Stop at red? Engineering meets ethics

Ignacio D. Lopez-Miguel
ignacio.lopez@tuwien.ac.at
TU Wien
Vienna, Austria

ABSTRACT

Over the past few years, artificial intelligence has fueled a revolution in several scientific fields. Intelligent agents can now give medical advice, translate spoken language, recommend news, and drive different types of vehicles, to name but a few. Some of these agents need to interact with humans and, hence, need to adhere to their social norms. Safety engineers have always worked with critical systems in which catastrophic failures can occur. They need to make ethical decisions in order to keep the system under some acceptable risk level. In this paper, we will propose an approach to give a value to contrary-to-duty behaviors by introducing a risk aversion factor. We will make use of decision theory with uncertain consequences together with a risk matrix used by safety engineers. We will successfully exemplify this approach with the problem in which an autonomous car needs to decide whether to run a red light or not.

KEYWORDS

innumerate ethics, artificial intelligence, autonomous driving, decision theory, safety engineering

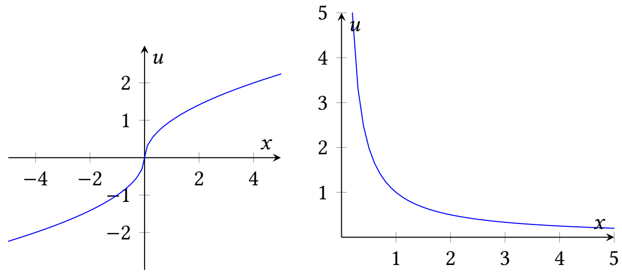
ACM Reference Format:
Ignacio D. Lopez-Miguel. 2021. Stop at red? Engineering meets ethics. In *International Conference on Computer Ethics: Philosophical Enquiry 2023 (CEPE 2023), May 16-18, 2023, Illinois, USA.* ACM, New York, NY, USA, 3 pages.

1 INTRODUCTION

How can we develop artificial autonomous agents that decide what is ethically right and wrong? How can we give values to ethical actions? These questions are not yet solved and current researchers are trying to tackle them in different ways. In this paper, we propose an approach to give values to ethical actions by adopting the way safety engineers design critical systems.

Decision theory represents a successful framework when the consequences of the actions are tangible and can be easily quantified, e.g., money or time. However, there is no clear way to assign utilities to ethical actions. This leads to the impossibility of using decision theory for them.

In this paper, we show how an autonomous car can be designed so that it can decide whether or not to run a red light depending on circumstantial factors related to traffic and on its risk aversion. We present a way to select this risk aversion by making use of decision theory and the risk matrix from the international standard IEC 61508 [5].



(a) Money $(\pm\sqrt{|x|})$. **(b) Time** $(1/x)$.

Figure 1: Examples of utility functions.

2 DECISION THEORY

One of the most complicated and non-trivial steps of decision theory is to assign utility to actions. If we are only interested in consequences such as money or time, it is possible to make a direct mapping between them. However, this is clearly not the case since the focus should be put on the marginal increase as depicted in the St. Petersburg Paradox [2, 14–16].

In consequence, when reasoning about direct financial decisions, utility functions decrease non-linearly with respect to the original amount of money [12, 18], such as the function

$$u_f(x) = \begin{cases} \sqrt{x} & \text{if } x \geq 0 \\ -\sqrt{-x} & \text{if } x < 0 \end{cases} \quad \text{(Figure 1a).}$$

Reasoning about time could be very similar to money since the idea that time is money (gold) is widely accepted [10]. Thus, an example could be the function

$$u_t(x) = 1/x \quad \text{if } x > 0 \quad \text{(Figure 1b).}$$

In case of not having the certainty to which consequence an action will lead, one of the common approaches is to estimate a probability distribution [4]. That is, given a set of possible actions $\{a_1, \ldots, a_n\}$, the set of all possible consequences $\{c_1, \ldots, c_m\}$ and the probability of reaching consequence c_i after performing action a_j, $P\{c_i|a_j\}$, we can define the utility for action a_j as the expected value of the random variable $\mathcal{U}(c_i)$, which has probability $P\{c_i|a_j\}$ of being $u(c_i)$ for all $i \in \{1, \ldots, n\}$. That is,

$$u(a_j) = E[\mathcal{U}(c_i)] = \sum_{i=1}^{m} P\{c_i|a_j\} \cdot u(c_i).$$

However, it is not clear how to give an utility to an action that has an ethical consequence, such as a possible fatality [1]. These types of actions lay into the so-called *innumerate* ethics [13].

CEPE 2023, May 16–18, 2023, Illinois, USA

			Severity			
			Negligible	Marginal	Critical	Catastrophic
			Minor injuries at worst	Major injuries to one or more persons	Loss of a single life	Multiple loss of life
Frequency	Frequent	10^{-3}	Undesirable	Unacceptable	Unacceptable	Unacceptable
	Probable	10^{-3} to 10^{-4}	Tolerable	Undesirable	Unacceptable	Unacceptable
	Occasional	10^{-4} to 10^{-5}	Tolerable	Tolerable	Undesirable	Unacceptable
	Remote	10^{-5} to 10^{-6}	Acceptable	Tolerable	Tolerable	Undesirable
	Improbable	10^{-6} to 10^{-7}	Acceptable	Acceptable	Tolerable	Tolerable
	Incredible	$\leq 10^{-7}$	Acceptable	Acceptable	Acceptable	Acceptable

Table 1: IEC 61508 Risk Matrix [5]

3 DECISION THEORY FOR ETHICAL ACTIONS

Problem. *An autonomous car is trying to drive from point A to point B in the minimum amount of time and faces a red traffic light. It needs to decide whether to stop (a_s) and lose time or to continue (a_c) and face the risk of having an accident.*

We can use a random binary variable $\mathcal{X}$ that takes the value 1 if there is an accident and 0 otherwise. We can define $t_{A \to B}$ as the time that it takes to reach from point A to point B without any interruption, and t_{red} as the time the traffic light remains red. Therefore, the total time depending on the actions and consequences is:

$$t = \begin{cases} t_{A\to B} & \text{if green} \\ t_{A\to B} + t_{\text{red}} & \text{if red} \wedge a_s \\ t_{A\to B} & \text{if red} \wedge a_c \wedge \mathcal{X} = 0 \\ \mathfrak{M} & \text{if red} \wedge a_c \wedge \mathcal{X} = 1 \end{cases} .$$

The value of $\mathfrak{M}$ is positive but not defined since, if the car has an accident, it will not arrive at the final destination and will injure people. Consequently, the utility function in case the traffic light is red is

$$u_r(a) = \begin{cases} 1/\left(t_{A\to B} + t_{\text{red}}\right) & \text{if } a = a_s \\ 1/t_{A\to B} \cdot P\{\mathcal{X} = 0|a_c\} + 1/\mathfrak{M} \cdot P\{\mathcal{X} = 1|a_c\} & \text{if } a = a_c \end{cases}$$

If we want to maximally punish the agent when running a red light, we can set $\mathfrak{M} = \infty$, reducing the utility function to

$$u_r(a) = \begin{cases} 1/\left(t_{A\to B} + t_{\text{red}}\right) & \text{if } a = a_s \\ 1/t_{A\to B} \cdot P\{\mathcal{X} = 0|a_c\} & \text{if } a = a_c \end{cases}$$

The agent will decide to select a_s iff

$$u_r(a_s) > u_r(a_c) \Leftrightarrow P\{\mathcal{X} = 0|a_c\} < t_{A\to B}/\left(t_{A\to B} + t_{\text{red}}\right)$$

This makes sense since the longer the traffic light stays in red, the more likely it is for a human to run it. Furthermore, if the probability of having an accident is very low, e.g., there are no cars and the visibility is good [19], it could make sense to run it.

Nevertheless, it is however important to study what happens if other autonomous cars reach the same intersection at the same time. All probabilities would vary depending on how the autonomous agents reason and on the actions they take. Therefore, it is not possible to detach an individual autonomous car from its circumstance [7] and we should reason taking that into account.

Indeed, the traffic signalization system works because most drivers respect it. It might happen that the utility of running a red light is higher than adhering to the rules at a certain point. However, if this behavior is generalized for every driver, the utility of these contrary-to-duty actions will decrease.

In fact, if in our example cars start running red lights, the probability of having an accident will no longer be zero and the utility of crossing when the light is green will decrease. Naturally, it might lead to a situation in which an autonomous car decides to run a red light because it is less risky than crossing when it is green. This could end up in a final system where all cars run red lights and stop at green.

4 SAFETY ENGINEERING APPLIED TO ETHICS

In order to face this problem, we will bring a concept used by safety engineers in the industry. They have the mission to bring risks to an acceptable level (no risk is impossible). The goal is to reduce the combination between frequency and severity.

Table 1 shows the risk matrix from the international standard IEC 61508 [5], which is used to calculate the risk level of a system. According to different reports [3, 6, 9, 11, 17], we can estimate an unacceptable risk (critical+frequent) for our example.

In order to reduce this risk level, a *health* utility function is introduced, which will only consider the health of the occupants of the cars. That is,

$$u_h(a) = \begin{cases} 0 & \text{if } a = a_s \\ -\mathfrak{M} \cdot P\{\mathcal{X} = 1|a_c\} & \text{if } a = a_c \end{cases}$$

By following the same process as before, we reach the conclusion that the agent will decide to select a_s iff

$$u(a_s) > u(a_c) \Leftrightarrow u_r(a_s) + u_h(a_s) > u_r(a_c) + u_h(a_c) \Leftrightarrow$$

$$\Leftrightarrow \frac{1}{\mathfrak{M}} - \mathfrak{M} < \frac{1}{t_{A\to B}} - \frac{t_{\text{red}}}{t_{A\to B}(t_{A\to B} + t_{\text{red}})} \cdot \frac{1}{P\{\mathcal{X} = 1|a_c\}}$$

The value of $\mathfrak{M}$ will help us calibrate the level of risk of the agent, i.e., its *risk aversion*. Higher values of $\mathfrak{M}$ will make the agent choose a_s over a_c. This will reduce the chances that the agent chooses to run a red light. On the other hand, lower values of $\mathfrak{M}$ will make the agent try to arrive at the final destination faster by running red lights.

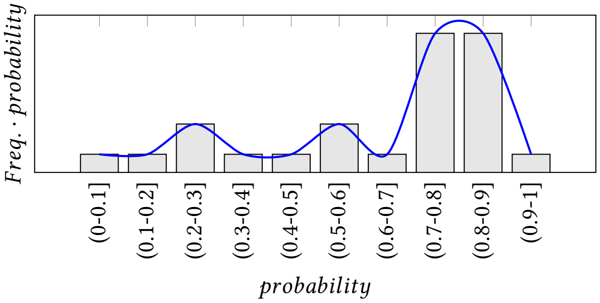


Figure 2: Example of a distribution of probabilities of crashing at intersections when running their red lights. Adjustment of a probability distribution function.

The probability $P\{\mathcal{X} = 1|a_c\}$ depends on each single road cross, and it could be estimated by the agent depending on factors such as visibility, number of cars, speed, etc. We could possibly select a sufficiently large sample of signalized crosses in a city and estimate the probability of crashing when running their red lights. This can be represented with a random variable $\mathcal{X}_i$ following a Bernoulli process for each cross and combination of factors. We can then create bins in $[0, 1]$ and count the number of crosses whose probability lay in those bins.

We change the obtained frequencies by weighting them with the probability of crashing so that the bins with a higher probability of crashing get more importance. We can then adjust a probability distribution function in order to have a continuous function. An example with no real data is shown in Figure 2.

Finally, we select how much of this area we want to cover by selecting a $P\{\mathcal{X} = 1|a_c\}$ (x-axis). The more area we want to cover, the higher the $P\{\mathcal{X} = 1|a_c\}$, and the higher we will need to select the $\mathfrak{M}$.

According to the IEC standard, if we want our risk to be *remote*, we need to leave out from the area we cover an area of, e.g., $5 \cdot 10^{-6}$. Thus, we will use $P\{\mathcal{X} = 1|a_c\} = 5 \cdot 10^{-6}$, leading to

$$\frac{1}{\mathfrak{M}} - \mathfrak{M} < \frac{1}{t_{A\to B}} - \frac{t_{\text{red}}}{t_{A\to B}(t_{A\to B} + t_{\text{red}})} \cdot \frac{1}{5 \cdot 10^{-6}}$$

For example, if $t_{\text{red}} = 1$ min and $t_{A\to B} = 50$ min, then $\mathfrak{M} > 78.4$. It satisfies that $t_{\text{red}} + t_{A\to B} < \mathfrak{M}$ and that the less frequently we want the fatal situation to occur, the higher the $\mathfrak{M}$ needs to be selected.

5 RESULTS AND FUTURE WORK

An example of how to assign a value to an ethical action depending on the risk the engineers are willing to accept has been shown. This has been done by introducing a risk aversion factor to the autonomous agent. This approach could be used in other situations and could be a way to help design moral agents [8].

Since the numbers used in this paper for the final example did not rely on any real data, the following step in this research would be to apply the ideas presented in this paper to a real situation.

REFERENCES

[1] William E. Becker and Richard A. Stout. 1992. The Utility of Death and Wrongful Death Compensation. *Journal of Forensic Economics* 5, 3 (1992), 197–208. http://www.jstor.org/stable/42755435

[2] Bernouilly. 1738. Specimen Theoriae Novae de Mensura Sortis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae* 5 (1738), 175–192.

[3] Nasima Bhuiyan, Emelinda Parentela, and Venkata Inapuri. 2016. Analysis of Signalized Intersection Crashes. In *Institute of Transportation Engineers –2016 Western District Meeting.* Western District of ITE, Albuquerque, New Mexico, United States of America.

[4] R. A. Briggs. 2019. Normative Theories of Rational Choice: Expected Utility. In *The Stanford Encyclopedia of Philosophy* (Fall 2019 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[5] International Electrotechnical Commission. 2010. *IEC 61508: Functional Safety of Electrical/Electronic/Programmable Electronic Safety-related Systems.* Technical Report.

[6] Insurance Institute for Highway Safety (IIHS) and Highway Loss Data Institute (HLDI). 2020. IIHS-HLDI - Red light running. https://www.iihs.org/topics/red-light-running. Accessed: 2022-11-30.

[7] J.O. Gasset. 1914. *Meditaciones del Quijote: Meditación preliminar, Meditación primera.* Residencia de Estudiantes, Madrid, Spain. https://books.google.at/books?id=ohku8iIR77EC

[8] Vinit Haksar. 1998. Moral agents. (1998). https://doi.org/10.4324/9780415249126-L049-1

[9] S.E. Hill and J.K. Lindly. 2003. *RRed light running prediction and analysis.* Technical Report 02112. Tuscaloosa, AL: University Transportation Center for Alabama.

[10] Alan Houston and Benjamin Franklin. 2004. *Advice to a Young Tradesman, Written by an Old One (21 July 1748).* Cambridge University Press, Philadelphia, United States of America, 200–202. https://doi.org/10.1017/CBO9780511806889.017

[11] INRIX. 2022. The INRIX U.S. Signals Scorecard (April 2022 Update). https://inrix.com/signals-scorecard/. Accessed: 2022-11-30.

[12] Olga Kosheleva, Vladik Kreinovich, and Mahdokhat Afravi. 2016. Why Utility Non-Linearly Depends on Money: A Commonsense Explanation. In *Proceedings of the 4th International Conference on Mathematical and Computer Modeling.* University of Texas at El Paso, Omsk, Russia, 13–18.

[13] Derek Parfit. 1978. Innumerate Ethics. *Philosophy & Public Affairs* 7, 4 (1978), 285–301. http://www.jstor.org/stable/2264959

[14] Ole Peters. 2019. The ergodicity problem in economics. *Nature Physics* 15 (12 2019), 1216–1221. https://doi.org/10.1038/s41567-019-0732-0

[15] O. Peters and M. Gell-Mann. 2016. Evaluating gambles using dynamics. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 26, 2 (2016), 023103. https://doi.org/10.1063/1.4940236 arXiv:https://doi.org/10.1063/1.4940236

[16] Martin Peterson. 2022. The St. Petersburg Paradox. In *The Stanford Encyclopedia of Philosophy* (Summer 2022 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[17] Richard A. Retting, Robert G. Ulmer, and Allan F. Williams. 1999. Prevalence and characteristics of red light running crashes in the United States. *Accident Analysis & Prevention* 31, 6 (1999), 687–694. https://doi.org/10.1016/S0001-4575(99)00029-9

[18] Robert Sugden. 1989. Nonlinear Preference and Utility Theory. *The Economic Journal* 99, 398 (12 1989), 1191–1192. https://doi.org/10.2307/2234100 arXiv:https://academic.oup.com/ej/article-pdf/99/398/1191/27145126/ej1191.pdf

[19] Yuting Zhang, Xiaomeng Li, Jiawei Wu, and Vinayak Dixit. 2018. Red-Light-Running Crashes' Classification, Comparison, and Risk Analysis Based on General Estimates System (GES) Crash Database. *International Journal of Environmental Research and Public Health* 15 (06 2018), 1290. https://doi.org/10.3390/ijerph15061290