A non-monotonic extension of Universal Moral Grammar Theory

Munneke, G.-J.; Szymanik, J.

Publication date
2015
Document Version
Final published version
Published in
COGSCI 2015 : 37th Annual Meeting of the Cognitive Science Society

A Non-monotonic Extension of Universal Moral Grammar Theory

Gert-Jan Munneke (G.J.Munneke@UvA.nl)
Institute for Logic, Language and Computation
University of Amsterdam
107 Science Park, Amsterdam, 1090 GE
The Netherlands
Jakub Szymanik (J.K.Szymanik@UvA.nl)
Institute for Logic, Language and Computation
University of Amsterdam
107 Science Park, Amsterdam, 1090 GE
The Netherlands

Abstract

We extend universal moral grammar theory (UMGT) with non-monotonic logic. Our experiment shows that such revision is necessary as it allows to account for the effects of alleviations and aggravations in moral reasoning. Our new theory updates UMGT from classical to non-monotonic logic, which reflects the incompleteness of information and uncertainty in actual human reasoning. In addition, it provides an explanation of the paradoxical findings in the moral dilemma of the Trolley problem and the Knobe effect.

Keywords: moral psychology; defeasible reasoning; universal moral grammar; non-monotonic logic

Introduction

Even though morality is such a fundamental feature of man and of great importance for societal well-being, little is still known about the moral mind. What are the principles that govern moral cognition? And what are the origins of our moral sense of right and wrong? Although these questions remain unanswered, a promising theoretical framework on how to approach these problems has recently been offered by *universal moral grammar theory* (Mikhail, 2007).

In this paper, we extend universal moral grammar theory with non-monotonic logic. This update from classical to modern logic allows the modeling of incompleteness of information and uncertainty in actual human reasoning. Our experiment shows that such revision is necessary as it allows to account for the effects of alleviations and aggravations in moral reasoning. Our new theoretical paradigm also offers an explanation of the paradoxical findings in the moral dilemma of the Trolley problem (Foot, 1967) and the Knobe effect (Knobe, 2003).

Universal moral grammar theory (UMGT)

Universal moral grammar theory is a paradigm for studying moral cognition that borrows concepts from Chomskian linguistics (Chomsky, 1969, 2002). It was proposed by Mikhail (2007), who has developed UMGT from an analogy between the study of morality and language that was put forward by Rawls (1999).

This analogy, in short, states that our moral faculty, just like our language faculty, allows for fast reflexive judgments on either whether a sentence is grammatical or whether a situation is deemed (im)moral. These systems mature relatively quickly in young children without the need for extensive training initiated by the social environment, which suggests that these faculties are, at least in part, an innate feature of human nature Mikhail (2011).

Many methodological assumptions behind UMG could be met with rightful skepticism but a full discussion is beyond the scope of this paper. What matters for our purpose is the modeling aspect of UMG which is aimed at capturing the different computational stages through which a person generates a moral judgment.

A sequential model of moral judgment

We will explain Mikhail's model of moral judgment with the example that he himself uses to explain the paradigm, namely the moral dilemma of the Trolley problem and a paradoxical dissociation between the way people respond to two of its variants (Thomson, 1985; Foot, 1967; Mikhail, 2007).

The trolley problem: Switch variant A train/trolley is about to hit and kill five people that are standing on the tracks. The only alternative is that a bystander pulls a switch which diverts the train unto a side-track, thus saving the 5 people. The caveat is that there is 1 person on that side-track that will be hit and killed as a side-effect. Is it permissible to pull the switch? Ninety percent of people say "yes" (Mikhail, 2007).

The trolley problem: Fat man variant A train/trolley is about to hit and kill five people that are standing on the tracks. The only alternative is that a bystander on a footbridge over the tracks shoves a fat man standing next to him off the bridge and in front of the train. This man will be hit and killed, but

will stop the train as well and thus result in saving the 5 men. Is it permissible for the bystander to shove the fat man in front of the train? Now 90 percent of the people say "no" (Mikhail, 2011).

Why does this flip in the moral judgment occur? The outcome in both variants is the same so a straightforward utilitarian explanation is problematic.

According to Mikhail's framework of UMGT, the reason behind this moral dissociation lies in the sequence of consecutive stages in which a moral judgment is constructed (Mikhail, 2011). The first stage exists in generating the temporal structure of the moral situation which states the order in which the atomic events of the situation occur. From this temporal structure a causal structure is obtained which specifies who/what does what to whom at which time-point with what result. This causal structure is conversed into a moral structure by determining which effects are considered good, bad or neutral. This structure is expanded upon into an intentional structure which determines which effects are directly intended and which are interpreted as foreseen but unwanted side-effects. The final stage is the deontic structure which is a logical derivation system that operates on classical logic, including the material implication (Mikhail, 2007). This system receives the results from the intentional structure as an input in the form of logical formulas and together with general world knowledge, this system allows the moral mind to derive whether an action is morally permissible or not.

The moral dissociation between the trolley problem variants then lies, according to Mikhail (2007), primarily in the difference in the intentional structure of the actions with bad consequences, but also in the "badness" of the initial action. Pulling the switch is a neutral, causal and directly intended means towards the good end of saving lives with the 1 death being an unintended side-effect. In contrast, shoving a person from a bridge involves the bad action of battery as a means towards a bad causal end of letting a person get killed by a train which itself is a bad means to the eventually positive end of saving lives.

The moral judgment of an action is formalized by Mikhail (2011) in the form of a logical equivalence, which is defined as classical material equivalence:

$$D(A) \leftrightarrow A(F_1, \ldots, F_n)$$

This formula states that an action A has deontic status D if and only if action A has features $F_1, \ldots, F_n$.The deontic status is a judgment like permissible or forbidden. The action is further specified as [S's V-ing at $t^{(\alpha)}]^c$ which means that a subject S performs a verb V at time-point t under circumstances c.

Non-monotonicity of deontic rules

Although the inference rules incorporate the possibility to take circumstances into account, we witness that there is no flexible way of incorporating contextual pragmatics detached from the action, nor of a flexible adaptation of derived beliefs in the deontic structure as new information becomes available. We propose that the pragmatic context of a moral situation plays a key role in the logical form of the inference rules as a whole. The way we update the system is by using default logic (Reiter, 1980; Berzati, 2007; Brewka et al., 1997). A classic example of an inference in default logic is: if Tweety is a bird, then Tweety can fly (Reiter, 1980). This inference is probably correct, unless Tweety turns out to be a penguin. The casual reasoner will assume by default that Tweety is not a penguin, or any other atypical bird that does not fly. The absence of evidence for such an abnormal bird is considered to be evidence of absence of such an abnormality.

These default inferences are also part of human reasoning in general (Stenning & Van Lambalgen, 2008), and of moral reasoning in particular (Horty, 1997). For example, if a person kills, then this action is forbidden. There are however exceptions to this rule, like self-defense. These circumstantial factors can alter the moral judgment of the entire situation, for the better but also for the worse. These excuses (alleviations) and aggravations are a key ingredient of moral reasoning. They are by default assumed to be absent, unless positive evidence in favor of their existence is available (Horty, 1997). Updating the inference rules from standard UMGT with this default reasoning gives us the following formalization:

$$D(A) \leftrightarrow A(F_1, \ldots, F_n) \wedge \neg ab$$

In which we have added the negation of abnormality, ab, which is required for action A to have deontic status D. More specifically we state that:

$$Bad(A) \wedge \neg alleviation \leftrightarrow Impermissible(A)$$

$$Good(A) \wedge \neg aggravations \leftrightarrow Permissible(A)$$

This means that the moral judgment of an action does not only depend on the goodness or badness of the action in and of itself but also of the contextual factors that surround it. A bad action can be excused and a good action can be nullified if it was, for instance, performed for the wrong reasons or with bad intentions. An interesting result is that new information on the existence of such abnormalities can invalidate moral inferences that used to be valid at an earlier step in the derivation when there was no evidence for such an abnormality.

For example, upon hearing that a burglar stole goods from a local pharmacy, we tend to judge this action as immoral. But when we later on learn that the man stole an expensive medicine from a pharmaceutic company because it was the only way for him to save the life of his wife, we tend to revise our initial judgment and some would even state that this man is now a hero (Kohlberg, 1981).

This new framework thus violates the property of monotonicity of classical logic under which standard UMGT operates and this new approach updates UMGT into a non-classical and flexible non-monotonic logic which fits better

with the actual dynamics and limitations of human reasoning (see Stenning & Van Lambalgen, 2008; Cummins, 1995; Nyamsuren & Taatgen, 2014).

This extension can incorporate the explanation of Mikhail for the moral dissociation in the Trolley problems by stating that the lack of direct intention can be interpreted as an alleviation. More interestingly, this new framework can also explain a finding in moral psychology that as of yet remained paradoxical, which is the Knobe effect (Knobe, 2003).

The Knobe effect The Knobe effect is the tendency of people to assign intentionality to a protagonist that commits an immoral act, even though the protagonist did strictly not intend the negative side-effect of his action. This tendency is absent when the side-effect is positive. Here is the original text from Knobe (2003):

> The vice-president of a company went to the chairman of the board and said, We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment. The chairman of the board answered, I dont care at all about harming the environment. I just want to make as much profit as I can. Lets start the new program. They started the new program. Sure enough, the environment was harmed.

Explaining this effect has received much attention but finding a robust explanation of the reasons behind this anomalous finding has proved to be difficult (Nichols & Ulatowski, 2007).

Our non-monotonic extension of UMG provides a hypothetical explanation. People assign intentionality because the lack of intentionality could be interpreted as an alleviation. People however insist that the action is immoral and therefor do not accept the lack of intention as a potential excuse, even though their statement is strictly in conflict with the factual state of the world. Moral cognition thus initiates the need for subjectively framing the situation in terms of alleviations and aggravations, even if it means that facts about the world have to be suppressed. This would also explain why the effect is absent when the side-effect is positive because in that case people do not need to protect there line of reasoning towards the moral judgment of "forbidden" as the action is now clearly permissible.

The standard and updated versions of UMGT give rise to some potential differences in their hypothetical explanations concerning the Knobe effect:

1. Standard UMG framework could explain the Knobe effect by referring to one of it's core principles, i.e. the principle of double effect (Quinn, 1989). This principle, from Catholic theology (Aquinas & Hutchins, 1952), posits that when a negative consequence of an action is a means to an end or the end in itself it is said to be *directly intended* by the agent, whereas an unwanted but foreseen side-effect is deemed to be *obliquely intended*. As the negative consequence in the Knobe effect seems to be a negative side-effect, people could be intuitively interpreting the question on intentionality as "obliquely intended" and thus their answer actually would be in line with the factual state of the world, albeit in a non-straightforward interpretation of the word intention. It remains unknown however how such an account would explain the lack of the Knobe effect in case of a positive side-effect.

   In contrast, the non monotonic-extension can explain the Knobe effect, as stated earlier, from the need to suppress an alleviation.

2. Furthermore, there is evidence that religious people are more prone to deem a morally questionable situation as impermissible (Shariff & Norenzayan, 2011). This increased tendency to judge actions as immoral should, according to our extension, increase the need to suppress any alleviations and thus lead to an increase in the assigned intentionality.

From these hypotheses we derive the following predictions:

1. Our theory predicts that there is a correlation between the assigned intentionality and the immorality rating assigned to the action of the chairman because the need to assign intentionality would rise in case a perpetrator would seek an alleviation, which is when his action was bad in and of itself. We also predict that when the negative side-effect is a lesser evil, that the assigned intentionality would drop as well.

2. If religious people are more likely to score the Knobe scenario as more immoral, then they should also assign more intentionality to the protagonist.

These predictions have been tested in an experimental vignette study by varying the severity of the negative side-effect (destroying one tree vs severely hurting the environment). As an exploratory effort we also varied whether the kind of agent (a loving father or a CEO) and whether the agent cared or not about the negative side-effect. In the standard scenario, the protagonist does "not care at all" about the negative side effect, which could be seen as an aggravation.

Experiment

Methods

Participants Two-hundred and forty-one US-residents with ages between 19 and 67 ($M = 32$, $SD = 10$), of which 144 males and 97 females; 73 were religious and 168 were not religious. The participants were M-Turk workers with an approval rating of 95% or higher.

Materials The stimuli consisted of a vignette in which the protagonist of the story performed an action which had a positive main effect and a negative side effect. We varied whether a) the action is deemed permissible due to the positive outcome outweighing the negative outcome, b) whether the protagonist was a CEO or a "loving father" and c) whether the agent cared about the negative outcome or not. See table

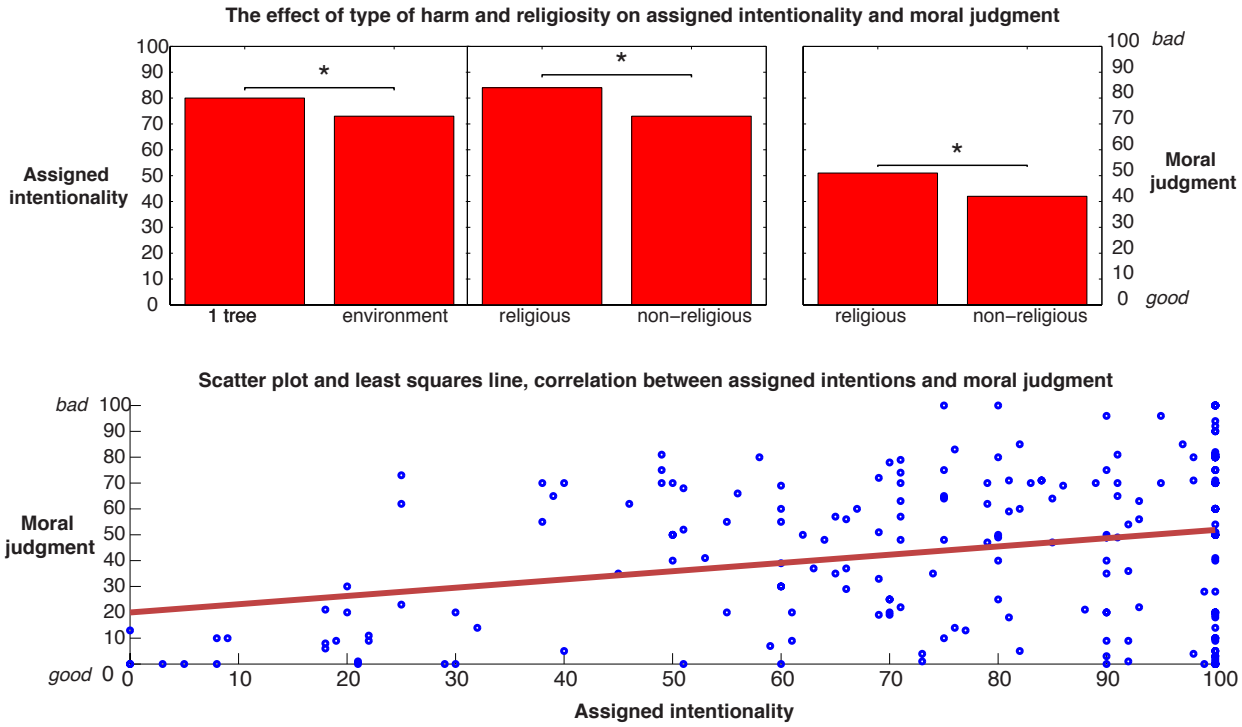

Figure 1: Main results. Stars depict statistical significant at $\alpha = .05$.

1 for the resulting vignette structure.

The 3 dichotomous independent variables resulted in 8 different stories which were administered in a between-subjects design.

Participants scored the degree to which the protagonist intended the negative side effect with a visual analogue scale ranging from 0 ("no intention at all") to 100 ("He fully intended the action"). Participants also indicated their moral acceptance of the action with a visual analogue scale ranging from 0 ("Neutral") to 100 ("Completely immoral, He is like a murderer"). The extreme labels were chosen in order to prevent ceiling effects and facilitate normally distributed data.

Control questions asked in a multiple-choice fashion whether the protagonist was a father or a CEO, whether harm consisted in destroying a tree or hurting the environment, and to simply answer 'yes' to a specific question.

An exit interview obtained information on country of residence, age, gender, religiosity (yes/no) and whether English is the native language.

Procedure The entire questionnaire lasted 2 minutes for which participants received $0.25.

Results and discussion

We excluded participants that did not pass the control questions ($N = 12$). The results are depicted in Figure 1.

There is an effect of type of harm on intentionality ($t(209) = -2.2$, $p = 0.03219$, $r = .15$). The lesser amount of harm ("destroying 1 tree") lead to a lower intentionality rating ($M = 73$, $SD = 31$), than the condition with a higher amount of harm ("severely destroying the environment") ($M = 80$, $SD = 22$).

There is correlation between the immorality score and the assigned intentionality ($r = 0.28$, $p < 0.001$, $r^2 = .078$).

There is an effect of religiosity on the moral judgment ($t(133) = -2.1$, $p = 0.04224$, $r = .17$). Religious people have a higher immorality rating ($M = 51$, $SD = 30$), than non-religious people ($M = 42$, $SD = 31$).

There is an effect of religiosity on intentionality ($t(185) = -3.4$, $p = 0.00082$, $r = .24$). Religious people have a higher intentionality rating ($M = 84$, $SD = 20$), than non-religious people ($M = 73$, $SD = 29$).

Both predictions are confirmed. There is a relation between the assignment of intentionality and the degree to which a bad action is deemed immoral. Furthermore we see that the stronger tendency to assign intentions in religious people is related to an overall stronger tendency to find a bad action immoral.

We did not observe any effect of the protagonist being a CEO or a loving father, nor did we find an effect of whether the protagonist cared or not about the negative side-effect of his decision. Apparently these manipulations are not perceived as salient aggravating or alleviating factors.

Protagonist	
Loving father	CEO of a big corporation
A loving father receives news of a friend that there is an opportunity to start a new program.	The CEO of a big company receives news of a friend that there is an opportunity to start a new program.
Kind of harm	
Destroying 1 tree	Severely hurting the environment
This program will make a lot of money but will also have a negative side effect of destroying 1 tree.	This program will make a lot of money but will also have a negative side effect of severely hurting the environment.
Care of the protagonist	
Cared	Did not care
The [CEO/loving father] cared about the damage but went ahead with the program.	The [CEO/loving father] did not care about the damage and went ahead with the program.

Table 1: Vignette structure of 8 variants on the Knobe scenario.

General discussion

We extended universal moral grammar theory (UMGT) with non-monotonic logic. We did so by replacing the classical material implication in the deontic rules with the implication according to default logic. This non-monotonic logic allowed us to model important contextual factors that influence moral reasoning. Acknowledging these factors—aggravations and alleviations—made it possible to re-frame Mikhail's explanation of the moral dissociation in the trolley problem. It furthermore allowed us to explain the paradoxical finding of the Knobe effect.

Our experiment on the Knobe effect showed that people indeed assign intentionality more strongly when there is a need to counter a potential excuse/alleviation. When an action is considered less deplorable, then the need for countering such an alleviation diminishes and the Knobe effect is attenuated. Furthermore we witnessed that religious people assign intentions more strongly; and as predicted, this was related to a stronger moral dismissal of immoral acts. Although this extension can explain existing paradoxical findings and provide novel predictions, some points do deserve some critical examination.

A non-theoretical issue concerns the validity of M-Turk workers as a research tool. Although the US M-Turk workers are not completely representative of the general US public, as is evidenced by the relatively low proportion of religious people, M-Turk workers are more diverse than standard laboratory samples (Buhrmester et al., 2011; Paolacci et al., 2010; Krantz & Dalal, 2000). M-Turk workers yield similar responses than typical samples (Buhrmester et al., 2011). Nonetheless, M-Turk workers can have a lack of attention (Gosling et al., 2004), for which we corrected by excluding workers that could not answer simple control questions.

A critical examination of the predictions that we derive from the standard and extended UMGT paradigms could result in doubt on whether the predictions follow necessarily and whether other predictions cannot be derived. Although it is true that both paradigms are somewhat flexible in the range of predictions that they allow for, it is still the case that aggravations and alleviations are a key feature of moral situations and therefor require distinct machinery in cognitive models of moral reasoning. Furthermore, the claim of these paradigms is not to fully understand moral cognition *ab initio* but rather to provide a paradigm in which the quest for this understanding can be pursued. Further research is therefor required in order to discover the categories of aggravations and alleviations and under what circumstances these are deemed valid.

Our experiment showed for instance that it does not matter whether the protagonist is framed as a "CEO of a big corporation" or a "loving father", nor whether he cared about the negative side-effects of his decisions. These potential aggravations are apparently not considered to be valid in the Knobe scenario. Future studies have to determine whether these framing effects would work under other conditions or whether they should not be considered as valid alleviations in general.

Our non-monotonic paradigm is furthermore not meant as a replacement of UMGT but rather as an extension that updates the paradigm as to to be more in line with the dynamic non-monotonic nature of human reasoning.

Future work should also explore other non-monotonic formalisms of moral reasoning. These can be realized with other non-monotonic logics (Berzati, 2007). Non-monotonicity could perhaps also be realized within the action tree's of standard UMGT, with a Bayesian modeling approach, with mental models (Bucciarelli et al., 2008; Bucciarelli & Johnson-Laird, 2005) or with constraint-based models (Simon & Holyoak, 2002). In addition to correspondence with empirical findings, models should also be efficient in terms of computational complexity as the brain has finite resources an has evolved to be efficient. Although non-monotonic logics generally give rise to hard problems, under some circumstances computational algorithms can be made tractable (Cadoli & Schaerf, 1993).

Our investigation shows how an interdisciplinary effort of combining insights from linguistics, experimental psychol-

ogy, ethics and modern logic can further our understanding of moral cognition.

References

Aquinas, T., & Hutchins, R. M. (1952). *The Summa Theologica of Saint Thomas Aquinas*. Encyclopedia Britannica.

Berzati, D. (2007). *Nonmonotonic reasoning: a unifying framework*. Nova Publishers.

Brewka, G., Dix, J., & Konolige, K. (1997). *Nonmonotonic reasoning: an overview* (Vol. 73). CSLI publications Stanford.

Bucciarelli, M., & Johnson-Laird, P. N. (2005). Naïve deontics: A theory of meaning, representation, and reasoning. *Cognitive Psychology*, *50*(2), 159–193.

Bucciarelli, M., Khemlani, S., & Johnson-Laird, P. N. (2008). The psychology of moral reasoning. *Judgment and Decision making*, *3*(2), 121–139.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, *6*(1), 3–5.

Cadoli, M., & Schaerf, M. (1993). A survey of complexity results for non-monotonic logics. *The Journal of Logic Programming*, *17*(2), 127–160.

Chomsky, N. (1969). *Aspects of the theory of syntax* (Vol. 11). MIT press.

Chomsky, N. (2002). *Syntactic structures*. Walter de Gruyter.

Cummins, D. D. (1995). Naive theories and causal deduction. *Memory & Cognition*, *23*(5), 646–658.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, *5*.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? a comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, *59*(2), 93.

Horty, J. F. (1997). Nonmonotonic foundations for deontic logic. In *Defeasible deontic logic* (pp. 17–44). Springer.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(279), 190–194.

Kohlberg, L. (1981). *Essays on moral development. vol. 1, the philosophy of moral development: moral stages and the idea of justice*. Harper & Row.

Krantz, J. H., & Dalal, R. (2000). Validity of web-based psychological research.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*, *11*(4), 143–152.

Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. Cambridge University Press.

Nichols, S., & Ulatowski, J. (2007). Intuitions and individual differences: The Knobe effect revisited. *Mind & Language*, *22*(4), 346–365.

Nyamsuren, E., & Taatgen, N. A. (2014). Human reasoning module. *Biologically Inspired Cognitive Architectures*, *8*, 1–18.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, *5*(5), 411–419.

Quinn, W. S. (1989). Actions, intentions, and consequences: The doctrine of double effect. *Philosophy & Public Affairs*, 334–351.

Rawls, J. (1999). *A theory of justice*. Harvard university press.

Reiter, R. (1980). A logic for default reasoning. *Artificial intelligence*, *13*(1), 81–132.

Shariff, A. F., & Norenzayan, A. (2011). Mean gods make good people: Different views of god predict cheating behavior. *The International Journal for the Psychology of Religion*, *21*(2), 85–96.

Simon, D., & Holyoak, K. J. (2002). Structural dynamics of cognition: From consistency theories to constraint satisfaction. *Personality and social psychology review*, *6*(4), 283–294.

Stenning, K., & Van Lambalgen, M. (2008). *Human reasoning and cognitive science*. MIT Press.

Thomson, J. J. (1985). The trolley problem. *Yale Law Journal*, *94*, 1395–1415.