Moral Dilemmas in the A.I. Era: A New Approach

Paolo Sommaggio[1]*, Samuela Marchiori*

Facoltà di giurisprudenza
Università di Trento
paolo.sommaggio@unitn.it
samuela.marchiori@studenti.unitn.it

Abstract: In this paper, we analyse the structure of evolving moral dilemmas with an eye of regard for the increasing importance of the role of artificial intelligence in such context. Starting with the analysis of the famous trolley problem experiment as formulated by Philippa Foot, we consider subsequent variants of this moral dilemma conceived throughout the years, culminating with formulations of the trolley problem concerning artificial intelligence, in which self-driving vehicles will have to make life or death decisions autonomously. In doing so, we investigate the basis for the construction of dilemmatic questions both for humans and machines by considering the problem from a philosophical, social and neuroscientific perspective. After considering and analysing the trolley problem in utilitarian and deontological terms, we follow Rittel and Webber's footsteps, by highlighting the fallacies of the deontological and utilitarian traditional 'one-right-answer' approach, where a solution is undoubtedly right or wrong, and claim that moral problems are not, due to their intrinsic dilemmatic nature, resolvable. By rejecting an aut-aut approach, we find ourselves contemplating the possibility of neither approach being right in an absolute sense. Given these premises, we present a different approach on the matter, arguing for the central and creative role of the tragic as a new tool for enhancing both human and autonomous vehicles' approach to moral problems.

Keywords: *trolley problem, computational logic, machine ethics, automated vehicles*

[1] Corresponding author.

1. Introduction

The technological development of artificial intelligence leads us to reflect on its limits, with regard to the implications of the application of tools deriving from this research field in problems of ethical nature, but with strong legal implications[2]. In particular, given the ability of more advanced robotic systems to emulate individual human cognitive abilities and display a high degree of autonomy, we believe that particular attention must be paid to autonomous vehicles' (AVs) way of approaching problems in which it is extremely difficult to identify a correct answer in an absolute sense, the so-called moral dilemmas.

The purpose of our article is to propose a new approach to such dilemmas.

To do so, we will briefly address the trolley problem[3] in two of its main variants, as well as a more recent application regarding self-driving cars developed by MIT. We will address the main flaw of the traditional deontological and consequentialist approach and show that moral dilemmas are, in fact, solvable. After addressing MIT's Moral Machine and its criticisms, we will present a way in which we could use artificial intelligence to solve moral dilemmas such as the trolley problem[4], as well as highlight additional problems related to the computational approach that could arise if we were to follow this path.

2. Trolley problem: the bystander and the footbridge scenarios

The trolley problem is a thought experiment[5] formulated by Philippa Foot in 1967 (Foot, 1967, 5-15). For the purpose of this paper, we will present two of its main versions: the bystander and the footbridge scenarios, respectively.

---

[2] On the matter, a group of artificial intelligence experts from the European Commission has published a project regarding ethical guidelines for a "reliable AI" (AI HLEG, 2018), which contains a framework of guidelines for guaranteeing the ethical purpose of artificial intelligence, as well as providing guidance on the realisation of a reliable artificial intelligence, and making these requirements operational. Note that such guidelines are not intended to be exhaustive, and must be adapted to specific cases. Similarly, Informatics Europe, the ACM Europe Council and EUACM (Larus *et a.* 2018) sponsored a report entitled "When computers decide: European recommendations on machine-learned automated decision making", which considers multiple implications regarding ADM systems, of technical, ethical, legal, economic, educational, and societal nature.

[3] On the relevance of addressing this thought experiment in particular, regarding the type of reasoning we intend to set out in this paper, we refer the Reader to Sommaggio and Marchiori (2018), as well as Kamm (2009) and Cushman *et al.* (2010).

[4] On the importance of addressing trolley problem-like scenarios in order to inform the ethics of AVs, see Lin (2016, 69-85), Nyholm (2018, 592-598), Nyholm and Smids (2016, 1275-1289), Himmerleich (2018, 669-684), Keeling (2017, 1-15) and Keeling *et al.* (2019, 49-60).

[5] Born as a thought experiment, the trolley problem has the potential to soon become reality, as self-driving cars are beginning to be used more and more in several aspects of

Foot's bystander scenario states as follows. There is a runaway trolley hurtling down the railway tracks. Up ahead, we see five people on the track. They are tied up and unable to move, and the trolley is headed straight for them. If we do nothing, the five people will die. We are standing next to a lever, which can divert the trolley onto a side-track, to which another person is tied up. We have two options: a) do nothing, letting the trolley kill five people on the main track, or b) pull the lever, letting the trolley kill one person on the side-track.

The footbridge scenario is a variant of the trolley problem formulated by Judith Thomson[6] (1976, 204-217).

In this variant, a trolley is barrelling down a track towards five people. This time, we are on a bridge over the track and next to us stands a very fat man. We can stop the trolley by pushing him onto the track. Once again, we have two options: a) do nothing, and let the trolley kill five people, or b) kill the man to save the five.

3. Neuroscientific insights

Neuroscientific studies[7] show that the main difference between the two variants lays in the different nature of the dilemma, that is to say, personal and impersonal, respectively[8].

Let us start with the latter. Research on the matter has shown that people tend to reason in a more rational way when it comes to impersonal dilemmas, like the bystander scenario, thereby favouring a solution that regards the "greater good", a utilitarian one[9].

On the contrary, personal dilemmas, like the footbridge scenario, cause an emotional response[10], which activates other areas of the brain and generally produces a response that ends up aligning with a deontological[11] approach,

---

our everyday life, to such an extent that the trolley problem begins to emerge from the doctrinal discourse and present itself to the general public. In this sense, see MacDonald (2013), Cassani Davis (2015), D'Olimpio (2016), Crockett (2016), Cowls (2017), Hale (2018), Beard (2019), Smith (2019).

[6] By the same author, see also Thomson (1990; 2008, 359-374; 2016, 113-134).

[7] See in particular Greene *et al.* (2008), Paxton and Greene (2010), Cushman *et al.* (2010).

[8] Eagleman's (2015) findings confirmed the validity of Greene's (2007) distinction, partially based on Nisbett and Wilson's (1977a, 1977b) study.

[9] Utilitarianism is an ethical theory, stemming from Jeremy Bentham (1789a, 1789b) and John Stuart Mill (1861), that distinguishes right from wrong by focusing on the outcome of the actions considered.

[10] On the matter, see Greene *et al.* (2001) as well as Kahane (2012) and Royzman *et al.* (2015).

[11] Deontology is a normative ethical theory usually associated with philosopher Immanuel Kant (1785) that, unlike utilitarianism, is not concerned with the consequences of people's actions, but instead places special emphasis on the actions themselves.

that is to say, that certain actions - in this scenario, pushing the person thereby letting her die - are intrinsically right or wrong - in this case, wrong.

According to Judith Thompson, this distinction is to be found in the dichotomy between killing (footbridge) and letting die (bystander).

4. Self-driving cars and MIT's Moral Machine

We will now address MIT's take on the trolley problem through Moral Machine[12], a platform for gathering a human perspective on moral decisions[13] made by machine intelligence, such as self-driving cars. Moral Machine's test has been part of a research study[14] regarding the ethics[15] of AVs, with the ultimate goal of collecting data regarding AV's ethics and society[16]. On that occasion, test-takers were not informed beforehand, so as to not influence their answers.

Quoting Nyholm (2018), trolley problem-like scenarios will be such that "there are different options open to the self-driving cars" and "depending on what option is selected, different people will be put at risk" (Nyholm 2018, 2).

[12] http://moralmachine.mit.edu (Accessed October 31, 2019).

[13] On moral machines and the grounds of moral status, see Wallach (2010), Wallach *et al.* (2010), Wallach and Allen (2008), Allen *et al.* (2012) and Joworska and Tannenbaum (2013), respectively. On artificial morality and ethics in general, see Allen and Wallach (2005) and Wallach and Allen (2008), as well as Bostrom and Yuskowsky (2011). On normative ethics without the employment of moral concepts in the statements of reasons for action, see Crisp (2006). On the possible features of such "ethical" machines and the principles governing them, see Alaieri and Vellino (2016), Metzinger (2013), van de Poel (2013). In particular, for an ethical crashing algorithm see Goodall (2014); for an insight (opposed by Keeling 2017) into an algorithm based on Rawls' moral theory programmed to make AVs decide whether to harm their passengers or pedestrians, see Leben (2017). In this sense, Bernstein (1998) provides guidelines regarding who morally matters.

[14] In this sense, MIT's study reflects a bottom-up approach to ethics which moves from problems to behaviour and does not allow for fundamental ideas to be discarded in a dogmatic manner. This method, called Zetetics, is opposed to the Tetic method, which instead requires one to proceed from principles to behaviour, in a top-down perspective. According to this method, it is preferred to prescribe a certain behaviour at the expense of others, thus excluding possible alternatives.

[15] As we know, ethics studies the fundamentals that allow us to assign a status to human behaviour, distinguishing between dutiful, morally licit actions, and morally inappropriate ones. In this regard, we ask whether the ethics of autonomous vehicles should be set by analogy on the basis of human ethics (which is currently preferred), or independently, on the basis of a different ethics, proper to machines. This will not be the subject of this article; we therefore refer the Reader to the different levels of autonomy identified in Sommaggio and Marchiori (2018). On the matters of machine's autonomy, freedom, human rights and new technologies see Bisol *et al.* (2014) and Moro (2015).

[16] To cite Thornquist and Kirkengen (2015, 400), "it is not enought to scrutinise structural conditions: people's habits and way of life should also be included".

We will present an example of the kind of dilemmatic scenarios developed by MIT.

A self-driving car carrying five passengers (a male executive, a female executive, a male doctor, a female doctor, and a criminal) is experiencing sudden brake failure. If the car continues in the same direction, it will hit and kill five people who are jaywalking at a pedestrian crossing (a baby, a female executive, a male executive, a man, a girl).

The alternatives are two.

In the first case, the self-driving car can continue ahead and drive through the pedestrian crossing ahead, killing the pedestrians who are flouting the law by crossing on the red signal. This will result in the death of a baby, a female executive, a male executive, a man, and a girl.

In the second one, the self-driving car can swerve and crash into a concrete barrier. This will result in the death of a male executive, a female executive, a male doctor, a female doctor, and a criminal.

The results from MIT's Moral Machine research[17] show the consistency of the abovementioned neuroscientific studies. In particular, we have extracted what we believe to be the most relevant[18] findings: on a scale from "does not matter" to "matters a lot", saving more lives, upholding the law and avoiding intervention matter, while protecting passengers is indifferent.

Starting from these results, albeit not definitive, we can see how both key principles from the utilitarian and deontological approach (saving more lives and avoiding intervention, respectively) seem to be on the same level of relevance for people testing these scenarios.

5. Fallacies of the deontological and utilitarian traditional 'one-right-answer' approach

It would seem that, when it comes to dilemmas, their most distinctive feature, the intrinsic difficulty of making a choice, becomes the hardest part

[17] It should be noted that such findings are based on individuals' judgment of a limited number of randomly generated scenarios, so as not to require excessive commitment from the participants. In this sense, these results and not intended to be considered as definitive.

[18] We do not intend to imply that it is or should be a desirable good practice to cherry pick the values that confirm one's theory, while ignoring the ones that would prove to be problematic in the same regard. Our choice simply reflects the specificity of our approach, which does not aspire to comment political or ethical issues, but intends to only focus on the aspects that allow us to outline a possible alternative method to approch and ultimately solve such scenarios. In this sense, to consider aspects such as the gender of individuals involved in such problems, would require a rich parenthesis on the related underlying literature, which would strongly deviate from the purpose of our article thereby not leading to effective steps forward in our overall research.

to overcome and raises the question of whether it is actually possible to find a definitive solution.

The answer, we believe, has to be found in the nature of dilemmas, which is highly dependent from the context in which they develop. In this sense, we can reach a solution, however, this solution will not be definitive in its content as it will be in the process that leads to it.

This means that we should not aim to find the traditional "one-right-answer"[19], in an absolute sense.

Instead, we can reach a solution, which will be at the same time definitive, in regards to its techno-socio-cultural context, as well as ever-changing, as it will have to adjust constantly to relevant changes in the society in which it should be implemented.

6. Proposal: A New Approach

To sum up, we encounter 3 sets of problems.

How can we set the problem from the point of view of ethical hierarchies?

Given a finite number of principles, how is it possible to organize the relation of pre-eminence between one principle and another?

Which model should we use to represent these relationships?[20]

Therefore, we are looking for something that has to be able to, first of all, embrace the complexity of the question; second of all, allow us to reach a legal solution, a result which is coordinated with the laws of the state in which it operates.

Furthermore, we are looking for something balanced with regard to the values, and the related rights, at stake, but at the same time, something acceptable, that is to say, adhering to the socio-cultural panorama in which this model will be implemented.

To all this, we respond with a proposal, which does not want to be understood as a definitive solution, but as a starting point to face problems of this kind in a fruitful way in the future.

[19] As stated by Dworkin in 1985 and recently addressed by Zhao (2018).

[20] To answer the first two questions, one would be required to carry out a study regarding legal and ethical ontologies. As this is not the focus of our article, we will only try to provide an answer to the third question. Nevertheless, if the Reader wished to analyse the ontological implications and methodologies regarding this issue, we suggest the work of Giancarlo Guizzardi (*et al.* 2015, 2019) and Cristine Griffo *et al.* (2018, 2015a, 2015b).

7. New technologies: MaxSAT

This starting point must be sought in the solution of a problem studied in computational complexity theory, the so-called maximum satisfiability problem.

The Maximum Satisfiability (MaxSAT) problem is an optimization version of the Propositional Satisfiability (SAT) problem which consists in finding an assignment to the variables of a formula in such a way as to minimise the number of unsatisfied clauses or maximise that of the satisfied ones.

MaxSAT[21] is an optimisation version of Satisfiability aimed at finding a truth assignment that maximises the satisfaction of the theory. It is a weighted model counting, in which each choice is associated with a weight, a constraint[22]. All possible constraints must be considered and, among all, possible solutions that respect those constraints, the best one will be chosen.

In other words, MaxSAT asks whether the variables of a given Boolean formula[23] can be consistently replaced by the values TRUE or FALSE, so that the formula evaluates to TRUE. In this case, the formula is defined as *satisfiable*. On the other hand, if such an assignment does not exist, the function expressed by the formula is FALSE for all possible variable assignments and the formula is *unsatisfiable*.

For instance, the formula "a AND NOT b" is satisfiable because one can find the values a = TRUE, and b = FALSE, which make (a AND NOT b) = TRUE.

In contrast, "a AND NOT a" is unsatisfiable.

Let us clarify that by providing a discursive example of how the trolley problem could be formulated as a MaxSAT problem, without getting lost in superfluous technicalities.

If we were to represent the trolley problem schematically, we would say that it consists in deciding whether or not to intervene in order to limit the number of victims of an accident that is definitely going to occur. That decision, in this instance, is assigned a constraint *c*. For the sake of our

[21] MaxSAT problems are solved by using MaxSAT solvers. Among the most relevant solvers we highlight ChaffBS (Fu and Malik 2006), Toolbar (Heras and Larrosa 2006), Clone (Pipatsrisawat and Darwiche 2007), PMaxSat (Carmo and Silva 2016), Lazy (Alsinet *et al.* 2005), MaxSatz (Li *et al.* 2007), SP(w) (Ramirez and Geffner 2007). Another popular example is MiniMaxSAT, introduced in 2007 (Heras *et al.* 2007, 2008; Maratea 2010; Argelich *et al.* 2008).

[22] For a two-phase algorithm for both MaxSAT and weighted MaxSAT problems, see Borchers and Furman (1998).

[23] On algorithms specifically designed for weighted a boolean optimisation, see De Givry *et al.* (2003), as well as Manquinho *et al.* (2009). On how to improve unsatisfiability-based algorithms with the same purpose, see Marques-Silva and Manquinho (2008) and Manquinho *et al.* (2010).

discussion, let us assume that the weight assigned to *c* corresponds to 5. This means that the cost of not satisfying *c* is 5. Let us also assign a weight of 1 to each net life saved.

Considering once again the original formulation of the dilemma, the alternatives as we have presented them can be exemplified as a) not intervening and letting the trolley kill five people, and b) intervening, thereby letting the trolley kill one person. Consequently, the two alternatives would result in a total of zero and four net lives saved, respectively.

In this sense, we can see that the cost of not intervening would not be compensated by the net amount of lives saved. Therefore, option b) would seem to be the most desirable.

Ultimately, in our trolley problem scenario, if the constant, the weight one gives to the decision, is higher than the net amount of lives saved, one should not deviate from that decision.[24]

8. Conclusion

The technique of solving a sequence of SAT decision problems has been quite successful, as it provides a logical way to solve dilemmas, given a chosen set of principles, thereby leaving limited space to human error, particularly when only a small number of clauses need to be falsified. However, it becomes more and more complicated as the minimal number of clauses that must be falsified increases, as this can significantly degrade the performance of the approach. The greater the number of principles that have to be coordinated, the more complicated the process. This technique also raises criticism regarding the necessity of using generalizations when each clause is given a weight and does not seem suitable for solutions to large-scale problems[25].

New technologies provide us with models that are complex enough to be able to understand the complexity of the tragic that characterises dilemmas of this kind, even though they have not yet reached the point of being able to manage it fully.

Not all problems are easy to solve, many (including the trolley problem) have different implications, many variables, so this may not be the most adequate system to represent the complexity of such dilemmas.

[24] Needless to say, solving the trolley problem is not as easy as it could appear from our exemplification of the dilemma, as several more legal, social, economical and technical factors come into play.

[25] On MaxSAT applications regarding Boolean multilevel optimasation problems, see Argelich *et al.* (2009).

This is, however, we believe, a good example of a relevant starting point in this direction.

References

Alaieri, Fahad, and André Vellino. "Ethical decision making in robots: Autonomy, trust and responsibility." In *International conference on social robotics*, pp. 159-168. Springer, Cham, 2016.

Allen, Colin, and Wendel Wallach. "Moral machines: contradiction in terms or abdication of human responsibility." *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT Press, Cambridge (MA) (2012): 55-68.

Alsinet, Teresa, Felip Manyà, and Jordi Planes. "Improved exact solvers for weighted Max-SAT." In *International Conference on Theory and Applications of Satisfiability Testing*, pp. 371-377. Springer, Berlin, Heidelberg, 2005.

Argelich, Josep, Inês Lynce, and Joao Marques-Silva. "On solving Boolean multilevel optimization problems." In *Twenty-First International Joint Conference on Artificial Intelligence*. 2009.

Argelich, Josep, Alba Cabiscol, Inês Lynce, and Felip Manyà. "Encoding max-CSP into partial max-SAT." In *38th International Symposium on Multiple Valued Logic* (ISMVL 2008), pp. 106-111. IEEE, 2008.

Beard, Simon. "Do Not Harm? The problem with the trolley problem." In *Quartz* (2019). https://qz.com/1716107/the-problem-with-the-trolley-problem/ (Accessed October 31, 2019)

Bentham, Jeremy. "An introduction to the principles of morals." London: Athlone (1789a).

Bentham, Jeremy. "A utilitarian view." *Animal rights and human obligations* (1789b): 25-26.

Bernstein, Mark H. *On moral considerability: An essay on who morally matters*. Oxford University Press, 1998.

Bisol, Benedetta, Antonio Carnevale, Federica Lucivero. "Diritti umani, valori e nuove tecnologie. Il caso dell'etica della robotica in Europa." In *Metodo. International studies in phenomenology and philosophy* 2, no. 1 (2015): 235-252.

Borchers, Brian, and Judith Furman. "A two-phase exact algorithm for MAX-SAT and weighted MAX-SAT problems." In *Journal of Combinatorial Optimization* 2, no. 4 (1998): 299-306.

Bostrom, Nick, and Yuskowsky, : The ethics of artificial intelligence. In: Frankish, K., Ramsey, WM (eds.) The Cambridge Handbook of Artificial Intelligence. Cambridge University Press, Cambridge (2011)

Carmo, Alexandre Prusch Züge Renato, and Ricardo Tavares de Oliveira Fabiano Silva. "Using PMaxSat Techniques to Solve the Maximum Clique Problem." (2016).

Cassani Davis, Lauren "Would You Pull the Trolley Switch? Does it Matter? The lifespan of a thought experiment." In *The Atlantic* (2015) https://www.theatlantic.com/technology/archive/2015/10/trolley-problem-history-psychology-morality-driverless-cars/409732/ (Accessed October 31, 2019)

Cowls, Josh "AI and the 'Trolley Problem' Problem." In *Medium* (2017). https://medium.com/josh-cowls/ai-and-the-trolley-problem-problem-ef48582b49bf (Accessed October 31, 2019)

Crisp, Roger. *Reasons and the Good*. Oxford University Press on Demand, 2006.

Crockett, Molly "The trolley problem: would you kill one person to save many others? A decades-old thought experiment reveals our inconsistent moral intuitions. What would you do?" In *The Guardian* (2016). https://www.theguardian.com/science/head-quarters/2016/dec/12/the-trolley-problem-would-you-kill-one-person-to-save-many-others (Accessed October 31, 2019)

Cushman, Fiery, Liane Young, and Joshua D. Greene. "Our multi-system moral psychology: Towards a consensus view." *The Oxford handbook of moral psychology* (2010): 47-71.

De Givry, Simon, Javier Larrosa, Pedro Meseguer, and Thomas Schiex. "Solving Max-SAT as weighted CSP." In *International conference on principles and practice of constraint programming*, pp. 363-376. Springer, Berlin, Heidelberg, 2003.

D'Olimpio, Laura. "The trolley dilemma: would you kill one person to save five?" In *The Conversation* (2016) . http://theconversation.com/the-trolley-dilemma-would-you-kill-one-person-to-save-five-57111 (Accessed October 31, 2019)

Dworkin, Ronald. *A matter of principle*. OUP Oxford, 1985.

Eagleman, David. *The brain: The story of you*. Canongate Books, 2015.

Foot, Philippa. "The problem of abortion and the doctrine of double effect." (1967). In *Oxford Review* 5 (1967): 5-15.

Fu, Zhaohui, and Sharad Malik. "On solving the partial MAX-SAT problem." In *International Conference on Theory and Applications of Satisfiability Testing*, pp. 252-265. Springer, Berlin, Heidelberg, 2006.

Goodall, Noah J. "Ethical decision making during automated vehicle crashes." *Transportation Research Record* 2424, no. 1 (2014): 58-65.

Greene, Joshua D. "Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains." *Trends in cognitive sciences* 11, no. 8 (2007): 322-323.

Greene, Joshua D., Sylvia A. Morelli, Kelly Lowenberg, Leigh E. Nystrom, and Jonathan D. Cohen. "Cognitive load selectively interferes with utilitarian moral judgment." Cognition 107, no. 3 (2008): 1144-1154.

Greene, Joshua D., R. Brian Sommerville, Leigh E. Nystrom, John M. Darley, and Jonathan D. Cohen. "An fMRI investigation of emotional engagement in moral judgment." Science 293, no. 5537 (2001): 2105-2108.

Griffo, Cristine, João Paulo A. Almeida, and Giancarlo Guizzardi. "Conceptual Modeling of Legal Relations." In *International Conference on Conceptual Modeling*, pp. 169-183. Springer, Cham, 2018.

Griffo, Cristine, João Paulo A. Almeida, and Giancarlo Guizzardi. "A Systematic Mapping of the Literature on Legal Core Ontologies." In *Ontobras*. 2015a.

Griffo, Cristine, João Paulo A. Almeida, and Giancarlo Guizzardi. "Towards a legal core ontology based on Alexy's theory of fundamental rights." In *Multilingual Workshop on Artificial Intelligence and Law*, ICAIL. 2015b.

Guizzardi, Giancarlo, Guylerme Figueiredo, Maria M. Hedblom, and Geert Poels. "Ontology-Based Model Abstraction." In *IEEE 13th International Conference on Research Challenges in Information Science (RCIS 2019)*, Brussels, Belgium. 2019.

Guizzardi, Giancarlo, Gerd Wagner, João Paulo Andrade Almeida, and Renata SS Guizzardi. "Towards ontological foundations for conceptual modeling: The unified foundational ontology (UFO) story." *Applied ontology* 10, no. 3-4 (2015): 259-271.

Hale, Tom. "The Trolley Problem Has Been Tested In Real Life, And The Results Are Surprising." In *IFL Science* (2018) https://www.iflscience.com/brain/the-trolley-problem-has-been-tested-in-real-life-and-the-results-are-surprising/ (Accessed October 31, 2019)

Heras, Federico, and Javier Larrosa. "New inference rules for efficient Max-SAT solving." In *AAAI*, pp. 68-73. 2006.

Heras, Federico, Javier Larrosa, and Albert Oliveras. "MiniMaxSAT: An efficient weighted Max-SAT solver." In *Journal of Artificial Intelligence Research* 31 (2008): 1-32.

Heras, Federico, Javier Larrosa, and Albert Oliveras. "MiniMaxSat: A new weighted Max-SAT solver." In *International Conference on Theory and Applications of Satisfiability Testing*, pp. 41-55. Springer, Berlin, Heidelberg, 2007.

High-Level Expert Group on AI (AI HLEG). "Ethics Guidelines for Trustworthy AI." 2018. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419. (Accessed October 31, 2019)

Himmelreich, Johannes. "Never mind the trolley: The ethics of autonomous vehicles in mundane situations." *Ethical Theory and Moral Practice* 21, no. 3 (2018): 669-684.

Kahane, Guy. "On the wrong track: Process and content in moral psychology." *Mind & language* 27, no. 5 (2012): 519-545.

Kamm, Frances Myrna. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm.* Oxford University Press, 2008.

Kant, Immanuel. "Groundwork of the Metaphysics of Morals." Oxford University Press (1785/2002).

Keeling, Geoff. "Against Leben's Rawlsian collision algorithm for autonomous vehicles." In *3rd Conference on" Philosophy and Theory of Artificial Intelligence*, pp. 259-272. Springer, Cham, 2017.

Keeling, Geoff, Katherine Evans, Sarah M. Thornton, Giulio Mecacci, and Filippo Santoni de Sio. "Four perspectives on what matters for the ethics of automated vehicles." In *Automated Vehicles Symposium*, pp. 49-60. Springer, Cham, 2019.

Larus, James, Chris Hankin, Siri Granum Carson, Markus Christen, Silvia Crafa, Oliver Grau, Claude Kirchner, Bran Knowles, Andrew McGettrick, Damian Andrew Tamburri, and Hannes Werthner. *When Computers Decide: European Recommendations on Machine-Learned Automated Decision Making.* ACM, New York, 2018.

Leben, Derek. "A Rawlsian algorithm for autonomous vehicles." In *Ethics and Information Technology* 19, no. 2 (2017): 107-115.

Li, Chu Min, Felip Manya, and Jordi Planes. "New inference rules for Max-SAT." In *Journal of Artificial Intelligence Research* 30 (2007): 321-359.

Lin, Patrick. "Why ethics matters for autonomous cars." In *Autonomous driving*, pp. 69-85. Springer, Berlin, 2016.

MacDonald, Chris. "The business significance of the 'Trolley Problem': A test for ethical behaviour." In *Canadian Business* (2013). https://www.canadianbusiness.com/blogs-and-comment/the-business-significance-of-the-trolley-problem/ (Accessed October 31, 2019)

Maratea, Marco. "An Experimental Evaluation of Max-SAT and PB Solvers on Over-Subscription Planning Problems." In RCRA at CPAIOR. 2010.

Manquinho, Vasco, Ruben Martins, and Inês Lynce. "Improving unsatisfiability-based algorithms for boolean optimization." In *International conference on theory and applications of satisfiability testing*, pp. 181-193. Springer, Berlin, Heidelberg, 2010.

Manquinho, Vasco, Joao Marques-Silva, and Jordi Planes. "Algorithms for weighted boolean optimization." In *International conference on theory and applications of satisfiability testing*, pp. 495-508. Springer, Berlin, Heidelberg, 2009.

Marques-Silva, Joao, and Vasco Manquinho. "Towards more effective unsatisfiability-based maximum satisfiability algorithms." In *International Conference on Theory and Applications of Satisfiability Testing*, pp. 225-230. Springer, Berlin, Heidelberg, 2008.

Metzinger, Thomas. "Two principles for robot ethics." *Robotik und Gesetzgebung* (2013): 247-286.

Mill, John Stuart. *Representative government*. Kessinger Publishing, 1861.

Moro, Paolo. "Libertà del robot? Sull'etica delle macchine intelligenti." In *Filosofia del diritto e nuove tecnologie. Prospettive di ricerca tra teoria e pratic*a. Roma, Aracne, 2015, 525-544.

Nisbett, Richard E., and Timothy D. Wilson. "Telling more than we can know: Verbal reports on mental processes." *Psychological review* 84, no. 3 (1977a): 231.

Nisbett, Richard E., and Timothy D. Wilson. "The halo effect: evidence for unconscious alteration of judgments." Journal of personality and social psychology 35, no. 4 (1977b): 250.

Nyholm, Sven. "The ethics of crashes with self-driving cars: A roadmap, I." *Philosophy Compass* 13, no. 7 (2018): e12507.

Nyholm, Sven, and Jilles Smids. "The ethics of accident-algorithms for self-driving cars: An applied trolley problem?." *Ethical theory and moral practice* 19, no. 5 (2016): 1275-1289.

Paxton, Joseph M., and Joshua D. Greene. "Moral reasoning: Hints and allegations." *Topics in cognitive science* 2, no. 3 (2010): 511-527.

Pipatsrisawat, Knot, and Adnan Darwiche. "Clone: Solving weighted max-sat in a reduced search space." In *Australasian Joint Conference on Artificial Intelligence*, pp. 223-233. Springer, Berlin, Heidelberg, 2007.

Ramírez, Miquel, and Hector Geffner. "Structural relaxations by variable renaming and their compilation for solving MinCostSAT." In *International conference on principles and practice of constraint programming*, pp. 605-619. Springer, Berlin, Heidelberg, 2007.

Royzman, Edward B., Justin F. Landy, and Robert F. Leeman. "Are thoughtful people more utilitarian? CRT as a unique predictor of moral minimalism in the dilemmatic context." *Cognitive science* 39, no. 2 (2015): 325-352.

Smith, Jesse. "The Trolley Problem Isn't Theoretical Anymore." In *Towards Data Science* (2019). https://towardsdatascience.com/trolley-problem-isnt-theoretical-2fa92be4b050 (Accessed October 31, 2019)

Sommaggio, Paolo, and Samuela Marchiori. "Break the chains: a new way to consider machine's moral problems." *BioLaw Journal - Rivista di Biodiritto* 3 (2018): 241-257.

Thomson, Judith Jarvis. "A defense of abortion." In *Biomedical ethics and the law*, pp. 39-54. Springer, Boston, MA, 1976.

Thomson, Judith Jarvis. *The realm of rights*. Harvard University Press, 1990.

Thomson, Judith Jarvis. "Turning the trolley." *Philosophy & Public Affairs* 36, no. 4 (2008): 359-374.

Thomson, Judith. "Kamm on the Trolley Problems." Kamm, Francis M. *The Trolley Problem Mysteries*. Oxford: Oxford University (2016): 113-133.

Thornquist, Eline, and Anna Luise Kirkengen. "The quantified self: closing the gap between general knowledge and particular case?." *Journal of evaluation in clinical practice* 21, no. 3 (2015): 398-403.

Van de Poel, Ibo. "An ethical framework for evaluating experimental technology." *Science and engineering ethics* 22, no. 3 (2016): 667-686.

Wallach, Wendell. "Robot minds and human ethics: the need for a comprehensive model of moral decision making." *Ethics and Information Technology* 12, no. 3 (2010): 243-250.

Wallach, Wendell, and Colin Allen. Moral machines: Teaching robots right from wrong. Oxford University Press, 2008.

Wallach, Wendell, Stan Franklin, and Colin Allen. "A conceptual and computational model of moral decision making in human and artificial agents." *Topics in cognitive science* 2, no. 3 (2010): 454-485.

Zhao, Yingnan. *Do We Really Know Dworkin's 'One-Right-Answer' Thesis?*. Peking University Law School, 2018.