COUNTERFACTUALS IN CRITICAL THINKING WITH APPLICATION TO MORALITY

Luís Moniz Pereira[1] and **Ari Saptawijaya**[2]

NOVA Laboratory for Computer Science and Informatics (NOVA LINCS)
Departamento de Informática
Universidade Nova de Lisboa, Portugal

Abstract

Counterfactuals are conjectures about what would have happened, had an alternative event occurred. It provides lessons for the future by virtue of contemplating alternatives; it permits thought debugging; it supports a justification why different alternatives would have been worse or not better. Typical expressions are: "If only I were taller…", "I could have been a winner…", "I would have passed, were it not for…", "Even if… the same would follow". Counterfactuals have been well studied in Linguistics, Philosophy, Physics, Ethics, Psychology, Anthropology, and Computation, but not much within Critical Thinking. The purpose of this study is to illustrate counterfactual thinking, through logic program abduction and updating, and inspired by Pearl's structural theory of counterfactuals, with an original application to morality, a common concern for critical thinking.
In summary, we show counterfactual reasoning to be quite useful for critical thinking, namely about moral issues.

Keywords: Critical Thinking, Counterfactual Reasoning, Abduction, Morality.

1 COUNTERFACTUAL REASONING

Counterfactual literally means contrary to the facts. Counterfactual reasoning involves thoughts on what could have happened, had some matter –action, outcome, etc.– been different in the past. Counterfactual thinking covers everyday experiences, like regret: "If only I had told her I love her!", "I should have studied harder"; or guilt responsibility, blame, causation: "If only I had said something sooner, then I could have prevented the accident". The general form is: "*If the* $\langle Antecedent \rangle$ *had been true, then the* $\langle Consequent \rangle$ *would have been true*".

Counterfactuals have been well studied in Linguistics, Philosophy, Physics, Ethics, Psychology, Anthropology, and Computation (Collins et al. 2004, Hoerl

[1]Email: lmp@fct.unl.pt
[2]Affiliated also with Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia.
Email: ar.saptawijaya@campus.fct.unl.pt

et al. 2011, Lewis 1973, Pearl 2000, Roese & Olson 2009), but oddly not much within Critical Thinking. However, people often think how things that matter to them might have turned out differently (Mandel et al. 2005). Researchers from psychology have asked: Why do people have such a strong tendency to generate counterfactuals? What functions does counterfactual thinking serve? What are the determinants of counterfactual thinking? What are its adaptive and psychological consequences?
Human's ability for such mental time travel relies on episodic memory. Without it humans would be unable to form a stable concept of self along time, and human cultures would not have been able to evolve.

In this paper, counterfactual reasoning is enacted using a three-step logic evaluation procedure (Pereira and Saptawijaya 2015c), inspired by the structure-based approach of Pearl (2000), viz., (1) ***Abduction***: to explain past circumstances in the presence of observed evidence, i.e., update and fix the circumscribing setting of exogenous background information, on the basis of the given evidence; (2) ***Action***: to adjust the logical causal model to comply with the antecedent of the counterfactual, i.e., to impose the truth of the antecedent's hypotheses by means of a forced intervention on the model; and (3) ***Prediction***: to predict if the counterfactual's consequent deductively follows, subsequently to steps 1 and 2, i.e., to compute the truth-value of the consequent in the modified intervened model.

The approach is realized by means of logic program abduction and updating. Abduction chooses from available hypotheses (the set A of *abducibles*) –the exogenous variables that constitute the situation's background– i.e., those abducibles or their negations, that best explain the observed given evidence O. An abduced explanation, E, is a subset of A that finds the specific values for exogenous variables, which lend an explanatory support to all currently observed evidence. Note that the abduction procedure guarantees the abduced explanation to be consistent, i.e., disallows both abducible a and its negation a^* to hold in explanation E.[3] Subsequent to abduction, updating modifies those

[3] In the sequel, starred atoms stand for their negations.

rules to be updated and fixes the initially abduced exogenous background context of the counterfactual statement. I.e., updates the knowledge base with some preferred explanation to the current observations and, additionally, the updating also permits causal intervention on the causal knowledge model, namely by means of hypothetical updates to the rules, achieved via reserved predicate $make$ (illustrated in examples below), so as to render the knowledge base consistently compliant with the antecedent of the counterfactual.

Consider an example (Byrne 2007): *Lightning hits a forest, and a devastating forest fire breaks out. The forest was dry, after a long hot summer.* Let us add more causes for forest fire, i.e., there are two possible alternative causes: storm –presuming the lightning– or barbecue. The model of this example consists in a set of abducibles $A = \{storm, barbecue, storm^*, barbecue^*\}$ and program P:

$fire \leftarrow barbecue, dry\_leaves.$ $\quad\quad$ $dry\_leaves.$

$fire \leftarrow barbecue^*, lightning, dry\_leaves.$ $\quad\quad$ $lightning \leftarrow storm.$

Take counterfactual statement:

If only there had not been lightning, then the forest fire would not have occurred.

Step 1: Given the observation $O = \{lightning, fire\}$, abduce its explanations E (a subset of A). Note that the observations assure us that both the antecedent and the consequent of the counterfactual were *factually* false. Two possible explanations for O (s and b standing for $storm$ and $barbecue$, resp.): $E_1 = \{s, b^*\}$ and $E_2 = \{s, b\}$. Say E_1 is preferred for consideration. Then fix its abduced background context for the counterfactual: i.e., update program P with E_1.

Step 2: Update program P, to get a new program T, by adding:

$make(lightning^*)$. % Intervention: *If only there had not been lightning*...

$lightning \leftarrow make(lightning)$. % Note that lightning or otherwise are

$lightning^* \leftarrow make(lightning^*)$. % now available only by intervention.

And, for irrelevancy and consistency, deleting: $lightning \leftarrow storm$.

Step 3: Verify if the conclusion "*the forest fire would not have occurred*" is true. Since $fire$ is not provable, '$not\ fire$' holds in the semantics of T for explanation $E_1 = \{s, b^*\}$ with intervention $make(lightning^*)$. The counterfactual is valid.

2 COUNTERFACTUALS IN MORALITY

Typically, people think critically about what they should or should not have done when they examine decisions in moral situations. It is therefore natural for them to engage in counterfactual thoughts of alternatives in such settings. Counterfactual thinking has been investigated in the context of moral reasoning, notably by psychology experimental studies (Byrne 2007), e.g., to understand the kind of critical counterfactual alternatives people tend to think of in contemplating moral behaviours, and the influence of counterfactual thoughts in moral judgment (Mandel et al. 2005, Roese & Olson 2009).

Morality and normality judgments typically correlate. Normality mediates morality with causation and blame judgments. The intervention controllability in counterfactuals mediates between normality, blame and cause judgments. The importance of control, namely the possibility intervention, is highlighted in theories of blame that presume someone responsible only if they had some control of the outcome (Weiner 1995).

As argued by Epstude and Roese (2008), the function of counterfactual thinking is not just limited to the evaluation process, but occurs also in the reflection one. Through evaluation, counterfactuals help correct wrong behaviour in the past, thus guiding future moral decisions. Reflection, on the other hand, permits momentary experiential simulation of possible alternatives, thereby allowing careful consideration before a decision is made, and to subsequently justify it.

The investigation in this paper pertains to how moral issues can innovatively be expressed with counterfactual reasoning by resorting to the aforementioned approach. In particular, its application for examining viewpoints on moral permissibility is scrutinized, exemplified by classic moral dilemmas from the literature on the Doctrine of Double Effect (DDE) (McIntyre 2004), and the Doctrine of Triple Effect (DTE) (Kamm 2006).

DDE is often invoked to explain the permissibility of an action that causes a harm by distinguishing whether this harm is a mere side effect of bringing about

a good result, or rather a means to bringing about the same good end (McIntyre 2004). In Hauser et al. (2007), DDE has been utilized to explain the consistency of judgments, shared by subjects from demographically diverse populations, on a series of moral dilemmas.

Counterfactuals may provide a general way to examine DDE in dilemmas, e.g., the classic trolley problem (Foot 1967), by distinguishing between cause and side effect of performing an action to achieve a goal. This distinction between causes and side effects may explain the permissibility of an action in accordance with DDE. I.e., if some morally wrong effect E happens to be a cause for a goal G that one wants to achieve by performing an action A, and E is not a mere side effect of A, then performing A is impermissible. The counterfactual form below, in a setting where action A is performed to achieve goal G, expresses this: "*If $not\ E$ had been true, then $not\ G$ would have been true.*"

The evaluation of this counterfactual form identifies permissibility of action A from its effect E, by identifying whether the latter is a necessary cause for goal G or a mere side effect of action A. That is, if the counterfactual proves valid, then E is instrumental as a cause of G, and not a mere side effect of action A. Since E is morally wrong, achieving G that way, by means of A, is impermissible; otherwise, not.

Note that the evaluation of counterfactuals in this application is considered from the perspective of agents who perform the action, rather than from anothers' (e.g., observers). Moreover, the emphasis on causation in this application focuses on agents' deliberate actions, rather than on causation and counterfactuals in general, cf. Pearl (2000) and Collins et al. (2004).

In the next examples, the aforementioned general counterfactual method is illustrated by taking off-the-shelf military morality cases (Scanlon 2008).

Consider "Terror Bombing", teb for short, which means: Bombing a civilian target during a war, thus killing many civilians, in order to terrorize the enemy, and thereby getting them to end the war. DDE affirms teb *impermissible*. On the other hand, "Tactical bombing" (tab) means: Bombing a military target, which will effectively end the war, but with the foreseen consequence of killing the same large number of civilians nearby. DDE affirms tab *permissible*.

Modeling Terror Bombing.

Take set of abducibles $A = \{teb, teb^*\}$ and program P:

$$end\_war \leftarrow terror\_civilians. \qquad terror\_civilians \leftarrow kill\_civilians.$$
$$kill\_civilians \leftarrow target\_civilians. \qquad target\_civilians \leftarrow teb.$$

Counterfactual: *If civilians had not been killed, the war would not have ended.*

The evaluation follows.

Step 1: Observations $O = \{kill\_civilians, end\_war\}$ with explanation $E = \{teb\}$.

Step 2: Produce program T from P:

$make(kill\_civilians^*)$. % Intervention: *If civilians had not been killed*...

$kill\_civilians \leftarrow make(kill\_civilians)$. % Killing civilians or otherwise is now

$kill\_civilians^* \leftarrow make(kill\_civilians^*)$. % available only by intervention.

And, for irrelevancy and consistency, delete: $kill\_civilians \leftarrow target\_civilians$.

Step 3: The counterfactual is valid since conclusion "*the war would not have ended*" is true. Indeed, '$not\ end\_war$' holds in the semantics of updated T plus E. Hence, the morally wrong action $kill\_civilians$ is an instrument to achieve the goal end_war. It is a cause of end_war by performing teb, and not a mere side effect of teb. Therefore, teb is DDE morally ***impermissible***.

Modeling Tactical Bombing.

Take set of abducibles $A = \{tab, tab^*\}$ and program P:

$end\_war \leftarrow target\_military.$ $kill\_civilians \leftarrow tab.$ $target\_military \leftarrow tab.$

The counterfactual is the same as above. The evaluation follows.

Step 1: Observations $O = \{kill\_civilians, end\_war\}$ with explanation $E = \{tab\}$.

Step 2: Produce T from P, obtaining same T as in the terror bombing's model.

And, for irrelevancy and consistency, now delete: $kill\_civilians \leftarrow tab$.

Step 3: The counterfactual is not valid, since its conclusion "*the war would not have ended*" is false. Indeed, end_war holds in the semantics of updated T plus E. Hence, the morally wrong $kill\_civilians$ is a just side effect of achieving the goal end_war. Therefore, tab is DDE morally permissible.

A more complex scenario can challenge this application of counterfactuals, to distinguish moral permissibility according to DDE vs. DTE. DTE (Kamm 2006) refines DDE particularly on the notion about harming someone as an intended means. That is, DTE distinguishes further between doing an action *in order* that an effect occurs and doing it *because* that effect will occur. The latter is a new category of action, which is not accounted for in DDE. Though DTE also classifies the former as impermissible, it is more tolerant to the latter (the third effect), i.e., it treats as permissible those actions performed just *because* instrumental harm will occur.

Kamm proposed DTE to accommodate a variant of the trolley problem, viz., the *Loop Case* (Thomson 1985):
A trolley is headed toward five people walking on the track, and they will not be able to get off the track in time. The trolley can be redirected onto a side track, which loops back towards the five. A fat man sits on this looping side track, whose body will by itself stop the trolley. Is it morally permissible to divert the trolley to the looping side track, thereby hitting the man and killing him, but saving the five?

This case strikes most moral philosophers that diverting the trolley is permissible (Otsuka 2008). Referring to a psychology study (Hauser et al. 2007), 56% of its respondents judged that diverting the trolley in this case is also permissible. To this end, DTE may provide the justification of its permissibility (Kamm 2006). Nonetheless, DDE views diverting the trolley in the Loop case as impermissible.

Modeling Loop Case.

Take set of abducibles $A = \{divert, divert^*\}$ and program P, where $save$, $divert$, hit, tst, mst stand for "save the five", "divert the trolley", "man hit by the trolley", "train on the side track", and "man on the side track", respectively:

$save \leftarrow hit.$ $hit \leftarrow tst, mst.$ $tst \leftarrow divert.$ $mst.$

Counterfactual: *If the man had not been hit by the trolley, the five people would not have been saved.* The evaluation follows.

Step 1: Observations $O = \{hit, save\}$ with explanation $E = \{divert\}$.

Step 2: Produce program T from P:

$make(hit^*)$. % Intervention: *If the man had not been hit by the trolley*...

$hit \leftarrow make(hit)$. % The man being hit by the trolley or otherwise is now

$hit^* \leftarrow make(hit^*)$. % available only by intervention.

And, for irrelevancy and consistency, now delete: $hit \leftarrow tst, mst.$

Step 3: The counterfactual is valid, since its conclusion "*the five people would not have been saved*" is true. Indeed, '$not\ save$' holds in the semantics of updated T plus E. Hence, hit, as a consequence of action $divert$, is instrumental as a cause of goal $save$. Therefore, $divert$ is DDE morally *impermissible*.

DTE considers diverting the trolley as permissible, since the man is already on the side track, without any deliberate action performed in order to place him there. In the above program, we have the fact mst ready, without abducing any ancillary action. The validity of the counterfactual "*if the man had not been on the side track, then he would not have been hit by the trolley*", which can easily be verified, ensures that the unfortunate event of the man being hit by the trolley is indeed the consequence of the man being on the side track. The lack of deliberate action (say, by pushing the man -- $push$ for short) in order to place him on the side track, and whether the absence of this action still causes the unfortunate event (the third effect) is captured by the counterfactual "*if the man had not been pushed, then he would not have been hit by the trolley*". This counterfactual is not valid, because the new observation $O = \{push, hit\}$ has no explanation: $push$ is not in the set of abducibles A, and moreover there is no

fact $push$ either. This means that even without this hypothetical but unexplained deliberate action of pushing, the man would still have been hit by the trolley (just because he is already on the side track). In summary, though hit is a consequence of div and instrumental in achieving $save$, no deliberate action is required to cause mst, in order for hit to occur. Hence $divert$ is DTE morally permissible.

In order to further distinguish moral permissibility with respect to DDE and DTE, we also consider a variant of the Loop case, viz., the *Loop-Push* case -- see also the Extra Push case in Kamm (2006). Differently from the Loop case, in this Loop-Push case the looping side track is initially empty, and besides the diverting action, an ancillary action of pushing a fat man in order to place him on the side track is additionally performed.

Modeling Loop-Push Case.

Take set of abducibles $A = \{divert, push, divert^*, push^*\}$ and program P:

$save \leftarrow hit.$ $\quad hit \leftarrow tst, mst.$ $\quad tst \leftarrow divert.$ $\quad mst \leftarrow push.$

Recall the counterfactuals considered in the discussion of DDE and DTE of the Loop case:

- *"If the man had not been hit by the trolley, the five people would not have been saved."* The same observation $O = \{hit, save\}$ provides an extended explanation $E = \{divert, push\}$. That is, the pushing action needs to be abduced for having the man on the side track, so the trolley can be stopped by hitting him. The same intervention $make(hit^*)$ is applied to the same transform T, resulting in a valid counterfactual: $not\ sav$ holds in the semantics of updated T plus E.
- "*If the man had not been pushed, then he would not have been hit by the trolley.*" The relevant observation is $O = \{push, hit\}$, explained by $E = \{divert, push\}$. Whereas this counterfactual is not valid in DTE of the Loop case, it is valid in the Loop-Push case. Given rule $push^* \leftarrow make(push^*)$ in the transform T and intervention $make(push^*)$, we verify that $not\ hit$ holds in the semantics of updated T plus E.

From the validity of these two counterfactuals it can be inferred that, given the diverting action, the ancillary action of pushing the man onto the side track causes him to be hit by the trolley, which in turn causes the five to be saved. In the Loop-Push, DTE agrees with DDE that such a deliberate action (pushing) performed in order to bring about harm (the man hit by the trolley), even for the purpose of a good or greater end (to save the five), is likewise impermissible.

3 CONCLUSIONS AND FURTHER WORK

Computational morality (Anderson & Anderson 2011, Wallach & Allen 2009) is a burgeoning field that emerges from the need of imbuing autonomous agents with the capacity of moral decision making to enable them to function in an ethically responsible manner via their own ethical decisions. It has attracted the artificial intelligence community, and brought together perspectives from various fields: philosophy, anthropology, cognitive science, neuroscience, and evolutionary biology. The overall result of this interdisciplinary research is not just important for equipping agents with some capacity for making moral judgments, but also to help better understand morality, via the creation and testing of computational models of ethical theories.

This paper presented a formulation of counterfactuals evaluation by means of logic program abduction and updating. The approach corresponds to the three-step process in Pearl's structural theory, despite omitting probability to concentrate on a naturalized logic. Furthermore, counterfactual reasoning has been shown quite useful for critical thinking, namely about moral issues, where (non-probabilistic) moral reasoning about permissibility is examined by employing this logic program approach to distinguish between causes and the side effects that are the result of agents' actions to achieve goals.

In Pearl's theory, intervention is realized by superficial revision, i.e., by imposing the desired value to the intervened node and cutting it from its parent nodes. This is also the case in the approach presented here, achieved by hypothetical

updates via the reserved predicate $make$. Other subtle ways of intervention may involve deep revision, realizable with logic programs, cf. Pereira et al. (2015a).

Logic program abduction was used in Kowalski (2011) and Pereira and Saptawijaya (2011) to model moral reasoning in various scenarios of the trolley problem, both from DDE and DTE viewpoints, sans counterfactuals. Abducibles are used to represent decisions, where impermissible actions are ruled out using an integrity constraint, and *a posteriori* preferences are eventually enacted to come up with a moral decision from the remaining alternatives of action. Subsequent work (Han et al. 2012) refines it with uncertainty of actions and consequences in several scenarios of the trolley problem by resorting to probabilistic logic programming P-log (Baral & Hunsaker 2007).

Side effects in abduction have been investigated in Pereira et al. (2013) through the concept of inspection points; the latter are construed in a procedure by 'meta-abducing' a specific abducible, $abduced(a)$, whose function is only checking that its corresponding abducible a is indeed already abduced elsewhere. Therefore, the consequence of the action that triggers this 'meta-abducing' is merely a side effect. Indeed, inspection points may be employed to distinguish a cause from a mere side effect, and thus may provide an alternative or supplement to counterfactuals employed for the same purpose.

Counterfactuals may as well be suitable to address moral justification, via 'compound counterfactuals': *Had I known what I know today, then if I were to have done otherwise, something preferred would have followed*. Such counterfactuals, by imagining alternatives with worse effect –the so-called *downward counterfactuals* (Markman et al. 1993)– may provide justification for what was done due to lack of the current knowledge. This is accomplished by evaluating what would have followed if the intent had been otherwise, other things (including present knowledge) being equal. It may justify that what would have followed is no morally better than the actual ensued consequence. We are currently investigating the application of counterfactuals to justify an exception for an action to be permissible (Pereira and Saptawijaya 2015b; Saptawijaya

and Pereira, 2015), which may lead to agents' argumentation following contractualism of Scanlon (1998).

ACKNOWLEDGEMENTS

Ari Saptawijaya is supported by FCT/MEC Portugal with grant SFRH/BD/72795/2010. We thank Emmanuelle-Anna Dietz for fruitful discussions.

REFERENCES

Anderson, M. & Anderson, S. L., editors (2011). *Machine Ethics*. Cambridge: Cambridge University Press.

Baral, C. & Hunsaker, M. (2007). *Using the probabilistic logic programming language P-log for causal and counterfactual reasoning and non-naive conditioning*. In Proceedings of IJCAI'07, Los Angeles: AAAI Press.

Byrne, R. M. J. (2007). *The Rational Imagination: How People Create Alternatives to Reality*. Cambridge, MA: MIT Press.

Collins, J., Hall, N., & Paul, L. A., editors (2004). *Causation and Counterfactuals*. Cambridge, MA: MIT Press.

Epstude, K. & Roese,N. J. (2008). *The functional theory of counterfactual thinking*. Personality and Social Psychology Review, 12(2):168–192.

Foot, P. (1967). *The problem of abortion and the doctrine of double effect*. Oxford Review, 5:5–15.

Han, T. A., Saptawijaya, A., & Pereira, L. M. (2012). *Moral reasoning under uncertainty*. In LPAR-18, LNCS 7180, pages 212–227. Berlin: Springer.

Hauser, M., Cushman, F., Young, L., Jin, R. K., & Mikhail, J. (2007). *A dissociation between moral judgments and justifications*. Mind and Language, 22(1):1–21.

Hoerl, C., McCormack, T., & Beck, S. R., editors (2011). *Understanding Counterfactuals, Understanding Causation: Issues in Philosophy and Psychology*. Oxford: Oxford University Press.

Kamm, F. M. (2006). *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford: Oxford University Press.

Kowalski, R. (2011). *Computational Logic and Human Thinking: How to be Artificially Intelligent*. Cambridge: Cambridge University Press.

Lewis, D. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.

Mandel, D. R., Hilton, D.J., & Catellani, P (2005). *The Psychology of Counterfactual Thinking*. New York, NY: Routledge.

Markman, K. D., Gavanski, I., Sherman,S.J., & McMullen, M.N. (1993).*The mental simulation of better and worse possible worlds*. Journal of Experimental Social Psychology, 29:87–109.

McIntyre, A. (2004). *Doctrine of double effect*. In The Stanford Encyclopedia of Philosophy.Zalta, E. N., editor, Center for the Study of Language and Information, Stanford University, Fall 2011 edition, 2004. http://plato.stanford.edu/ archives/fall2011/entries/double-effect/

Otsuka, M. (2008). *Double effect, triple effect and the trolley problem: Squaring the circle in looping cases*. Utilitas, 20(1):92–110.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge, MA: Cambridge University Press.

Pereira, L. M., Dell'Acqua, P., Pinto, A. M., & Lopes, G. (2013). *Inspecting and preferring abductive models*. In The Handbook on Reasoning-BasedIntelligent Systems. Nakamatsu, K. & Jain, L. C., editors, pages 243–274. London: World Scientific Publishers.

Pereira, L. M., Dietz, E.-A., & Hölldobler, S. (2015a). *Abductive Framework for Counterfactual Reasoning in Logic Programming*. Available from: http://centria.di.fct.unl.pt/~lmp/publications/online-papers/counterfactuals.pdf

Pereira, L. M. & Saptawijaya. A. (2011). *Modelling Morality with Prospective Logic*. In Machine Ethics. Anderson, M. & Anderson,S. L., editors, pages 398–421. Cambridge, MA: Cambridge University Press.

Pereira, L. M. & Saptawijaya. A. (2015b). *Abduction and Beyond in Logic Programming with Application to Morality*. Available from: http://centria.di.fct.unl.pt/~lmp/publications/online-papers/abduction&beyond.pdf

Pereira, L. M. & Saptawijaya, A. (2015c). Counterfactuals in Logic Programming with Applications to Agent Morality. Available from: http://centria.di.fct.unl.pt/~lmp/publications/online-papers/moral_counterfactuals.pdf

Roese, N.J. & Olson, J. M., editors (2009). *What Might Have Been: The Social Psychology of Counterfactual Thinking*. New York, NY: Psychology Press.

Saptawijaya, A. & Pereira, L. M. (2015). Logic Programming Applied to Machine Ethics. Available from: http://centria.di.fct.unl.pt/~lmp/publications/online-papers/lp_app_mach_ethics.pdf

Scanlon, T. M. (1998). *What We Owe to Each Other.* Cambridge, MA: Harvard University Press.

Scanlon, T. M. (2008). *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge, MA: Harvard University Press.

Thomson, J. J. (1985). *The trolley problem*. The Yale Law Journal, 279:1395–1415.

Wallach, W. & Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.

Weiner, B. (1995). *Judgments of Responsibility: A Foundation for a Theory of Social Conduct*. The Guilford Press.