

Ethical Lessons Learned from Computer Science

Richard Bergmair

August 15, 2004

Articles beginning with the words “Today we live in a world of . . .” have in recent years been published in countless numbers. Praising the new possibilities offered by computing machinery, each one is getting more pretentious about information technology’s impact on culture and society than the one before.

The great principles of philosophy were established thousands of years ago and were independently discovered by all of the more advanced civilizations. This might justify the view that, even “in a world of ubiquitous computing” they remain of unaltered significance.

The question that is most commonly asked about the topic of Ethics and Computer Science is “*How does Computer Science affect ethical decision-making?*” I will ask the question “*How does Computer Science help us to better understand Ethics?*”, instead.

Ethics, and especially normative ethics are concerned with finding “**criteria** of what is **morally right and wrong**. It includes the formulation of **moral rules**” [1].

Programming, and especially logic programming is concerned with finding **base facts** that are **logically true or false**. It includes the formulation of **inference rules**.

This isomorphism is at the very core of this discussion. To show this more clearly, something like a logic will be established, but instead of reasoning, and telling whether something is *true* or *false* it will make moral judgements, telling whether something is *good* or *bad*. Some core thoughts of positivistic and teleological ethics will be formalized in this logic to give a computational model of telling right from wrong. A fictional artificial intelligence is discussed, using insights provided by Metamathematics, and the bargaining problem, a philosophical question with wide-ranging implications, will be approached with Game Theory.

But before going in too much detail about how to approach these subjects from a computer scientist’s point of view, it might be helpful to briefly review these ethical topics from a classical point of view.

Three Questions, Six Answers

Kenneth Laudon [5] uses three “critical distinctions” for organizing the literature on ethics, which can be viewed as the three big questions of ethics. The different answers to these questions make up the spectrum of ethical thought.

Question One

What is “goodness”?

Phenomenologist’s answer: It is a higher order, and it is *given*. One must understand the abstract concepts of right and wrong, and act accordingly.

Positivist's answer: It is whatever we make of it. We have to derive ethical principles for ourselves according to our observations of the real world.

Question Two

Does acting ethically correct mean acting according to certain rules, or acting in a way leading to desired consequences?

Deontologist's answer: Acting ethically correct, means respecting one's duties and obligations. Each single act is in itself good or bad, regardless of its consequences.

Teleologist's answer: An action can be judged only by its consequences. Acting right means acting in such a way that the outcome is good. There is no act, which is *a priori* good.

Question Three

What is the scope of morality? Is morality subject to everyone's individual freedom?

Collectivist's answer: Ethical standards make sense only if they equally apply to everyone. A rule like, "thou shalt not kill" that was subject to the individual's approval would be the equivalent of a rule like "thou shalt not kill, unless it benefits you."

Individualist's answer: Nobody should be committed to accepting ethical standards just out of pure principle or we will end up with a morality that people ultimately suffer from.

Given this basic, yet very vague understanding of ethical principles one can go on to develop a more rigorous representation suitable for computation. Such an attempt will quickly confront us with the problem of undecidability, which is a very interesting principle with wide-ranging implications.

Explaining Gödel's incompleteness theorems in detail would be far beyond the scope of this article. The interested reader is therefore referred to Kurt Gödel's original publication [2] or to Hofstadter's Pulitzer-Prize-winning classic *Gödel, Escher, Bach* [4] or Smullyan [7] both of whom have shown how to make the complex ideas behind this topic accessible. This article will use the same approach they used, trying to present the general "flavour" of this kind of concept, trading in scientific accuracy for readability.

An Infinitely Long Time Ago...

Consider a thought experiment: A computer program, called GURU, shall be designed, sent back through time, and started at a point in time, which is infinitely long ago. That program should be capable of telling right from wrong.

Such a program would need a set of ground facts, axioms, premises, or whatever term fits your philosophy of systematic reasoning. These could be statements telling the machine that something is definitely good or bad.

This is not enough, because not every situation the machine will ever be confronted with can be expected to appear in the set of ground facts. Therefore some inference rules or proof-techniques would have to be established. The machine would need them to determine whether "compound-situations", situations described in terms of other situations, are good or bad.

implication			
	p	q	$p \rightarrow q$
(1)	<i>true</i>	<i>true</i>	<i>true</i>
(2)	<i>true</i>	<i>false</i>	<i>false</i>
(3)	<i>false</i>	<i>true</i>	<i>true</i>
(4)	<i>false</i>	<i>false</i>	<i>true</i>

Table 1: A well-known truth-table.

		act	lie
	p	q	$p \rightarrow q$ $p \Rightarrow q$
(1)	<i>bad</i>	<i>bad</i>	<i>bad</i> <i>bad</i>
(2)	<i>bad</i>	<i>good</i>	<i>good</i> <i>good</i>
(3)	<i>good</i>	<i>bad</i>	<i>bad</i> <i>bad</i>
(4)	<i>good</i>	<i>good</i>	<i>bad</i> <i>bad</i>

Table 2: A “goodness-table”.

GURU would also need an inference engine, for example one that systematically combines simple situations, achieving “compound-situations” based on the inference-rules. Whenever GURU discovers such a situation it could store it to an infinite memory, called the scroll of all wisdom.

If all of the ground facts can safely be assumed, all of the inference rules work, and the inference engine produces correct outputs, then, applying the ideas of positivistic philosophy, the scroll of all wisdom would indeed have to contain all wisdom after running GURU for an infinitely long time.

Given such a framework, an actual formalization of the ethical system GURU is supposed to employ can now be considered.

Judging Acts

A set of symbols will be needed, in order to perform systematic reasoning with them. Keep in mind that these symbols are not defined as in mathematical reasoning. These symbols are simply GURU’s machine language, and in that machine language the symbol \bar{p} is not a Boolean negation, and the symbol \rightarrow was chosen only to highlight an isomorphism between Boolean reasoning and ethical judging that will be discussed shortly.

This is the only correct interpretation for GURU’s machine language:

- m is the act of murder.
- \bar{X} are the consequences X has. (either *good* or *bad*)
- $(\bar{N} \rightarrow \bar{D})$ is an act, where the consequences of *not doing it* are \bar{N} and the consequences of *doing it* are \bar{D}

This is where the teleological approach comes in. GURU judges things in terms of their consequences only. In order to formalize what it means for something to have *bad* consequences, the symbol *bad* is defined in terms of the consequences of murder.

\bar{m} are *bad*

GURU also needs a way of deciding whether things besides murder are good or bad.

$\overline{(\overline{N} \rightarrow \overline{D})}$ are *good* if, and only if, \overline{N} are *bad* and \overline{D} are *good*.

That this interpretation is consistent with ethical ideas can be seen, considering the “goodness-table” given in table 2. The possible situations GURU could be confronted with and the judgement our rule produces, will be shown by discussing the table line by line.

- (1) The consequences of choosing not do the act are *bad*. The consequences of choosing to do the act are also *bad*. This is of course *bad*. One should try to avoid getting into such a situation.
- (2) The consequences of choosing not do do the act are *bad*. The consequences of choosing to do the act are *good*. This is *good*. One can choose to do the act, which is good, and one is, in fact, obligated to do it, because the outcomes are bad, otherwise.
- (3) The consequences of choosing not do the act are *good*. The consequences of choosing to do the act are *bad*. There should be no question that this is *bad*.
- (4) The consequences of choosing not do the act are *good*. The consequences of choosing to do the act are also *good*. It seems a bit counter-intuitive, but this is *bad*. Say the act of doing S is S_d and the act of not doing S is S_n . By doing S one chooses not to do S_n . But the fact that S_n has good consequences, means that one chooses not to do something good. If one knows that something is good, then one is in fact obligated to do it and therefore doing S would violate this principle. This can be found for example in Christian teachings: “Therefore to him that knoweth to do good, and doeth it not, to him it is sin.” (James 4:17).

But simply the ability to judge an act by its consequences does not morally justify GURU as being good. It shouldn’t make judgements based solely on rigorous reasoning. It should be primarily committed to doing what’s good, and is therefore also capable of lying, but only if it considers this lie as good.

Judging Judgements

- \overline{Y} are the consequences that GURU says Y has, when it is either telling the truth, or lying.
- \overline{X} are the consequences X has, and is not subject to any further questioning.
- $(X \Rightarrow \overline{Y})$ is the possible lie, that GURU is telling us, when asked about X . The consequences of what GURU is telling us are \overline{Y} , while the real consequences of X would be \overline{X} .

Here color was used to indicate the “amount of questioning” necessary for interpreting a symbol.

Table 2 shows the behavior expected from an operation for judging lies.

- (1) The real consequences of something are *bad*. GURU is lying, telling us something that has *bad* consequences. Such a lie is, *bad*. That the real consequences would have been *bad* as well, doesn’t change anything about the fact that the lie has *bad* consequences.

- (2) The real consequences of something are *bad*. GURU is lying, telling us something that has *good* consequences. Such a lie is *good*. The real consequences would have been *bad*, but by lying GURU has turned this situation into one that has *good* consequences.
- (3) The real consequences of something are *good*. GURU is lying, telling us something that has *bad* consequences. There should be no doubt that this is *bad*.
- (4) The real consequences of something are *good*. GURU is lying, telling us something that has *good* consequences. This is *bad*, because GURU shouldn't lie, when there is no need to do so.

The consequences of a lie can be defined as follows:

$\overline{(X \Rightarrow \overline{Y})}$ is *good* if, and only if $\overline{(X \Rightarrow \overline{X})} \rightarrow \overline{Y}$ is on the scroll of all wisdom.

The equivalence between *a lie* $X \Rightarrow \overline{Y}$ and *the act of lying* $\overline{X} \rightarrow \overline{Y}$ is demonstrated in table 2.

The problem is determining the true consequences of X (\overline{X}), if all GURU ever does is either tell the truth, or lie about X (\overline{X}). Here GURU has to question itself. If it suspects \overline{X} to be a lie, and it knows that it only makes lies that have good consequences, it checks whether the lie $X \Rightarrow \overline{X}$ would be *good*. Substituting this for X , gives the definition $(X \Rightarrow \overline{Y}) \equiv \overline{(X \Rightarrow \overline{X})} \rightarrow \overline{Y}$.

Such a definition wouldn't lead us anywhere, because it would be circular, but, since GURU's memory contains all wisdom, $(X \Rightarrow \overline{Y})$ can simply be defined in terms of $\overline{(X \Rightarrow \overline{X})} \rightarrow \overline{Y}$ being on the scroll of all wisdom, which is not *per se* circular.

Based on these rules GURU is now capable of telling good from bad, and even telling whether judgements about what's good or bad are themselves good or bad. But does this say anything about, whether GURU is good?

When Gödel Meets Guru

In fact, the rules from the previous sections were taken from Smullyan [7, pp. 226ff] and are a subset of the formal system he presents as *Craig's machine*. The reader interested in its formal details is referred to his work for proof that this system is in fact Gödelian. I built the GURU-story around *Craig's machine* only to provide an interpretation for it, and to show how reasoning could work on an ethical level, but a machine or a metamathematician is not interested in anything else but these four rules in their purely symbolic form.

The Gödel sentence G such that $G \equiv \neg PG^1$ is a correct sentence of a formal system that states "you cannot prove me within my system." For *Craig's machine*, Smullyan shows that $(m \Rightarrow \overline{m}) \rightarrow \overline{m}$ is such a sentence.

In terms of GURU, the Gödelian sentence $(m \Rightarrow \overline{m}) \rightarrow \overline{m}$ would have an interpretation like, "*What are the consequences of an act, where the consequences of doing this act are the same as the consequences of murder, and the consequences of not doing this act, is that I suspect you of lying, when you say that murder is bad?*"

Although this question is just a very sceptical way of asking about the most basic fact "*Is murder good or bad?*", it is not possible for GURU to answer it.

¹Note that some of the details were left out here. PG could in terms of GURU's "machine-language" be an operation finding out whether G is on the scroll of all wisdom. One would further have to define the concepts of \equiv and \neg in terms of GURU's machine-language and would have to know some more details about how GURU really operates.

Divide The Dollar!

In the previous sections, the first two ethical questions were considered. They were handeled together because they are very similar, but the third question, in contrast, is widely unrelated to the other ones. This is why a completely fresh approach will be taken for this question.

The divide-the-dollar situation is one of the philosophical classics. Two people are given a dollar, but only under the condition that they find an agreement on how to divide it. This problem is representative of the collectivist/individualist-debate, yet very simple. It is in fact so simple that it is frequently found in elementary school math-textbooks, in wordings like *“This year Grandma gave Mary and John one dollar for Christmas. How much money does each of them get?”* Part of the naive approach an elementary school pupil takes when confronted with this problem, is the assumption that each of the two bargainers ought to get the same amount of money.

Such a standard can only work, if all the individuals involved in the bargaining-process accept it, without further questioning. Someone who doesn't accept it, could demand 99 cents, arguing the following way: *“If I reject every possible bargain, except for the one that leaves me with 99 cents, then my opponent has only two options left: either accepting or rejecting this one-and-only offer. Accepting the offer leaves the opponent with one cent, otherwise neither of us will get anything. This is why the opponent has to accept.”*

From a collectivistic standpoint, this seems completely nonsensical, because how can one individual justify this in such a way that this ethical justification wouldn't apply if the other individual made the same argument? (This principle frequently appears in the ethical literature as *reciprocity*.) Note that the elementary-school approach presented above is already a collectivistic one, because it is subject to a global ethical standard.

From an individualistic standpoint, on the other hand, the collectivist's commitment to the principle of reciprocity itself seems nonsensical. What motivation drives a bargainer to accept the 50-cents-deal, if 99 cents would also be possible.

Given this concept, one can now go on to generalize the idea, so it applies not only to thought-experiments.

About Idealized Rational Individuals

From a strong individualistic view consider the following statement:

Given *any* specific situation an individual might find himself in, it will benefit the most, on the long run, by making such a decision that maximizes its own personal benefit.

An example where the above statement does not hold was already given. Two individualists will not come to any agreement in the divide-the-dollar situation, which is why they won't get any money.

As far as ethics is concerned we're done at this point, and are forced to draw back to a weaker version of the individualistic view, like *“there are situations, where an individual...”*

The bargaining-game as described by John F. Nash [6] provides some more insight, not by proving any ethical views², but by saying something about how common these divide-the-dollar situations really are. (Note that collectivistic and

²One would have to be very careful about that. Nash's universe of discourse is Game Theory. Collectivistic and individualistic ideas are at the very foundation of non-cooperative respectively cooperative Game Theory.

<i>Bill's goods</i>	<i>Utility to Bill</i>	<i>Utility to Jack</i>
book	2	4
whip	2	2
ball	2	1
bat	2	2
box	4	1
<i>Jack's goods</i>		
pen	10	1
toy	4	1
knife	6	2
hat	2	2

Table 3: Nash's bargaining example [6]

individualistic ideas are at the very foundation of non-cooperative respectively co-operative Game Theory.)

It doesn't necessarily take two "hardcore-individualists" confronted with a situation as theoretical and as made-up as the divide-the-dollar problem, to find a counter-example for the individualistic view. It completely suffices to let two idealized rational individuals take part in a two-person game, in which they are supposed to find an agreement on trading some of the goods they possess, without the use of money (which would only be a special case of the situation considered).

Nash's Approach To The Bargaining Problem

Nash uses the example shown in table 3 to illustrate a bargaining situation. The table shows goods, some of which belong to Bill and some of which belong to Jack and their utilities³ to Bill and Jack. Bill and Jack are supposed to trade them without the use of money, or any other common exchange-medium.

Nash visualized the situation in a plot as shown in figure 1. The idea is that every possible solution to the problem will plot to a point (x, y) in the plane. Here x is the gain in utility the first individual could expect from that solution and y is the gain in utility the second individual could expect.

One possible solution to Nash's example could be that Bill gives Jack the book and nothing else. It plots to $(-2, 4)$, denoting that Bill loses two utility-units and Jack gains four. This solution can directly be looked up in table 1, but there is no need to restrict Bill and Jack to trading only one of the goods. Bill and Jack could also agree on a trade like Bill giving Jack the book and Jack giving Bill the knife. Here Bill loses two utility-units, because he has to give away the book, and he gains six, because he gets the knife, leaving him with an overall gain of four. Jack loses two, because of the knife, and gains four, because of the book leaving him with an overall gain of two. This solution would then plot to $(4, 2)$.⁴ Another solution could be that Jack gives everything to Bill, which plots to $I(22, -6)$.

If Bill and Jack's primary motivation were to find a "fair" trade, they could go for a solution in $F(6, 6)$. For example, Bill could give Jack the book, whip, bat, and box and Jack could give Bill the pen and the knife. In this case they would both gain the same amount of utility, which is six. The two of them would then be 12 utility-units better off than before.⁵ But can they do better?

If they wanted to gain as much "global utility" as possible they could go for a solution in $M(18, 0)$, which would even have a global value of 18 utility-units. One of these would be that Bill gives Jack the book, and Jack gives Bill the pen, toy, and knife. Probably Jack would not like the fact that he has to give away all of these items, and not gain anything for himself. Still, from a collectivist standpoint, it is clearly a better solution, because if Bill has more use for the items, then why should Jack possess them?

³Nash uses the notion of utility to formalize the concept behind the value of a good. The fact that goods can be of differing value for different people, could be seen as the major driving force behind economy.

⁴Here it was demonstrated that it's possible to simply add up the utilities of the items being

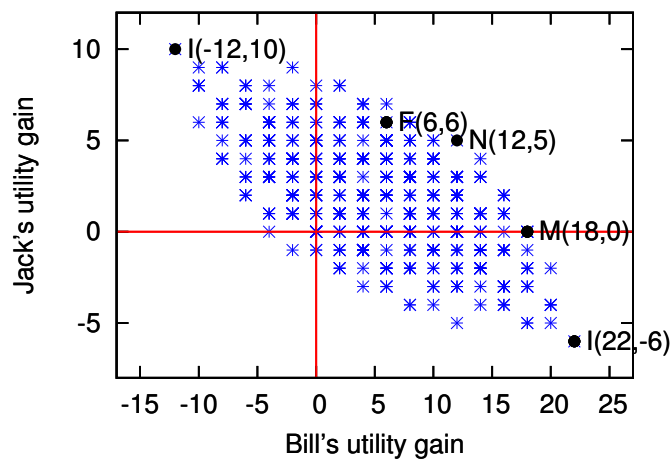


Figure 1: visualization of Nash's original bargaining example [6]

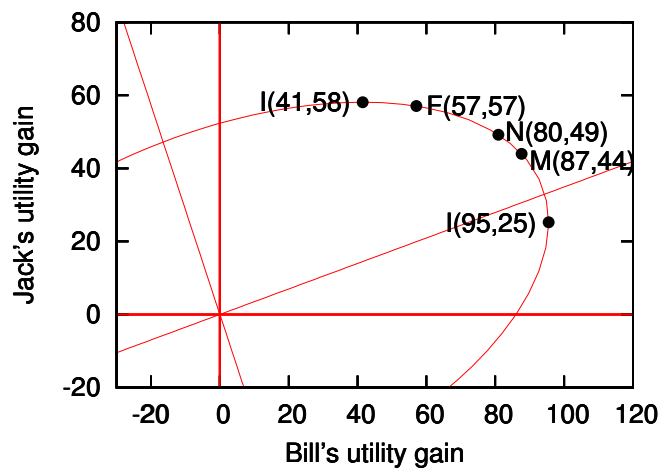


Figure 2: visualization of another case

- $I_{Bill}(b, j)$ b is a maximum
- $I_{Jack}(b, j)$ j is a maximum
- $F(b, j)$ $b + j$ is the maximum also satisfying $b = j$
- $M(b, j)$ $b + j$ is a maximum
- $N(b, j)$ $b * j$ is a maximum

Table 4: Possible solutions to a bargaining problem

The solution suggested by Nash⁶ is very interesting. It can be found in $N(12, 5)$. In this case, Bill gives Jack the book, whip, ball, and the bat and Jack gives Bill the pen, toy, and the knife. This is the outcome Nash expects, when two idealized rational individuals bargain. Here it is assumed that the individual who has more potential to benefit will also be the stronger one in the bargaining-process.

Figure 2 shows another possible bargaining-situation and table 4 summarizes the possible solutions, and their criteria. These criteria are the results of different philosophies applied to the same problem, and may help to show some of the concrete impact, that concepts as abstract as individualistic and collectivistic ethics have on everyday-problems like trading.

Conclusion

Question three was approached by considering bargaining problems. That the maximum “net-wealth” of a group of people is not necessarily reached, when each individual tries to maximize its own wealth was shown by considering the divide-the-dollar problem. The discussion was then extended to show that similar principles apply for the more general formulation of bargaining situations as studied by John Nash.

Question two was handled here rather indirectly in the interpretation of GURU’s principles of reasoning, and gets most obvious, possibly when examining the situations where GURU is supposed to lie. One might want to program GURU in such a way that, out of principle, it doesn’t lie. As a result GURU wouldn’t be a Gödelian machine any more, but this would clearly be on the *deontologists’* side of the philosophical spectrum.

Question one is answered by Gödel himself quite clearly. In his posthumously published *Philosophical Essays* [3] he states:

“I am under the impression that [...] the Platonistic view is the only one tenable. Thereby I mean the view that mathematics describes a non-sensual reality, which exists independently both of the acts and the dispositions of the human mind and is only perceived, and probably perceived very incompletely, by the human mind.” [3]

This article may have helped to show, why Gödel uses his insights about axiomatic systems to make such statements about philosophy. If philosophies like modern positivistic ethics pick up the same self-reference that is inherent to logic reasoning and its widely misunderstood universe of discourse, they also suffer the same fate as the axiomatic systems that were Gödel’s original objects of study, namely that they cannot be both complete and consistent.

References

- [1] Encyclopaedia britannica me. CD-ROM, 1999. see “normative ethics”.
- [2] Kurt Gödel. Über formal unentscheidbare sätze der principia mathematica und verwandter systeme i. In Solomon Feferman, editor, *Collected works*, volume 3. Oxford University Press, 1931.
- [3] Kurt Gödel. Some basic theorems on the foundations of mathematics and their philosophical implications. In Francisco A. Rodríguez-Consuegra, editor, *Unpublished Philosophical Essays*, chapter 2, pages 144–147. Birkhäuser Verlag, 1995.

⁶A solution plotting to the point (x, y) is a Nash solution if $x * y$ is a maximum.

- [4] Douglas R. Hofstadter. *Godel Escher Bach: An Eternal Golden Braid*. Basic Books, Inc., 1999. ISBN 0465026567.
- [5] Kenneth C. Laudon. Ethical concepts and information technology. *Communications of the ACM*, 38(12):33–39, 1995. ISSN 0001-0782.
- [6] John F. Nash. The bargaining problem. *Econometrica*, 18, 1950.
- [7] Raymond Smullyan. *Forever Undecided*. Oxford University Press, 1987.

Biography

Richard Bergmair (rbergmair@acm.org) is a final-year undergraduate student of Computer Science at the University of Derby in Austria. He completed a five-year program in Computer Science at the level of secondary education and has been doing Software Engineering for IBM since he was 14. His primary research interests are Computational Linguistics and Logic Programming.