Optimality Theoretic Ethics

Abstract

In generative grammar it is common for cross-linguistic constraints on the language faculty to come into conflict with each other. In the Principles and Parameters Model, there exist principles which apply universally to language itself along with parameters (turned on or off) that generate the full range of the world's languages (Chomsky 6). However, in Optimality Theory, a constraint can be violated to satisfy a more important constraint. Instead of parameterization, Optimality Theory uses constraint ranking to determine the optimal surface form of a word or sentence for a given input (Legendre 2). Though theoretical linguistics and ethics may seem like disparate areas of study, there are many parallels between them, which support the idea that ethics can be analyzed within an Optimality Theoretic framework. This framework is a representative model of human moral decision making that serves to explain the underlying reasons for a decision or moral stance. In its current form, the model does not dictate whether one moral decision is objectively *better* than another.

1 Introduction

The Optimality Theory cognitive architecture has three main components - GEN, EVAL, and CON. GEN takes in the input and generates the set of possible outputs known as the candidates. CON provides the decision-making criteria in the form of a constraint ranking which is used to decide between candidates. EVAL chooses the optimal candidate based on the constraint ranking. Each of these cognitive mechanisms serves a purpose within Optimality-Theoretic Ethics. In linguistics, GEN populates the candidate set with possibilities from any of the world's languages. Thus, in Optimality-Theoretic Ethics, the candidates can be generated from any possible moral decision or moral stance. In linguistics, EVAL tries to maximize *markedness* which refers to the grammaticality of the word/sentence, within the given constraint ranking. Thus, in Optimality-Theoretic Ethics, EVAL attempts to optimize the *moral good*. In linguistics, CON is made up of constraints on the language faculty, but in Optimality-Theoretic Ethics, CON consists of constraints drawn from established moral theories such as Utilitarianism (Mill), Kantian

Ethics (Kant), and societal norms. Constraints can be derived from an applied ethics standpoint as well. I will appeal to other aspects of these theories over the course of this argument. The crucial parallel between the two theories, though, is that constraints *can* be violated. So here, moral claims are not considered right or wrong, but instead *most right* and *least wrong* under a given ranking. Take for example, abortion. In Optimality-Theoretic Ethics, the claim is not that abortion is right or wrong, but instead that there are two different constraint rankings which can generate both moral views.

1. SANCTITYOFLIFE >> AUTONOMY

2. AUTONOMY >> SANCTITYOFLIFE

Each moral question, then, must be viewed as an instance of *constraint conflict*, just as is the case within linguistic Optimality Theory. In addition to the similarities in the architecture of linguistic OT and OT Ethics, both language and values are acquired. The acquisition of values (right and wrong) may not be as seamless and remarkable as child language acquisition, yet it is clear that children generally acquire the values that their parents and society hold. This serves as another argument for the plausibility of OT Ethics. Finally, Optimality Theory is considered a meta-linguistic theory: it can apply to each aspect of the language faculty (phonology, syntax, and semantics). If this is true, it follows that the theory might govern even broader cognitive processes such as ethical decision making.

2 Previous Work

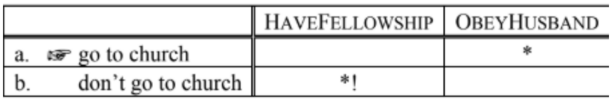
In their work, *Optimality Theory for Ethical Decision Making*, Steve and Monica Parker showed how one might go about applying OT to moral questions. Specifically, they examined the Bible, a trove of moral laws which are notorious for coming into conflict with one another. These two constraints form the basis of the constraint conflict below (Parker):

1. OBEYHUSBAND - Christian wives must submit to their husbands.

2. FELLOWSHIP - Christians must have fellowships with other believers.

These two obligations come into conflict in the case that a woman might have an *unbelieving* husband who commands her not to go to church. They resolve this conflict with the following

OT tableau: This example demonstrates that an Optimality-Theoretic representation of moral



	HAVEFELLOWSHIP	OBEYHUSBAND
a. ☞ go to church		*
b. don't go to church	*!	

Figure 1: An OT-Ethics Tableau

questions can in fact aid in adjudicating a conflict of constraints, yet the analogy between linguistic Optimality Theory and Optimality-Theoretic Ethics is not yet complete. In linguistic Optimality Theory, there are two types of constraints; *markedness* constraints are related to the well-formedness of the surface form, while *faithfulness* constraints are related to the *incongruity* of the input and the output. If the output is too distant from the input, it will violate a *faithfulness* constraint.

3 The Faithfulness Analogy

In order to fully explain *faithfulness* constraints, here I will first include an example from linguistics. Then, I will present examples in OT Ethics to elucidate the concept of faithfulness constraints.

3.1 Faithfulness in OT Syntax

The following is a technical example from syntactic theory highlighting the function of faithfulness constraints. The example comes from OT Syntax and is intended to further elucidate the concept of faithfulness. The example revolves around a linguistic phenomenon called the *expletive subject* that occurs in the English sentence, *It rained*, where *it* is an expletive subject (Legendre 4). Consider the following data from English and Italian:

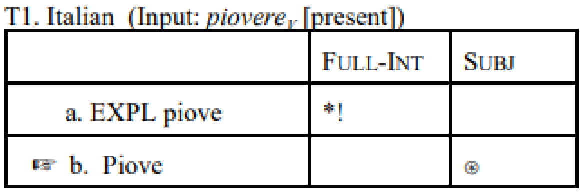
a It rained. (English)

b Piove. (Italian)

Both sentences here mean the same thing: *It is raining outside.*, however they differ in their surface form realization. There are two constraints that are active in this example.

1. FULLINTERPRETATION - lexical items must contribute to the interpretation of the structure, violated when there is an expletive

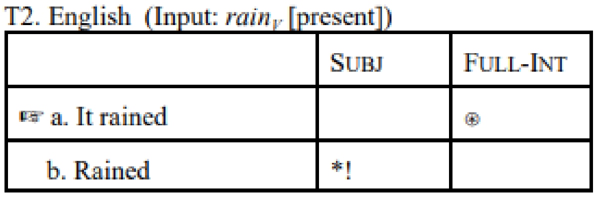
2. Subject - clauses have subjects as the highest specifier (aka: a sentence must have a subject), violated when there is no subject in an clause



T1. Italian (Input: *piovere*$_V$ [present])

	Full-Int	Subj
a. EXPL piove	*!	
☞ b. Piove		⊛

Figure 2: English



T2. English (Input: *rain*$_V$ [present])

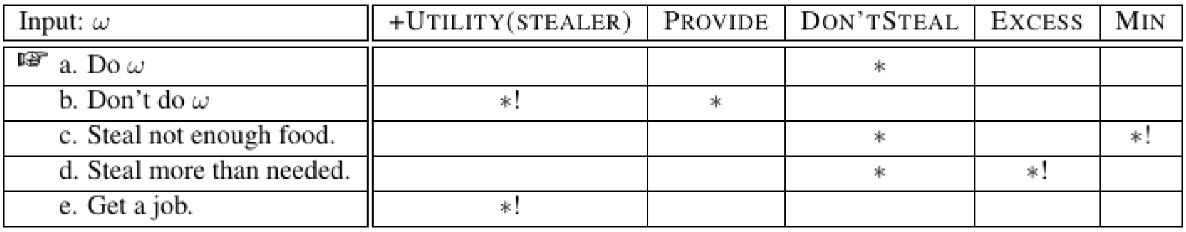
	Subj	Full-Int
☞ a. It rained		⊛
b. Rained	*!	

Figure 3: Italian

The first constraint is a faithfulness constraint while the second is a markedness constraint. The constraints conflict in the case of a *weather verb* which doesn't select a thematic argument. Either the surface form violates Subject or it violates FullInterpretation. Thus, we can generate two tableau, one for English and one for Italian. The ranking for English shows that Subject must higher than FullInterpretation - this means that the language would rather include a subject in a clause than include a lexical item which contributes no content to the sentence (it as an expletive subject). Under re-ranking, the Italian grammar is easily generated demonstrating a language *typology*. Crucially, the faithfulness constraint mandates that the input is realized in a *particular way*. It is this effect that I will attempt to replicate in OT Ethics.

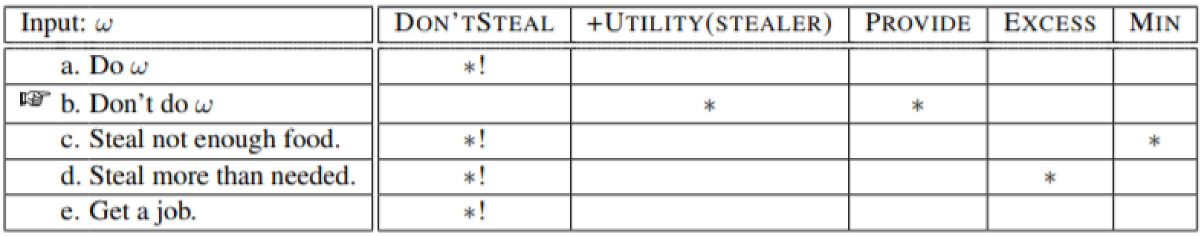
3.2 Form of the Input

Before presenting examples of OT Ethics faithfulness constraints, the form of the input must be defined. There are many ways to specify the input in OT-Ethics, yet I will posit for sake of an appeal to authority and convention that it would be reasonable to adopt the form of the maxim, put forth by Immanuel Kant. The following is a Kantian maxim for a simple moral choice: *I will* ***steal food*** *in order to* ***feed my family*** *in the circumstances that* ***I am too poor to afford it and my family is hungry.*** Under this form of the Input, the decision space is well defined so that there is less ambiguity in the adjudication process.



Input: ω	+UTILITY(STEALER)	PROVIDE	DON'TSTEAL	EXCESS	MIN
☞ a. Do ω			$*$		
b. Don't do ω	$*!$	$*$			
c. Steal not enough food.			$*$		$*!$
d. Steal more than needed.			$*$	$*!$	
e. Get a job.	$*!$				

Figure 4: The ranking which selects candidate a.



Input: ω	DON'TSTEAL	+UTILITY(STEALER)	PROVIDE	EXCESS	MIN
a. Do ω	$*!$				
☞ b. Don't do ω		$*$	$*$		
c. Steal not enough food.	$*!$				$*$
d. Steal more than needed.	$*!$			$*$	
e. Get a job.	$*!$				

Figure 5: The ranking which selects candidate b.

3.3 Degree

This example will generate a tableau for the example given above. One concern in the theory is the generation of potential candidate outputs. One might posit that there is an *infinite* number of possible candidates. However, the faithfulness constraints eliminate much of the candidate space. The constraints are the following:

1. DEGREE(EXCESSIVE) - violated by excessive action

2. DEGREE(MINIMAL) - violated by a insufficient action

3. +UTILITY(STEALER) - violated when the stealers utility decreases

4. DONOTSTEAL - violated when the action includes stealing

5. PROVIDE(FAMILY) - violated when the agent does not provide for their family

Based on these constraints, then, it will be possible to generate two constraint rankings - one in which the right course of action is to steal, and one in which it is to not steal.

ω = *I will **steal food** in order to **feed my family** in the circumstances that **I am too poor to afford it.***

In the Figure 4, the candidates c. and d. violate the faithfulness constraints, but it must be kept in mind that EXCESS and MINIMAL have already eliminated other candidates like killing or robbing a bank that are either not relevant or not appropriate - in many ways, in OT-Ethics, the faithfulness constraint represents the degree of response of the candidate action/decision in relation to the input. In Figure 4 they serve as fatal violations for c. and d. since a. only has one violation. Now under re-ranking, it is possible to select a different optimal output demonstrating that OT Ethics is able to represent a moral typology for the given example. Under the second ranking, we are able to select for candidate b. instead of a. and thus we have shown that the difference in the two candidates is not a question of moral good, but rather, a question of constraint ranking.

3.4 Specificity

In order to demonstrate the SPECIFICITY faithfulness constraint, I will consider the case of modest dressing laws. Here, it is not possible to justify the MODESTY constraint as a societal norm. Whether the MODESTY constraint is a moral duty or even a moral virtue will not be relevant here as OT Ethics does not seek to find the objectively ”best” ranking - instead it defines a moral typology. The level of specificity of the input, has a role in the realization of the constraints, and the output. The MODESTY constraint must be realized as a spectrum - not all moral claims can be represented through ranking so one must posit that MODESTY as a constraint has a feature called GRADIENT which requires the instantiation of graded spectrum to go along with the tableau. Here I will model the moral typology of *modesty* in the United States, Saudi Arabia, and Iran. The MODESTY constraint is realized as a graded spectrum where each gradation on the line is realized numerically as a modesty value- thus, invoking a harmonic grammar inspired approach here (Smolensky). The constraints are as follows:

1. AUTONOMY - violated when an agent’s personal freedom is violated
2. MODESTY with the feature GRADIENT - violated when a modesty law is violated
3. SPECIFICITY - violated when a candidate has a different level of specificity than the input

Typically a maxim is formulated as a proposed course of action from an agent-based perspective, however, that doesn’t quite make sense in this scenario, thus the principle will be demonstrated on a broader scale in an attempt to represent the moral stance of a *society*.

Input: ω	SPECIFICITY	MODESTY	AUTONOMY
a. Wear no head covering		*! (1)	
b. Wear the hijab		*! (2)	*
c. Wear the niqab		*! (3)	*
☞ d. Wear the burka			*
e. Dress modestly	*!	?	?

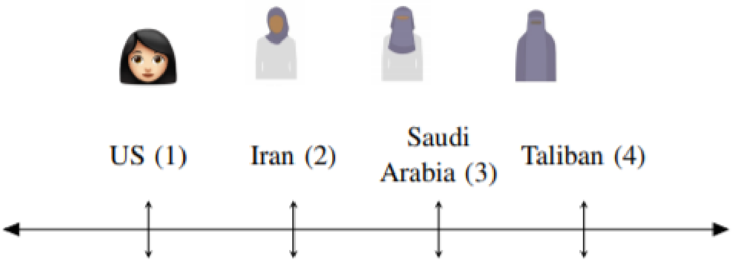


Figure 6: A tableau containing the SPECIFICITY constraint and GRADIENT.

ω = *A woman should **dress modestly** in order to **please God**, in particular, by covering her **head and face.***

Here, we see that since the INPUT specifies that the head *and* face ought to be covered, any candidate to the left of *Taliban* on the modesty gradient will result in a violation of the MODESTY constraint, which in this case must be greater than or equal to 4. Since candidate e. lacks specificity, it is not clear whether or not the MODESTY or AUTONOMY constraints are violated, though due to the ranking, this candidate is eliminated by the faithfulness constraint on SPECIFICITY. Here the broad typology of MODESTY is represented; the particular moral view represented here is that of the Taliban, a more traditional, conservative, Islamic regime. It is clear then, that a different reordering of the MODESTY and AUTONOMY constraints can generate the various moral typologies for modesty. This examples has also demonstrated the SPECIFICITY constraint and GRADIENT feature.

4 Summary

Optimality Theory is a useful framework for representing ethical questions in which various constraints come into conflict. The framework doubly serves as a representational tool for conceiving of various moral typologies around the world and throughout history as simply a reordering of constraints. Additionally, I hope to have uncovered a potentially valid, but somewhat tenuous, analogy from the concepts of *markedness* and *faithfulness* in linguistic theory to those of OT Ethics. Specifically, within the faithfulness category, I employed the use of the constraints, MINIMAL, EXCESS, SPECIFICITY, and the feature known as GRADIENT. The constraints, EXCESS and MINIMAL can be reduced to one constraint called DEGREE.

5 Evidence from Compatibilism

In Daniel Dennett's paper, *On giving libertarians what they say they want*, a compatibilist account of decision making is given (Dennett 295). Dennett claims that first, ”a consideration-generator whose output is to some degree undetermined produces a series of considerations, some of which are immediately rejected as irrelevant by the agent (consciously or unconsciously)”. The *consideration-generator* that Dennett posits is analogous to the GEN component of the optimality-theoretic cognitive architecture. Then, he claims, ”Those considerations ... [with] a more than negligible bearing on the decision...ultimately serve as predictors and explicators of the agent's final decision.” Interestingly, this description closely tracks with the account of OT-Ethics given above by which the interaction of CON and EVAL determine the optimal moral decision. Thus, the candidates are generated both consciously and subconsciously by GEN. Candidates that violate SPECIFICITY and/or a DEGREE constraint may also be subconsciously eliminated. Finally, Dennett adds that, ”the model ...permits moral education to make a difference, without making all the difference”, indicating that there is some level of flexibility in the CON module. Moral education is certainly proof of this, however there is also evidence from moral psychology that CON is amenable to reranking based on external factors.

5.1 Moral Licensing

Though its effects may be exaggerated (Kuper), *moral licenscing* is an important psychological phenomenon by which public morally good behavior can license subsequent morally bad behavior in individuals. To explain this phenomenon within an OT-Ethics framework, one must posit some degree of constraint flexibility that would allow the +UTILITY(SELF) constraint (self-interest) to *float* to the top of the ranking as the result of an individual's belief that they have ”done their due” and thus are licensed to behave badly.

6 Applications

The Optimality Theoretic Ethics framework can be applied to the field of AI Safety. In this area of research, there is a much discussed problem which raises concerns about how humanity can reliably impart its values to a future artificial intelligence/superintelligence (Bostrom). This is known as the *value-alignment problem*. It has become rather pertinent, perhaps sooner than technologists might have anticipated, due to the advancement of narrow AI machines such as the self driving car. Here, I'll take preliminary steps toward demonstrating that Optimality-Theoretic Ethics framework could be used to encode humanity's values into an artificially intelligent system (once these values had been decided upon by humanity). Specifically, I will examine the increasingly relevant problem of self driving cars and the Trolley Problem.

6.1 Value Acquisition

In the Optimality-Theoretic view of language acquisition, children are born with an innate language faculty that includes mechanisms such as GEN, to generate the candidate space, and EVAL, to select the winning candidate based on the constraint ranking, as well as a built in representation of constraints on the language faculty. Under this view, language acquisition is reduced to a question of using stimuli to determine the correct constraint ranking for a given language. Now, this premise can be extended to value acquisition by a seed AI in Optimality-Theoretic Ethics. As a seed AI moves through its virtual world, it observes decisions and then based on this stimulus is able to learn the optimal constraint ranking. Of course, in this conception of the value acquisition, the seed AI may only learn the values of the society whose decisions it is exposed to. Interestingly, this means that a seed AI will also be able to make mistakes - if an AI makes a decision that is considered by humans to be non-optimal, we would be able to label it as such and thus give the AI a chance to re-rank its *innate* constraints. This mirrors the human ability to learn from mistakes. This analysis further supports the original *applicability* claim (that OT applies to ethics), but also lends credence to the idea that OT Ethics is applicable to AI safety.

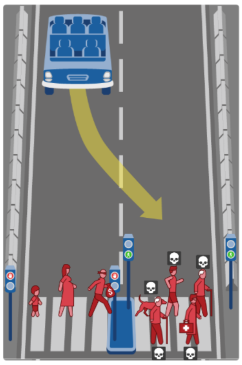


Figure 7: A scenario from MIT's Moral Machine.

6.2 Moral Machine

In response to advances in self-driving car technology, MIT has created a thought-provoking version of the trolley problem to help assess societal attitudes toward self driving cars - I will use just one of their thirteen examples to further explain the OT-Ethics framework. In this example (Awad), on the left side of the street are the following agents : *a pregnant woman, a criminal, and a boy.* On the right side of the street are: *two elderly men, a dog, a male doctor, a male athlete*. Thus, the constraints are the following:

1. GENDER - a preference toward protecting women from danger

2. AGE - a preference toward protecting children from danger

3. ΣUTILITY with the feature, GRADIENT - violated when there is a decrease in the net utility to society

4. JUSTICE - violated when agents are not held to account

5. SELF-PRESERVATION - violated when an agent does not preserve itself

It is important to note that the gradient refers to only the ΣUTILITY constraint and attempts to classify the amount of utility that each type of agent might provide to society as a whole - this is a rather cynical point of view to take; to argue that a doctor provides more utility to society than a old man or child, to most would seem to be quite a detached and harsh moral judgment, yet this is what utilitarianism demands. However, this is why the constraints AGE and GENDER which have been persistent values throughout human history are included. They represent the sentiment that within a disaster scenario, women and children should be given

Input: ω	SELF-PRESERVATION	AGE	GENDER	ΣUTILITY	JUSTICE
☞ a. Swerve toward right side				$*$ (-17)	$*$
b. Continue straight		$*$!	$*$	$*$ (-8.5)	
c. Crash into median (∅)	!$*$				$*$

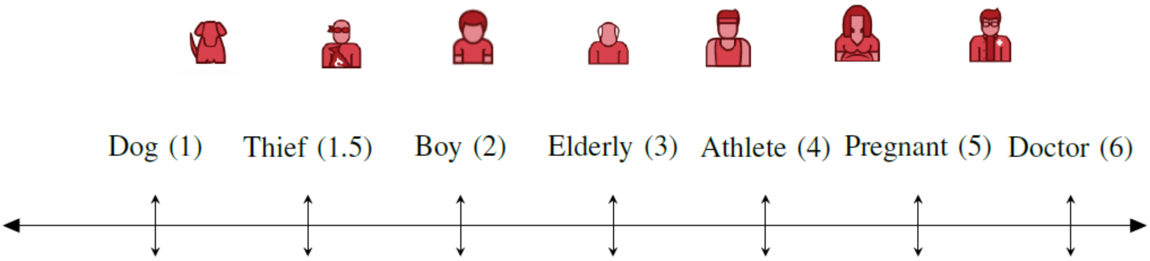


Figure 8: A tableau for the example given above with a GRADIENT.

preferential treatment and aid. After all, an optimal solution to the value alignment problem is one that succeeds in imparting human values onto a machine - preferential treatment to women and children is a rather uncontroversial human value. Additionally, the existence of these constraints is supported by Peter Singer's distinction between *preferences* and *pleasures* by which the constraints of AGE and GENDER would be considered preferences and not pleasures. This distinction is largely used as a *go-around* for the rigidness of utilitarianism - take Peter Singer's example: A fish struggles to free itself from a barbed hook in its mouth, despite causing more pain to itself because it has a preference for doing so (in the face of a decrease in utility) (Singer 95). In this case, SELF-PRESERVATION is used to eliminate the NULL-PARSE candidate (the do nothing option), but it certainly might be possible to re-rank this constraint at the end for an empty self-driving car. Finally, the JUSTICE constraint represents the idea that those that have broken the law, inherently have less value, and this constraint is ranked last. It is violated by the agents in the leftmost lane since they are crossing illegally and since one of them is a thief. States in the US and countries around the world have different ideas of how they believe self-driving cars should be regulated and how they should behave in a trolley-problem scenario. The above OT Ethics tableau allows for these varying ideals to be represented from the same finite set of constraints by tinkering with the constraint ranking.

7 Conclusion

Having expanded on work from Parker and Parker by extending the Optimality-Theoretic Ethics framework to the linguistic concept of *faithfulness* I hope to have further supported the idea that

Optimality Theory can apply to a non-linguistic discipline. In this view, Optimality Theory is a cognitive architecture that governs not only the language faculty, but other parts of the mind as well: in this case, values and decisions. Evidence from the compatibalist view of free-will and from social psychology also lend credence to this model along with the many parallels with the original linguistic theory, including the idea of *acquisition*. The framework then allows for one to conceive of a future AI (either narrow, or superintelligent) whose actions could be modeled by OT Ethics - just as the human language faculty is modeled by OT.

8 Acknowledgments

Much credit is due to Amalia Gnanadesikan for suggesting several of the examples used above to me by email. She was also (as far as I know) the progenitor of the idea itself. Steve and Monica Parker also helped greatly in reviewing early drafts of this paper. Their paper, *Optimality Theory for Ethical Decision Making* which explored this idea from a Biblical perspective, provoked much thought on this topic.

9 References

[1] Chomsky, Noam. The Minimalist Program. 20th ed., MIT Press, 2015. JSTOR, www.jstor.org/stable/j.ctt17kk8xd. Accessed 24 Jan. 2020.

[2] Legendre, Géraldine. (2001). An introduction to optimality theory in syntax. Optimality-Theoretic Syntax.

[3] Mill, J. (2014). Utilitarianism (Cambridge Library Collection - Philosophy). Cambridge: Cambridge University Press.

[4] Kant, Immanuel, and H J. Paton. Groundwork of the Metaphysic of Morals. 1964. Print.

[5] Parker, S. and Parker, M., 'Optimality Theory and Ethical Decision Making', in Work Papers of the Summer Institute of Linguistics, University of North Dakota Session, vol. 48 (2004),

[6] Paul Smolensky and Geraldine Legendre. The harmonic mind: From neural computation to Optimality-Theoretic grammar. 2 vols. Cambridge, Mass.: MIT Press, 2006.

[7] Dennett, Daniel C. (1978). On giving libertarians what they say they want. In Brainstorms. MIT Press

[8] Kuper, Niclas, and Antonia Bott. "Has the Evidence for Moral Licensing Been Inflated by Publication Bias?." PsyArXiv, 16 May 2018. Web.

[9] Bostrom, Nick. ”Superintelligence: Paths, Dangers, Strategies, Reprint ed.” (2016).

[10] Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J.D., Shariff, A., Bonnefon, J., Rahwan, I. (2018). The Moral Machine experiment. Nature, 563, 59-64.

[11] Singer, P. (2011). Practical Ethics. Cambridge: Cambridge University Press.