumption, etc.) is certainly a value to aim for. And *more* efficiency improves the usability and decreases running costs and consumption of energy. (Let us focus on the last aspect, 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 *could* be unwanted consequences that are not obvious and deserve a deep analysis, and efficiency is therefore more of an instrumental value rather than an intrinsic value.

An innovation that improves efficiency prompts many customers to buy the new version of the product instantly. Isn't that good? They will save energy from now on. But before that, the production of the new devices *costs much 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 the initial extra costs pay 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 may 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 I wait, then my running costs will (unnecessarily!) be higher, until I finally buy the newer version, therefore I better buy now. 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, "online" has nothing to do with the internet. The term 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 imply that *people* intuitively apply such strategies. One can assume that the supplier has more competence also in the economical aspects, and hence the supplier has 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 can both prevent many users from too many purchases and still increase the profit, since many customers will still buy.

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", the design of products that deliberately limits their lifetimes although they could physically be working for longer periods. Such a strategy 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 (suppliers, 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, but we cannot provide final answers to them. Just think about them.

- 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 consequences ("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 intended ones. Nobody can be blamed for a bad outcome that was really unpredictable, even if this person's action provably generated this outcome in a causal chain of events. (Note that a deontological position is applied here.) But how big is the obligation to investigate all possible (likely) 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: 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 is a cut-off point – determined by what?

- Ethical responsibility is, in general, larger than just accountability. As an example: The person that runs the system is responsible for avoiding possible harm caused by that system. But already the designer of the system should make predictable, potentially harmful situations impossible.

- 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 be considered in decision making? Certainly this meta-decision also involves some ethical values.

Formalize 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 autonomuous AI systems the right behaviour in a wealth of situations? Complete formalization is clearly far too ambitious, but some more modest goals are realistic, e.g., 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 a robot ethics (see, e.g., the conference series WeRobot).

Attempts to formalize limited parts of ethics and studying their logical implications also helps reveal inconsistencies and gaps in ethical theories.

Aggregating Utilities?

This matter is discussed in some ethics literature. It has an optimization flavour, so it should also be fun (without forgetting the serious background) for CS people. It shows that the utilitarianism principle of increasing happiness for a maximum number of people is not as clear as it may appear at first glance.

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 predict their "happiness". This assumption is, of course, already an idealization, but it can be a reasonable approximation of truth in many cases. For every action we get a vector of n utilities, where n is the number of individuals. Now, what action is the best? This is quite clear if some utility vector *dominates* all others, i.e., is component-wise the largest. Then we should choose this one.

Otherwise, one standard approach would be to maximize the sum (or equivalently: the 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. A utility monster cannot "represent" the average happiness of a population.

An obvious idea to avoid such disfavour is to maximize the minimum of all n utilities instead. 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 in skewed cases. But also the increase of the median can disfavour many other individuals.

A natural special case of our abstract setting appears when fixed total amounts 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 in reality seem to be often concave. 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. (The mathematical details are simple and are omitted here; think about them.)

The question what 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.

One could object that the whole discussion is an artifact that origins from an attempt to compare incomparable objects, that is, to turn a partial order into a total order. But the point is that decisions of this type do appear in reality, no matter whether we like this fact 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 the 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 is merely a function of these n numbers, is an abstraction and simplification. It 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: Also ask yourself what has been "abstracted away". Consider models only as a guide, do not automatically follow the optimization results just because they were output by a program, but check them against reality.

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 considered to be instrumental values, and their moral rights are respected in the end.

Examples of Ethical Issues in (Computer) Science

Nowadays ethical questions in CS are even a frequent topic in the media. Also, 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, they are mainly intended as inspiration. Suggestions of further examples are welcome, too.

Pick some of the issues and start thinking: How would you approach the questions? Can you state the ethical problems more explicitly and detailed? What options do you see to solve these problems? What is your favourite solution? And which factors, general principles, and ethical standards have guided your reasoning? Would you revise your opinions after re-thinking from a more general perspective?

Perhaps 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 task would be manageable and the benefits for many people would be obvious?

- If some resources are scarce, where do we put them, what 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; anonymity and privacy of data.

- Risk versus cost: There can be a trade-off between costs and reliability of a product or service. 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 might have detected (for instance, prediction of earthquakes)? One of the subtle points is: How can one communicate the intrinsic uncertainty of the prediction itself, in a way that people without a clue of science can understand? 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 that. 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, and complex systems cannot simply be switched off.

- 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 becomes 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; 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 is that the internal algorithms are based on models that overlook circumstances that a person can still recognize in the particular case. – However, this strict view is not always realistic: In real-time applications no human can possibly be fast enough. Even in "slow" decision processes, programs may take more objective and fair decisions than biased humans. In any case one cannot bypass 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 should not be applied *only* because this makes health care cheaper. Here 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 principle worth considering 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? The 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 as well: 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 do in 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 absolutely 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 necessary changes to eliminate unwanted side effects. However, the situation is different for some Machine Learning algorithms: For principal reasons it is not transparent how they achieve their results. If this were transparent and we knew the decision criteria, we would not need a learning algorithm at all, and we could directly implement these criteria.

- Programs that easily provide solutions to complex design tasks may reduce the application of human creativity and instead favour uninspired 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 and bank loans ...

- 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 should I bother?" However, what can companies, authorities, and others do with all my personal data? Furthermore, already the consciousness that one *could* possibly be watched may change behaviour, for instance, people become more cautious and conformistic (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 ways that harm the users. Search terms entered in search engines may be analyzed, too, and possibly lead to wrong conclusions. Is it only the users that are responsible for the selection of data they give away? Moreover, data of friends allow to draw conclusions about persons who are careful with their own data.

- 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.)

- Nudging is a way to get other people (users of a system, etc.) to take decisions that the designer of the system wants. They have the full freedom of choice, but the desired options are made easier, for instance, due to the way the options are presented. When is nudging ethically correct?

- 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-safety trade-off in 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 then 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 it 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 identifica-

tion with the work and own development, and workers may not even know what the complete task is.

- 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 saying 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 solution. 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 etc.) stay together; people follow the "herd" or some ad-hoc leaders. On top of that, some exits may be blocked unexpectedly. A naive optimization problem 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) 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?

Explicitly Ethical Computer Science?

Ethics of technology should not be misunderstood as 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.

On the positive side, 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. This 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 since 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 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/`