OPTIMALITY THEORY AND THE THREE LAWS OF ROBOTICS[1]

Gabriel de Avila Othero[2]

Once, while reading an article that criticized Optimality Theory (Rennison 2000), I thought the following remark to be quite interesting: "OT is a theory of anything and everything that it is applied to" (p. 140). This seems to be true about OT. When you first start studying OT, you soon realize that some of its principles can really be applied to different domains, not only to linguistic domains. Actually, some of the OT principles can be applied to basically any kind of decision we need to take in life. And that was exactly the characteristic that I found to be the most attractive in the theory. After all, OT is about **solving conflict situations**, usually situations with **different kinds of demands**. And these demands are not necessarily grammatical or linguistic in nature. We can apply OT to other areas and domains of our lives.

I believe the first text that I have read on OT was a short text by Elan Dresher (Dresher 1996) in which he tells an anecdote about Jewish elders discussing Jewish laws and rituals back in the first century A. D. The story shows us how the 'spirit' of OT in solving problems involving different demands could help the wise Jewish elders. Later I also read in a Linguistics blog that 'OT is not so much a theory of phonology or syntax as a philosophy of life. Life makes conflicting demands, and to satisfy some we must violate others.'[3] And this, in my opinion, is a great merit of the theory. After all, wouldn't it be great if we could explain language phenomena following the basic principles and ways of reasoning we so often use to solve our everyday problems?

OT is a theory that deals with **constraints** that can be violated whenever needed. For instance, if a principle A says '**sleep eight hours per night**' and a principle B says '**wake up at 7 a.m. to go to work**', we face three logical alternatives:

[1] Thanks to Leda Bisol and Eric Bakovic for their comments on this text.
[2] Postdoc in Linguistics at Universidade Federal do Rio Grande do Sul – UFRGS; CNPq. E-mail: gab.othero@gmail.com .
[3] www.garyfeng.com/wordpress/2006/02/21/the-rise-of-optimality-theory-in-first-century-palestine/

i. I can obey both principles and go to bed at 11 p.m. to wake up at 7 a.m.;
ii. I can go out at night and return at 2 a.m. and still obey principle A: I sleep my eight hours, wake up at 10 a.m. and violate principle B;
iii. I can go out at night and return at 2 a.m. and still obey principle B: I sleep only five hours – violating principle A – in order to wake up at 7 a.m. and go to work.

It all depends on which I consider to be the most important principle in the **hierarchy** I want to adopt, i.e. I have to organize the ranking of constraints and respect the most important one. It can be **A >> B** (A is more important than B) or **B >> A** (B is more important than A). A graphical way to represent this ranking is the following[4]:

Candidates		**Principle A** (8 hours of sleep)	**Principle B** (wake up at 7:00)
i.	I go to bed at 11 p.m. and wake up at 7 a.m.		
ii.	I go to bed at 2 a.m. and wake up at 10 a.m.		*
iii.	I go to bed at 2 a.m. and wake up at 7 a.m.	*	

If I believe both constraints are equally important (**A <<>> B**), then the best scenario for me is (i) and I will go to bed at 11 p.m. to wake up at 7 a.m. However, if eventually I go to bed at 2 a.m. (that is, if candidate (i) is no longer available), I will have to make a choice: if principle A is more important to me, then (ii) is the best candidate scenario: I will be late for work (i.e. I will violate principle B) in order to have my eight hours of sleep (i.e. in order to satisfy principle A). On the other hand, if I believe principle B is more important, then the optimal candidate for me will be candidate (iii): I will give up some hours of sleep (violating principle A) in order to wake up at 7 a.m. and not be late to work (i.e. in order to satisfy principle B).

The moral of the story is the following: all constraints must be **ranked**, and it is perfectly possible to **violate** a less important constraint if necessary in order to respect a **higher constraint** in my ranking.

But what do Isaac Asimov and the Three Laws of Robotics from the title have to do with this?

[4] The star (*) marks a violation to the principle.

Let's see. Asimov created the **three laws of Robotics**:[5]

1) A robot may not harm a human being or, through inaction, allow a human being to come to harm.
2) A robot must obey any orders given to it by human beings, except where such orders would conflict with the First Law.
3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

At first sight, these laws look trivial. When we take a closer look, though, we can see that they are very well designed and planned. And – most importantly for the purposes of our text – they have the 'OT way of thinking'. Let's see why.

These three laws of Robotics could have been perfectly expressed as classical logic rules, such as (1'), (2') and (3') below:

$\forall$x, x(robot), x must[6]

1') not harm a human being;

2') obey the orders given to it by a human being;

3') protect its own existence.

The instructions in (1'), (2') and (3') seem clear enough for machine understanding. However, let us imagine the following scenario: I have a robot and I command it to jump from a bridge. What would the robot do? Would it obey my command, as it is stated in (2'), and jump? Or would it remain still, not doing anything in order to protect its own existence, following rule (3')? And what would the robot do if I ordered it to kill somebody? Would it obey my orders and kill a person? Or would the robot follow (1'), not harming, then, a human being? And finally, what would happen if I'd ask the robot to kill *me*? Would it obey my command?

[5] These laws became especially famous because of the movie *I, robot* (2004). They were first published by Asimov in his short story *Runaround* in 1942. The idea of using the laws as an example of Optimality Theory is not mine and is not new; McCarthy (2002: 3-4) and (2008: 12-13) also cite Asimov's Laws: "To draw an analogy from ethics, optimality is more like moral relativism or the Three Laws of Robotics than the Ten Commandments; it is about being the best among a choice of options, not about being objectively perfect." (McCarthy 2002: 3-4). Giannakoulopoulos (2001) and Trommer (2008) have also developed similar comparisons. I thank Eric Bakovic for warning me about that.

[6] Or: 'for every x, if x is a robot, then x must'.

If the three laws of Robotics were implemented as (1'), (2') and (3'), the robot would probably halt in an infinite loop, since it would try to follow all the orders at the same time, one contradicting the other. That would most certainly damage the 'positronic' brain of the robot.

The interesting part in the three laws of Robotics as proposed by Asimov is the part where it is stated 'obey this law **unless it contradicts a more important law**'. Asimov used in his laws of Robotics the same kind of reasoning we use to solve conflicts in OT. Notice that there is a dominance relation among the three laws: **Law 1 >> Law 2 >> Law 3**.

Hence, if I command a robot to jump from a bridge, *it will jump*. We can represent that with the following table (or ***tableau***).[7]

What happens if I order a robot to jump from a bridge?

Candidates		Law 1	Law 2	Law 3
a.	☞ The robot jumps from the bridge			*
b.	The robot does not jump from the bridge		*	
c.	The robot throws *me* from the bridge	*	*	
d.	The robot does not do anything		*	

This tableau shows us that the robot will not halt in an infinite loop. On the contrary, it will select the **optimal candidate** among all logical possibilities – the optimal candidate being the one which violates the lowest law, i.e. candidate (a). Candidates (b) and (d) violate a higher law (Law 2); and candidate (c) violates both Laws 1 and 2.

And what would happen if I ordered the robot to kill a person? We can see that in the tableau below:

Candidates		Law 1	Law 2	Law 3
a.	The robot 'kills' itself		*	*
b.	The robot kills a person	*		
c.	The robot kills *me*	*	*	
d.	☞ The robot does not do anything		*	

[7] The symbol ☞ indicates the optimal candidate, the option that will be chosen as the best option by the robot.

The answer: the robot would not do anything; not because it halts in a loop, but because it will follow the hierarchy of the rules. Even though this (lack of) action disrespects the second law of Robotics – candidate (d) –, candidates (b) and (c) violate the highest law on the ranking, and candidate (a) violates two laws.

Finally, what would happen if I ordered a robot to kill me?

Candidates		Law 1	Law 2	Law 3
a.	The robot 'kills' itself		*	*
b.	The robot kills a person	*		
c.	The robot kills *me*	*	*	
d.	☞ The robot does not do anything		*	

Again, the optimal candidate says that the robot wouldn't be able to do anything. That is its only logical choice, in an OT-like analysis. By programming the robots with these laws, Asimov prevented the robots from harming any human being. And he was a pioneer in Science Fiction for that – he wrote his stories about harmless intelligent robots, unlike most of the sci-fi(\terror) movies and stories there were (and there still are) at the time. Asimov's robots are logically unable to harm any human being. Even before his time, Asimov expressed an 'OT way of thinking'.

Where to go from here? There are some very good introductory books on OT, such as Archangeli & Langendoen (1997), Kager (1999), McCarthy (2002), and McCarthy (2008). And an excellent source of material on OT is available at ROA (*Rutgers Optimality Archive*), at http://roa.rutgers.edu.

REFERENCES

ARCHANGELI, Diana; LANGENDOEN, D. Terence (eds.) (1997) *Optimality Theory: an overview*. Oxford: Blackwell.

DRESHER, Elan. (1996) The Rise of Optimality Theory in First Century Palestine. In: *GLOT International* 2, 1/2, January/February.

GIANNAKOULOPOULOS, Andreas P. (2001) *Frog Leaps and Human Noises. An Optimality Theory Approach to Cultural Change*. Institute for Logic, Language and Computation, Universiteit van Amsterdam. Technical report.

KAGER, René. (1999) *Optimality Theory*. Cambridge: Cambridge University Press.

McCARTHY, John. (2002) *A thematic guide to Optimality Theory*. Cambridge: Cambridge University Press.

McCARTHY, John. (2008) *Doing Optimality Theory*. Malden, MA and Oxford: Blackwell.

RENNISON, John. (2000) OT and TO: On the status of OT as a theory and a formalism. *The Linguistic Review* 17, 2-4.

TROMMER, Jochen. (2008) A Crash Course in Optimality Theory, EGG 2008.