Ethical Decision Making under the Weak Completion Semantics

Steffen Hölldobler
International Center for Computational Logic, Technische Universität Dresden, Germany
and North-Caucasus Federal University, Stavropol, Russian Federation
sh@iccl.tu-dresden.de

Abstract

The weak completion semantics is a novel computational theory based on logic programs. It is extended to deal with equalities, which is a prerequisite to represent and reason about actions and causality as in the fluent calculus. This is discussed in the context of ethical decision making. In order to decide questions about the moral permissibility of actions, counterfactuals need to be considered. Somewhat surprisingly, this can be straightforwardly done in the extended approach.

1 Introduction

The weak completion semantics (WCS) is a novel cognitive theory. Its original idea is based on [Stenning and van Lambalgen2008] who proposed to model human reasoning tasks by, firstly, reasoning towards a normal logic program to represent the reasoning task and, secondly, by reasoning with respect to the least model of the normal logic program. Unfortunately, Stenning and van Lambalgen's approach contained a technical bug which was corrected in [Hölldobler and Kencana Ramli2009].

The WCS is based on many techniques and methods from logic programming and computational logic. However, these techniques and methods are usually tweaked a little bit in order to model human reasoning tasks adequately. For example, programs are not completed in the sense of [Clark1978], but only weakly completed. Instead of the semantic operator introduced in [Fitting1985], a modified operator introduced in [Stenning and van Lambalgen2008] is used. Instead of the three-valued Kripke-Kleene logic used in [Fitting1985, Stenning and van Lambalgen2008], the three-valued Łukasiewicz logic [Łukasiewicz1920] is used. Because of the latter, normal logic programs admit a least model and reasoning is performed with respect to this model (see [Hölldobler and Kencana Ramli2009]).

The approach has been applied to various human reasoning tasks like the suppression task [Byrne1989, Dietz *et al.*2012], the selection task [Wason1968, Dietz *et al.*2013], and human syllogistic reasoning [Khemlani and Johnson-Laird2012, Oliviera da Costa *et al.*2017]. In fact, WCS performed better on the human syllogistic reasoning tasks than all 12 cognitive theories discussed in [Khemlani and Johnson-Laird2012]. As all human reasoning tasks are solved within one framework, the WCS is an integrated and computational cognitive theory. We are unaware of any other theory of this kind and with such a wide variety of applications.

Recently, ethical decision making has received much attention as autonomous agents become part of our daily life. In particular, we were inspired by [Pereira and Saptawijaya2016], who studied computational models of machine ethics. Various ethical problems are implemented as logic programs and these programs can be queried for moral permissibility. Unfortunately, their approach does not provide a general method to account for ethical dilemmas and is not integrated into a cognitive theory about human reasoning.

The problems studied in [Pereira and Saptawijaya2016] were trolley problems or variants thereof like the bystander case. In these problems, actions with direct and indirect effects must be considered. Hence, in order to model and reason about these problems within the WCS, the WCS must be extended to deal with actions and causality. We have chosen the fluent calculus [Hölldobler and Schneeberger1990] for modeling actions and causality because it treats fluents as resources which can be consumed and produced. This property is shared with Petri networks [Hölldobler and Jovan2014], the latter of which have already been used in computational models for human reasoning [Barrett2010].

In the fluent calculus [Hölldobler and Schneeberger1990] states are represented as multisets of fluent. Multisets are represented with the help of a binary function symbol $\circ$ written infix and a constant 1 such that $\circ$ is commutative, associative, and 1 is its unit element. For example, the multisets $\dot{\{}\,\dot{\}}$ and $\dot{\{}a, b, b\dot{\}}$ are represented by the fluent terms 1 and $a \circ b \circ b$, respectively. In order to deal with function symbols like $\circ$ in the WCS, we need to extend WCS to handle equality. Luckily, as shown in [Dietz Saldanha *et al.*2018] the key properties of the WCS, viz. the existence of a least model and the fact that this model can be computed as the least fixed point of an appropriate semantic operator, hold also for logic programs with equality.

In this paper, we will focus on the representation of the bystander case. We will show how to represent this problem in the extended approach. In particular, we formalize a purely utilitarian view [Bentham2009] and the doctrine of double effect [Aquinas1988]. In order to decide which action is morally permissible in the bystander case we need to reason about a counterfactual [Nickerson2015]. It turns out, that this can be straightforwardly done in the extended approach.

2 The Weak Completion Semantics with Equality

We assume the reader to be familiar with the WCS as presented in [Hölldobler2015, Dietz Saldanha *et al.*2017]. In the weak completion semantics with equality (WCSE) a logic program $\mathcal{P}$ is considered together with a set $\mathcal{E}$ of equations. As shown in [Jaffar *et al.*1984], $\mathcal{E}$ defines a finest congruence relation on the set of ground terms. Let $[t]$ denote the congruence class defined by the ground term t. For example, $[a \circ b \circ b] = [b \circ a \circ b] = [b \circ b \circ a \circ 1]$. Furthermore, let $[p(t_1, \ldots, t_n)]$ be an abbreviation for $p([t_1], \ldots, [t_n])$, where p is an n-ary relation symbol and all t_i, $1 \leq i \leq n$, are ground terms. $[p(t_1, \ldots, t_n)] = [q(s_1, \ldots, q_m)]$ if and only if $p = q$, $n = m$, and $[t_i] = [s_i]$ for all $1 \leq i \leq n$. For example, $[p(a \circ b \circ b, 1)] = [p(b \circ a \circ b, 1 \circ 1]$. We consider $\mathcal{E}$-interpretations and $\mathcal{E}$-models as usual (see e.g. [Jaffar *et al.*1984]).

As shown in [Dietz Saldanha *et al.*2018], a logic program $\mathcal{P}$ together with a set $\mathcal{E}$ of equation has a least $\mathcal{E}$-model under the three-valued Łukasiewicz logic [Łukasiewicz1920]. This model is the least fixed point of the following semantic operator: Let I be an $\mathcal{E}$-interpretation. We define $\Phi_{\mathcal{P}}^{\mathcal{E}}(I) = \langle J^\top, J^\bot \rangle$ where

$$\begin{array}{rcl} J^\top & = & \{[A] \mid \text{there exists } A \leftarrow \mathit{Body} \in \mathrm{g}\mathcal{P} \\ & & \quad \text{and } I(\mathit{Body}) = \top\}, \\ J^\bot & = & \{[A] \mid \text{there exists } A \leftarrow \mathit{Body} \in \mathrm{g}\mathcal{P} \\ & & \quad \text{and for all } A' \leftarrow \mathit{Body} \in \mathrm{g}\mathcal{P} \\ & & \quad \text{with } [A] = [A'] \\ & & \quad \text{we find } I(\mathit{Body}) = \bot\}, \end{array}$$

and g$\mathcal{P}$ denotes the set of all ground instances of clauses occurring in $\mathcal{P}$.

One should observe that the set $\mathcal{E}$ of equations is built into the computation of the $\Phi_{\mathcal{P}}^{\mathcal{E}}$-operator: In the computation of $J^\top$, if a ground atom A is mapped to true because it is the head of a rule whose body is true, then all members of the congruence class containing A are mapped to true. Likewise, in the computation of $J^\bot$ we do not only have to consider all rules with head A, but all rules whose head A' is in the same congruence class as A, and if A is mapped to false, then all members of the congruence class containing A are mapped to false.

3 The Bystander Case

A trolley, whose conductor has fainted, is headed towards two people walking on the main track.[1] *The banks of the track are so steep that these two people will not be able to get off the track in time. Hank is standing next to a switch, which can turn the trolley onto a side track, thereby preventing it from killing the two people. However, there is a man standing on the side track. Hank can change the switch, killing him. Or he can refrain from doing so, letting the two die. Is it morally permissible for Hank to change the switch?*

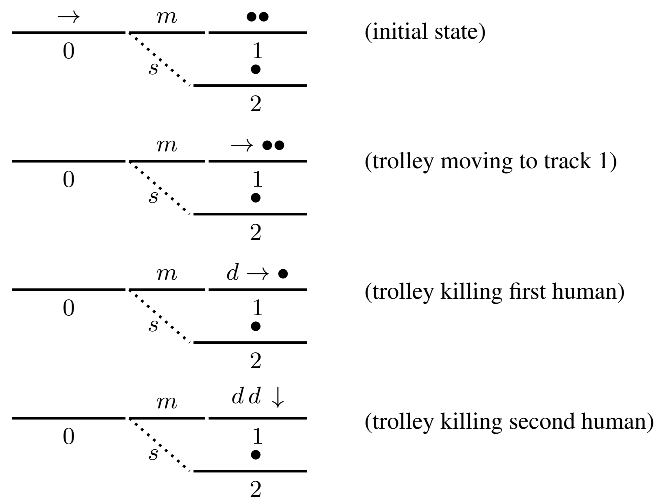

Figure 1: The bystander case (initial state) and its ramifications if Hank decides to do nothing, where ↓ denotes that no further action is applicable.

The case is illustrated in Figure 1 (initial state). The tracks are divided into segments 0, 1, and 2, the arrow represents that the trolley t is moving forward and that the track is clear (c), the switch is in position m (main) but can be changed into position s (side), and a bullet above a track segment represents a human (h) on this track. t, c, and h may be indexed to denote the track to which they apply. In addition, we need a fluent d denoting a dead human.

We choose to represent a state by a pair of multisets consisting of the casualties in its second element and all other fluents in its first element. Multisets are represented by so-called *fluent terms* in the fluent calculus, i.e., the initial state of the bystander case is the pair

$$(t_0 \circ c_0 \circ m \circ h_1 \circ h_1 \circ h_2, 1) \tag{1}$$

of fluent terms. The casualties are represented in the second element of (1) by the constant 1 encoding the empty multiset. Initially, there are no casualties, but casualties will play a special role when preferring one action over another as will be discussed later in this section. The first element of (1) encodes the multiset $\dot{\{}t_0, c_0, m, h_1, h_1, h_2\dot{\}}$.

There are two kinds of actions, the ones which can be performed by Hank (the direct actions $donothing$ and $change$), and the actions which are performed by the trolley (the indirect actions $downhill$ and $kill$). We will represent the actions by the trolley explicitly with the help of a five-place relation symbol $action$ specifying the preconditions, the name, and the immediate effects of an action. As a state is represented by two multisets, the preconditions anf the immediate effects

[1] Note that in the original trolley problem, five people are on the main track. For the sake of simplicity, we assume that only two people are on the main track.

have also two parts:

$$\begin{array}{l} action(t_0 \circ c_0 \circ m, 1, downhill, t_1 \circ c_0 \circ m, 1) \leftarrow \top \\ action(t_0 \circ c_0 \circ s, 1, downhill, t_2 \circ c_0 \circ s, 1) \leftarrow \top \\ \\ action(t_1 \circ h_1, 1, kill, t_1, d) \leftarrow \top \\ action(t_2 \circ h_2, 1, kill, t_2, d) \leftarrow \top \end{array}$$

If the trolley is on track 0, this track is clear, and the switch is in position m, then it will run downhill onto track 1 whereas track 0 remains clear and the switch will remain in position m; if, however, the switch is in position s, the trolley will run downhill onto track 2. If the trolley is on either track 1 or 2 and there is a human on this track, it will kill the human leading to a casualty.

The possible actions of Hank are the base cases in the definition of causality:[2]

$$\begin{array}{l} causes(donothing, t_0 \circ c_0 \circ m \circ h_1 \circ h_1 \circ h_2, 1) \leftarrow \top \\ causes(change, t_0 \circ c_0 \circ s \circ h_1 \circ h_1 \circ h_2, 1) \leftarrow \top \end{array} \quad (2)$$

The recursive case of the definition of causality is given as

$$\begin{array}{l} causes(A, E_1 \circ Z_1, E_2 \circ Z_2) \ \leftarrow \\ \quad action(P_1, P_2, A', E_1, E_2) \wedge \\ \quad causes(A, P_1 \circ Z_1, P_2 \circ Z_2) \wedge \\ \quad \neg ab(A'). \end{array} \quad (3)$$

It checks whether in a given state $(P_1 \circ Z_1, P_2 \circ Z_2)$ an action A' is applicable, which is the case if the preconditions (P_1, P_2) are contained in the given state. If this holds, then the action is executed leading to the successor state $(E_1 \circ Z_1, E_2 \circ Z_2)$, where (E_1, E_2) are the direct effects of the action A'. In other words, if an action is applied, then its preconditions are consumed and its direct effects are produced. Such an action application is considered to be a ramification [Thielscher2003] with respect to the initial, direct action performed by Hank. Hence, the first argument A of *causes* is not changed. The execution of an action is also conditioned by $\neg ab(A')$, where ab is an abnormality predicate. Such abnormalities were introduced in [Stenning and van Lambalgen2008] to represent conditionals as licenses for inference. In this example, there is nothing abnormal known with respect to the actions $downhill$ and $kill$ and, consequently, the assumptions

$$\begin{array}{l} ab(downhill) \leftarrow \bot \\ ab(kill) \leftarrow \bot \end{array}$$

are added to the program. But we can imagine situations, where the trolley will only cross the switch if the switch is not broken.[3]

[2] In the original version of the fluent calculus, *causes* is a ternary predicate stating that the execution of a plan transfers an initial into a goal state. Its base case is of the form $causes(X, [], X)$, i.e., the empty plans transforms arbitrary states X into X. Generating models bottom up using a semantic operator one has to consider all ground instances of this atom, which is usually too large to consider as a base case for human reasoning episodes. The solution presented in this paper overcomes this problem in that we only have a small number of base cases depending on the number of options an agent like Hank may consider.

[3] If the switch is broken, the trolley may derail. Such a scenario can be modeled in WCSE as well, but it is beyond the scope of this paper to discuss it in detail.

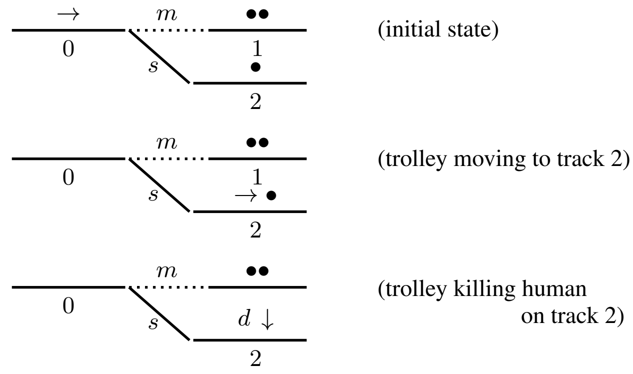

Figure 2: The bystander case (initial state) and its ramifications if Hank decides to change the switch. One should observe that now the switch points to the side track.

Let $\mathcal{P}$ be the program consisting of the clauses mentioned in this section so far and $\mathcal{E}$ be the set of equations specifying that $\circ$ is associative, commutative, and 1 being its unit element. Hank has the choice to do nothing or to change the switch. Depending on his decision, the trolley will execute its actions which are computed as ramifications in the fluent calculus [Thielscher2003]. If Hank is doing nothing, then the least $\mathcal{E}$-model of $\mathcal{P}$ – which is equal to the least fixed point of $\Phi_{\mathcal{P}}^{\mathcal{E}}$ – is computed by iterating $\Phi_{\mathcal{P}}^{\mathcal{E}}$ starting with the empty interpretation $\langle \emptyset, \emptyset \rangle$. The following equivalence classes will be mapped to true in subsequent iterations:[4]

$$\begin{array}{l} [causes(donothing, t_0 \circ c_0 \circ m \circ h_1 \circ h_1 \circ h_2, 1)] \\ [causes(donothing, t_1 \circ c_0 \circ m \circ h_1 \circ h_1 \circ h_2, 1)] \\ [causes(donothing, t_1 \circ c_0 \circ m \circ h_1 \circ h_2, d)] \\ [causes(donothing, t_1 \circ c_0 \circ m \circ h_2, d \circ d)] \end{array}$$

They correspond precisely to the four states shown in Figure 1. No further action is applicable to the elements of the final congruence class. The two people on the main track will be killed.

On the other hand, if Hank is changing the switch, then the least fixed point of $\Phi_{\mathcal{P}}^{\mathcal{E}}$ contains

$$[causes(change, t_2 \circ c_0 \circ s \circ h_1 \circ h_1, d)].$$

The two people on the main track will be saved but the person on the side track will be killed. This case is illustrated in Figure 2.

The two cases can be compared by means of a *prefer* clause:

$$\begin{array}{l} prefer(A_1, A_2) \ \leftarrow \\ \quad causes(A_1, Z_1, D_1) \wedge \\ \quad causes(A_2, Z_2, D_1 \circ d \circ D_2) \wedge \\ \quad \neg ab_{prefer}(A_1) \\ ab_{prefer}(change) \leftarrow \bot \\ ab_{prefer}(donothing) \leftarrow \bot \end{array}$$

Comparing D_1 and $D_1 \circ d \circ D_2$, action A_2 leads to at least one more dead person than action A_1. Hence, A_1 is preferred over A_2 if nothing abnormal is known about A_1.

[4] The first two iterations of $\Phi_{\mathcal{P}}^{\mathcal{E}}$ are shown in detail in the Appendix.

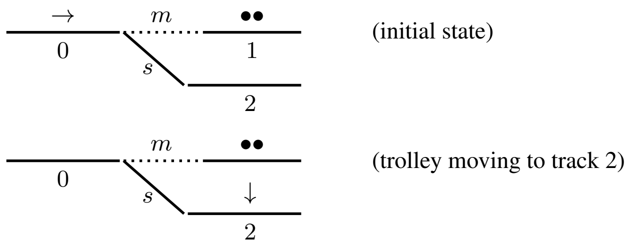

Figure 3: The bystander case (initial state) and its ramifications if Hank is considering the counterfactual.

Under an utilitarian point of view [Bentham2009], the *change* action is preferable to the *donothing* action as it will kill fewer humans. On the other hand, we know that a purely utilitarian view is not allowed in case of human casualties. Hank may ask himself: *Would I still save the humans on the main track if there were no human on the side track and I changed the switch?* This is a counterfactual. But we can easily deal with it in WCSE by starting a new computation with the additional fact

$$causes(change, t_0 \circ c_0 \circ s \circ h_1 \circ h_1 \circ c_2, 1) \leftarrow \top. \quad (4)$$

Comparing (2) and (4), h_2 has been replaced by c_2. There is no human on track 2 anymore and, hence, this track is clear. This is a minimal change necessary to satisfy the precondition of the counterfactual. In this case, the least $\mathcal{E}$-model of the extended program will contain

$$[causes(change, t_0 \circ c_0 \circ s \circ h_1 \circ h_1 \circ c_2, 1)].$$

This case is illustrated in Figure 3. Using

$$\begin{array}{l} permissible(change) \leftarrow \\ \quad prefer(change, donothing) \wedge \\ \quad causes(change, t_2 \circ c_0 \circ s \circ h_1 \circ h_1 \circ c_2, 1) \wedge \\ \quad \neg ab_{permissible}(change) \\ ab_{permissible}(change) \leftarrow \bot \end{array}$$

allows Hank to conclude that changing the switch is permissible within the doctrine of double effect [Aquinas1988].

4 Discussion

We have extended the WCS to WCSE and we have shown how the bystander case can be modeled in the extended approach. We believe that the methods and techniques can be applied to all ethical decision problems discussed in [Pereira and Saptawijaya2016]. In [Dietz Saldanha *et al.*2018] we have already considered the footbridge and the loop case. Moreover, we have applied the doctrine of triple effect [Kamm2006] to distinguish between direct and indirect intentional killings. Currently, we are working out the details for all problems. For us it is important that all these problems can be discussed within the presented framework and are compatable to our solutions for other human reasoning tasks like the suppression and the selection task.

On the other hand, there are many open questions. The examples discussed in this paper are hand-crafted and we would like to develop an extension, where examples taken from the moral machine project (`moralmachine.mit.edu`) can be automatically treated under WCSE. We also would like to generalize the reasoning such that if an action does something good and nothing abnormal is known, then it is permissible. This, however, requires a formalization of 'something good' and very likely a formalization of 'something bad'. And, we should have a closer look at counterfactuals and minimal change.

Acknowledgements I'd like to thank Dominic Deckert, Emmanuelle-Anna Dietz Saldanha, Sybille Schwarz, and Lim Yohanes Stefanus for jointly developing the weak completion semantics with equality.

Appendix

Let $\mathcal{P}$ be the program developed in Section 3 and $\mathcal{E}$ be the set of equations specifying that $\circ$ is associative, commutative, and 1 being its unit element. Let $I_0 = \langle \emptyset, \emptyset \rangle$ be the empty interpretation. Suppose Hank has decided to do nothing. Then,

$$\Phi^{\mathcal{E}}_{\mathcal{P}}(I_0) = I_1 = \langle I_1^\top, I_1^\bot \rangle,$$

where

$$\begin{array}{ll} I_1^\top = \{ & [causes(donothing, t_0 \circ c_0 \circ m \circ h_1 \circ h_1 \circ h_2, 1)], \\ & [action(t_0 \circ c_0 \circ m, 1, downhill, t_1 \circ c_0 \circ m, 1)], \\ & [action(t_0 \circ c_0 \circ s, 1, downhill, t_2 \circ c_0 \circ s, 1)], \\ & [action(t_1 \circ h_1, 1, kill, t_1, d)], \\ & [action(t_2 \circ h_2, 1, kill, t_2, d)] \ \}, \\ I_1^\bot = \{ & [ab(downhill)], \\ & [ab(kill)] \ \}. \end{array}$$

Considering the body of (3) we find that both possible ground instances of $ab(A')$, viz. $ab(downhill)$ and $ab(kill)$, are false under I_1 and, consequently, their negations are true under I_1. The only ground instance of

$$causes(A, P_1 \circ Z_1, P_2 \circ Z_2) \quad (5)$$

being true under I_1 is

$$causes(donothing, t_0 \circ c_0 \circ m \circ h_1 \circ h_1 \circ h_2, 1). \quad (6)$$

Hence, we are searching for a ground instance of

$$action(P_1, P_2, A', E_1, E_2)$$

being true under I_1 such that the ground instance of P_1 is contained in $t_0 \circ c_0 \circ m \circ h_1 \circ h_1 \circ h_2$ and the ground instance of P_2 is contained in 1. There are four candidates in I_1. The only possible ground instance of an action meeting the conditions is

$$action(t_0 \circ c_0 \circ m, 1, downhill, t_1 \circ c_0 \circ m, 1). \quad (7)$$

Comparing the second arguments of (5) and (6) with the first argument of (7) we find that

$$P_1 = t_0 \circ c_0 \circ m$$

and

$$Z_1 = h_1 \circ h_1 \circ h_2.$$

Likewise, comparing the third arguments of (5) and (6) with the second argument of (7) we find that $P_2 = 1$ and $Z_2 = 1$. Combining Z_1 with the fourth argument of (7) and, likewise, combining Z_2 with the fifth argument of (7) we learn that

$$causes(donothing, t_1 \circ c_0 \circ m \circ h_1 \circ h_1 \circ h_2, 1)$$

must be true under $\Phi^{\mathcal{E}}_{\mathcal{P}}(I_1)$.

References

[Aquinas, 1988] T. Aquinas. Summa Theologica II-II, q. 64, art. 7, "Of Killing". In W. P. Baumgarth and R.J. Regan, editors, *On Law, Morality, and Politics*, pages 226–227. Hackett Publishing Co., Indianapolis, 1988.

[Barrett, 2010] L. Barrett. *An Architecture for Structured, Concurrent, Real-time Action*. PhD thesis, Computer Science Division, University of California at Berkeley, 2010.

[Bentham, 2009] J. Bentham. *An Introduction to the Principles of Morals and Legislation*. Dover Publications Inc., 2009.

[Byrne, 1989] R. M. J. Byrne. Suppressing valid inferences with conditionals. *Cognition*, 31:61–83, 1989.

[Clark, 1978] K. L. Clark. Negation as failure. In H. Gallaire and J. Minker, editors, *Logic and Databases*, pages 293–322. Plenum, New York, 1978.

[Dietz *et al.*, 2012] E.-A. Dietz, S. Hölldobler, and M. Ragni. A computational logic approach to the suppression task. In N. Miyake, D. Peebles, and R. P. Cooper, editors, *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pages 1500–1505. Cognitive Science Society, 2012.

[Dietz *et al.*, 2013] E.-A. Dietz, S. Hölldobler, and M. Ragni. A computational logic approach to the abstract and the social case of the selection task. In *Proceedings Eleventh International Symposium on Logical Formalizations of Commonsense Reasoning*, 2013. `commonsensereasoning.org/2013/proceedings.html`.

[Dietz Saldanha *et al.*, 2017] E.-A. Dietz Saldanha, S. Hölldobler, and I. Lourêdo Rocha. The weak completion semantics. In C. Schon and U. Furbach, editors, *Proceedings of the Workshop on Bridging the Gap between Human and Automated Reasoning – Is Logic and Automated Reasoning a Foundation for Human Reasoning?*, volume 1994, pages 18–30. CEUR-WS.org, 2017. `http://ceur-ws.org/Vol-1994/`.

[Dietz Saldanha *et al.*, 2018] E.-A. Dietz Saldanha, S. Hölldobler, S. Schwarz, and L. Y. Stefanus. The weak completion semantics and equality. In *Proceedings of the 22nd International Conference on Logic for Programming, Artificial Intelligence, and Reasoning (LPAR-22)*. EPiC series in Computing, 2018. (to appear).

[Fitting, 1985] M. Fitting. A Kripke–Kleene semantics for logic programs. *Journal of Logic Programming*, 2(4):295–312, 1985.

[Hölldobler and Jovan, 2014] S. Hölldobler and F. Jovan. Advanced Petri nets and the fluent calculus. In S. Hölldobler, A. Malikov, and C. Wernhard, editors, *Proceedings of the Young Scientists' International Workshop on Trends in Information Processing*, volume 1145 of *CEUR Workshop Proceedings*, pages 15–24. CEUR-WS.org, 2014. `http://ceur-ws.org/Vol-1145/`.

[Hölldobler and Kencana Ramli, 2009] S. Hölldobler and C. D. P. Kencana Ramli. Logic programs under three-valued Łukasiewicz's semantics. In P. M. Hill and D. S. Warren, editors, *Logic Programming*, volume 5649 of *Lecture Notes in Computer Science*, pages 464–478. Springer-Verlag Berlin Heidelberg, 2009.

[Hölldobler and Schneeberger, 1990] S. Hölldobler and J. Schneeberger. A new deductive approach to planning. *New Generation Computing*, 8:225–244, 1990.

[Hölldobler, 2015] S. Hölldobler. Weak completion semantics and its applications in human reasoning. In U. Furbach and C. Schon, editors, *Bridging 2015 – Bridging the Gap between Human and Automated Reasoning*, volume 1412 of *CEUR Workshop Proceedings*, pages 2–16. CEUR-WS.org, 2015. `http://ceur-ws.org/Vol-1412/`.

[Jaffar *et al.*, 1984] J. Jaffar, J-L. Lassez, and M. J. Maher. A theory of complete logic programs with equality. In *Proceedings of the International Conference on Fifth Generation Computer Systems*, pages 175–184. ICOT, 1984.

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

[Khemlani and Johnson-Laird, 2012] S. Khemlani and P. N. Johnson-Laird. Theories of the syllogism: A meta-analysis. *Psychological Bulletin*, 138(3):427–457, 2012.

[Łukasiewicz, 1920] J. Łukasiewicz. O logice trójwartościowej. *Ruch Filozoficzny*, 5:169–171, 1920. English translation: On Three-Valued Logic. In: *Jan Łukasiewicz Selected Works*. (L. Borkowski, ed.), North Holland, 87-88, 1990.

[Nickerson, 2015] R. S. Nickerson. *Conditional Reasoning*. Oxford University Press, 2015.

[Oliviera da Costa *et al.*, 2017] A. Oliviera da Costa, E.-A. Dietz Saldanha, S. Hölldobler, and M. Ragni. A computational logic approach to human syllogistic reasoning. In G. Gunzelmann, A. Howes, T. Tenbrink, and E. J. Davelaar, editors, *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, pages 883–888, Austin, TX, 2017. Cognitive Science Society.

[Pereira and Saptawijaya, 2016] L. M. Pereira and A. Saptawijaya. *Programming Machine Ethics*. Springer, Berlin, Heidelberg, 2016.

[Stenning and van Lambalgen, 2008] K. Stenning and M. van Lambalgen. *Human Reasoning and Cognitive Science*. MIT Press, 2008.

[Thielscher, 2003] M. Thielscher. Controlling semi-automatic systems with FLUX (extended abstract). In C. Palamidessi, editor, *Logic Programming*, volume 2916 of *Lecture Notes in Computer Science*, pages 515–516. Springer, Berlin, Heidelberg, 2003.

[Wason, 1968] P. C. Wason. Reasoning about a rule. *The Quarterly Journal of Experimental Psychology*, 20:273–281, 1968.