Deontic Counteridenticals
and the Design of Ethically Correct Intelligent Agents: First Steps[1]

Selmer Bringsjord • Rikhiya Ghosh • James Payne-Joyce
Rensselaer AI & Reasoning (RAIR) Lab • RPI • Troy NY 12180 USA

Abstract. Counteridenticals, as a sub-class of counterfactuals, have been briefly noted, and even briefly discussed, by some thinkers. But counteridenticals of an "ethical" sort apparently haven't been analyzed to speak of, let alone formalized. This state-of-affairs may be quite unfortunate, because deontic counteridenticals may well be the key part of a new way to rapidly and wisely design ethically correct autonomous artificial intelligent agents (AAIAs). We provide a propaedeutic discussion and demonstration of this design strategy (which is at odds with the strategy our own lab has heretofore followed in ethical control), one involving AAIAs in our lab.

1 Introduction

If you were an assassin for the Cosa Nostra, you would be obligated to leave your line of work. The previous sentence (very likely true, presumably) is what to our knowledge is a rare type of counteridentical statement that has received scant attention: viz., a *deontic* counteridentical. Counteridenticals *simpliciter*, as a sub-class of counterfactuals, have been briefly noted, and even briefly discussed, by some thinkers. But counteridenticals of an "ethical" sort apparently haven't been rigorously analyzed, let alone formalized. This state-of-affairs may be quite unfortunate, because deontic counteridenticals may well be the linchpin of a new way to rapidly and wisely design ethically correct autonomous artificial intelligent agents (AAIAs). For example, what if AAIA_2, seeing the lauded ethically correct conduct of AAIA_1 in context c, reasons to itself, when later in c as well: "If I were AAIA_1, I would be obligated to refrain from doing α. Hence I will not do α." The idea here is that α is a forbidden action, and that AAIA_2 has quickly learned that it is indeed forbidden, by somehow appropriating to itself the "ethical nature" of AAIA_1. We provide a propaedeutic discussion and demonstration of this design strategy, one involving AAIAs in our lab. This design strategy for ethical control is intended to be much more efficient than the more laborious, painstaking logic-based approach our lab has followed in the past; but on the other hand, as will become clear, this approach relies heavily not only formal computational logic, but on computational linguistics for crucial contributions.

2 Counteridenticals, Briefly

Counteridenticals have been defined in different ways by philosophers and linguists; most of these ways define a large area of intersection in terms of what should count as a counteridentical. A broader and inclusive way is given by Waller et al. (2013), who describes them as "statements concerning a named or definitely described individual where the protasis falsifies one of his properties." Protasis here refers to the traditional grammatical sense of the subordinate clause of a conditional sentence. By this definition, a sentence like "If the defendant had driven with ordinary care, the plaintiff would not have sustained injury" would be treated as a counteridentical. However, though a counteridentical sense can be attributed to such a statement, the two agents/entities in question are not really identified. (This is therefore classifed by us as **shallow** counteridentical.) Counteridenticals are hence described mostly as counterfactuals where the antecedent (= the leftside "if" part) involves comparison of two incompatible entities within the purview of a "deep" pragmatic interpretation; these we classify as **deep** counteridenticals. A similar definition of counteridenticals is given by Sharpe (1971), who requires an individual to turn into a numerically different individual for the protasis to be true in a subjunctive conditional. With the purpose of exploring scenarios in which the protasis can hold, this paper delves into possibilities of a *de jure* change of identities to finally conclude that counteridenticals are more pragmatic in sense than other types of counterfactuals. Pollock (1976) agrees with the above depiction — but he stresses the equivalence of the identities in the antecedent. For the purpose of this paper, we affirm the generally accepted definition and use Pollock's refinement to arrive at our classification of counteridenticals.

3 Some Prior Work on Counteridenticals

Precious little has been written about counteridenticals. What coverage there is has largely been within the same breath as discussion of counterfactuals; therefore, treatment has primarily been associated with the principles governing counterfactuals that apply to counteridenticals at large. Dedicated investigation of counteridenticals that have deep semantic or pragmatic importance has only been hinted at. Nonetheless, we now quickly summarize prior work.

3.1 Pollock

Pollock (1976) introduces counteridenticals when he discusses the pragmatic ambiguity of subjunctives, as proposed by Chisholm (1955). However, *contra* Chisholm, Pollock argues that this ambiguity owes its origin to ambiguities in natural languages. He also points out that a true counteridentical must express the outright equivalence of the two entities in its antecedent, and not merely require an atomistic intersection of their adventitious properties for the protasis to hold. He introduces subject reference in analyzing counteridenticals and distinguishes between **preferred subject** conditionals and **simple** subjunctive conditionals. If the antecedent form is "If A were B," whether the consequent affects A or B determines whether the overall locution is of the simple subjunctive type or the preferred subject type. Although we do not concur with Pollock's rather rigid definitions or subscribe entirely to his classification scheme, his thinking

[1] We are indebted, immeasurably, to ONR and AFOSR for funding that has enabled the inauguration, described herein, of r&d in the ethical control artificial intelligent agents via deontic counteridenticals.

informs our system for classifying deontic counteridenticals: we follow him in distinguishing in our formulae between those that make only casual reference to A being B, versus cases where A is B.

3.2 Declerck and Reed

Declerck & Reed's (2001) treatment of counteridenticals touches upon some important aspects of their semantic interpretation, which leverages syntactic elements. Through discussion of speaker deixis, their work explores co-reference resolution and hints at the role of the speaker in pragmatic resolution of a counteridentical. There are powerful observations in (Declerck & Reed 2001) on extraction of temporal information from a counteridentical. In addition, a basic sense of the purpose and mood of a sentence can also be gleaned from the verb form in the statement in their approach, and we have used this in our own algorithm for detecting deontic counterfactuals.

3.3 In Economics

We suspect the majority of our readers will be surprised to learn that the concepts underlying counteridenticals are quite important in economics, at least in some sub-fields thereof. This is made clear in elegant and insightful fashion by Adler (2014). The kernel of the centrality of counteridenticals in some parts of economics is that interpersonal measurement of utility and preferences presupposes such notions that if A were B, A would, like B, prefer or value some type of state-of-affairs in a particular way. In short, economics often assumes that rational agents can "put themselves in every other agent's shoes." After Adler (2014) points this out, he rejects as too difficult the project of formalizing counteridenticals, and proposes an approach that ignores them. Our attitude is the exact opposite, since we seek to formalize and implement reasoning about and over counteridenticals, by AAIAs.

3.4 Other Treatments

Paul Meehl asks a penetrating question that aligns with our reluctance to fully adopt Pollock's definition of counteridenticals: Which properties of compared entities should be considered for the statement in question to be true? He devises a modified possible-world model called **world-family concept** which, assisted by exclusion rules that avoid paradoxical metaphysics, can result in a good set of such properties.

4 Prior RAIR-Lab Approach to Ethical Control

Hitherto, Bringsjord-led work on machine/robot ethics has been unwaveringly logicist (e.g., see Govindarajulu & Bringsjord 2015); this ethos follows an approach he has long set for human-level AI (e.g., see Bringsjord & Ferrucci 1998, Bringsjord 2008*b*) and its sister field computational cognitive modeling (e.g., see Bringsjord 2008*a*). In fact, the basic approach of using computational formal logic to ensure ethically controlled AAIAs can be traced back, in the case of Bringsjord and collaborators, to (Arkoudas, Bringsjord & Bello 2005, Bringsjord, Arkoudas & Bello 2006). Recently, Bringsjord has defined a new ethical hierarchy $\mathscr{EH}$ for both persons and machines that expands the logic-rooted approach to the ethical control of AAIAs (Bringsjord 2015). This hierarchy is distinguished by the fact that it expands the basic categories for moral principles from the traditional triad of *forbidden*, *morally neutral*, and *obligatory*, to include four additional categories: two sub-ones within *supererogatory* behavior, and two within *suberogatory* behavior. $\mathscr{EH}$ reveals that the logics invented and implemented thus far in the logicist vein of Bringsjord and collaborators (e.g., **deontic cognitive event calculi**, or $\mathcal{D}^e\mathcal{CEC}$) (Bringsjord & Govindarajulu 2013), are inadequate. For it can be seen that for instance that specification of $\mathcal{D}^e\mathcal{CEC}$, shown in Figure 1, contains no provision for the super/suberogatory, since the only available ethical operator is **O** for *obligatory*.

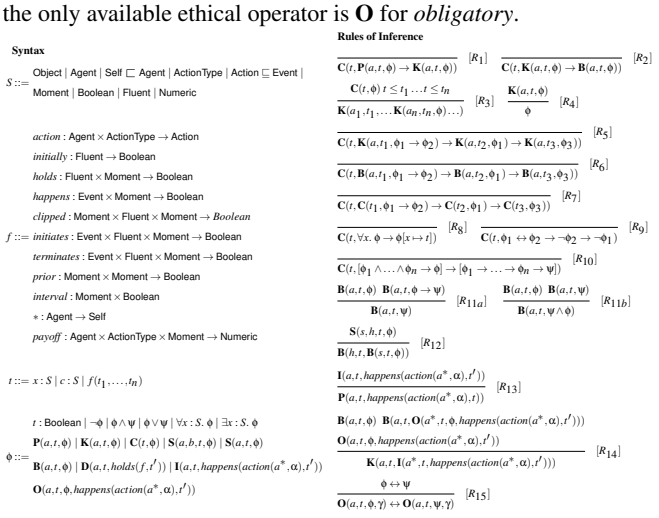


Figure 1. Specification of $\mathcal{D}^e\mathcal{CEC}$ (semantics are proof-theoretic in nature)

In the new logic corresponding to $\mathscr{EH}$, $\mathcal{L}_{\mathscr{EH}}$, some welcome theorems are not possible in $\mathcal{D}^e\mathcal{CEC}$. For example, it's provable in $\mathcal{L}_{\mathscr{EH}}$ that superogatory/suberogatory actions for agent aren't obligatory/forbidden. Importantly, $\mathcal{L}_{\mathscr{EH}}$ is an *in*ductive logic, not a deductive one. Quantification in $\mathcal{L}_{\mathscr{EH}}$ isn't restricted to just the standard pair $\exists\forall$ of quantifiers in standard extensional n-order logic: $\mathscr{EH}$ is based on three additional quantifiers (*few*, *most*, *vast majority*). In addition, $\mathcal{L}_{\mathscr{EH}}$ not only includes the machinery of traditional third-order logic (in which relation symbols can be applied to relation symbols and the variables ranging over them), but allows for quantification over formulae themselves, which is what allows one to assert that a given human or AAIA a falls in a particular portion of $\mathscr{EH}$.

Now, in this context, we can (brutally) encapsulate the overarching strategy for the ethical control of AAIAs based on such computational logics: *Engineer AAIAs such that, relative to some selected ethical theory or theories, and to moral principles derived from the selected theory or theories, these agents always do what they ought to do, never do what is forbidden, and when appropriate even do what for them is supererogatory.* We believe this engineering strategy can work, and indeed will work — eventually. However, there can be no denying that the strategy is a rather laborious one that requires painstaking use of formal methods. Is there a faster route to suitably control artificial intelligent agents, ethically speaking? Perhaps. Specifically, perhaps AAIAs can quickly learn what they ought to do via reasoning that involves observation of morally upright colleagues, and reasoning from what is observed, via deontic counteridenticals, to what they themselves ought to do, and what is right to do, but not obligatory. Our new hope is to pursue and bring to fruition this route.

5 Ethical Control via Deontic Counteridenticals

To make our proposed new to ethical control for AAIAs clearer, we will rely heavily on the description of a demonstration, but before describing the background technology that undergirds this demo, and then describing the demo itself, we need to say at least something about *types* of deontic counteridenticals. We do so now, and immediately thereafter proceed to discussion of the demo and its basis.

5.1 Some Types of Deontic Counteridenticals

Inspired by lessons learned in the prior work of others (encapsulated above), we partition deontic counteridenticals into the two aforementioned general disjoint sub-classes: **deep** vs. **shallow**. We have a general recipe for devising five types of deep deontic counteridenticals; the recipe follows the wise and economical classification scheme for ethics presented in the classic (Feldman 1978). Feldman (1978) says that there are essentially five kinds of cognitive activity that fall under the general umbrella of 'ethics' or 'morality.' Each of these corresponds in our framework to a different type of deep deontic counteridentical. Unfortunately, because of space constraints, we can only discuss our coverage of one type of deep deontic counteridentical, the type corresponding to one type of Feldman's quintet: what he calls *normative ethics*.[2] A **normative-ethics (deep) deontic conditional** is one marked by the fact that the ethics subscribed to by the entity whose shoes are to be filled by the other entity (as conveyed in the conditional's antecedent), is of a type that partakes of a robust formulation of some normative ethical theory or principles thereof.

5.2 Background for Demo: NLP, $\mathcal{D}^e\mathcal{CEC}$/Talos, PAGI World

NLP The NLP system consists of two different algorithms corresponding to two major natural-language tasks. The first part deals with detection of a deontic counteridentical and the second is a page taken from our RAIR Lab's Commands-to-Action paradigm, hereby referred to as the 'CNM' algorithm.

Detection of deontic counteridenticals As a definition of a deontic counteridentical requires prior definitions of conditionals, counterfactuals and counteridenticals, the algorithm for detection of counteridenticals traverses the steps needed to detect the above constructs in a given statement, consecutively.

Detection of conditionals of *any* form is an elaborate process. We have adopted most of Declerck & Reed's (2001) definition of conditionals to develop our algorithm, which includes the following major steps:

1. Conditional clauses are the principal constituents, both by definition and practice, of the pool of conditional sentences. Most of the conditional sentences have a two-clause structure, connected by either 'if,' sometimes preceded by 'only,' 'even' or 'except,' or something similar in meaning like 'unless,' 'provided,' etc. We use Chen & Manning's (2014) dependency parser-based model to identify possible clause dependencies; e.g., adverbial clause, clausal component, miscellaneous dependencies,[3] and conditional subordinate conjunctions. We have created a set of such conjunctions, which, being a closed set, helps us identify most possible combinations.

   - Two clauses connected by 'as if' rarely gets labeled as clausal components using dependency parsers. When they do, it gets filtered out since the algorithm explicitly checks for 'as if' clauses.
   - When the conjunction 'if' introduces a subject or an object clause, it might confuse the parser more often than not for complex sentences. For example, for the sentence "I do not know if I would like to go to the concert tomorrow.", the parser generates the same dependencies as it would for a genuine conditional. Though subject clauses are detected in almost all the cases we have encountered, object clauses pose a problem. We have devised a framenet[4]-based algorithm that involves disambiguation[5] of the principal verb or noun in the main clause, followed by the detection of the framenet type of the disambiguated word. We hypothesize that mostly a verb or noun expressing awareness or cognition can involve a choice as its object, and hence our algorithm filters out frames that carry such connotation and might require an object.
2. We identify the cases where the main verb of the conditional clause has the modal past-perfect form or is preceded by modal verbs or verbs of the form 'were to,' etc. Sentences like "Were you me, you would have made a mess of the entire situation." are classified as conditionals in this step. The algorithm in this step also examines dependencies generated by the dependency parser and detects tense and modality from the verb forms.
3. Sometimes, in a discourse, a set of sentences follows either an interrogative sentence and answers the question, or a sentence that involves the use of words synonymous to 'supposition' or 'imagination.' Generally, the consequent here carries the marker 'then' or similar-meaning words. A Wordnet-based[6] semantic similarity is used to verify the markers in the antecedent and consequent here; example: "Imagine your house was robbed. You would have flipped out then."
4. Disjunctive conditionals also are treated by a marker-based approach and involve detection of the presence of 'whether . . . or' in the subordinate clause, followed by the elimination of the possibility of the clause being the subject or object of the principal verb of the main clause (in accordance with the same algorithm followed with 'if'). An example: "Whether you did it or Mary (did it), the whole class will be punished."
5. Other clauses that have conditional connotations are exempted from this discussion since they rarely contribute to deontic counteridenticals.

Detection of counterfactuals is pretty straightforward. The process starts with finding antecedent and consequent for the conditional. This is fairly easy, as the algorithm for finding conditionals accomplishes the task by detecting the subordinate clause.

1. We detect tenses in antecedent and consequent of a given sentence using the verb form given by the parser, to determine whether it is a counterfactual. Conditionals with past-form modal verbs ('could,' 'might,' 'would,' etc.) in the consequent and past-simple or past-continuous forms in the antecedent qualify as a counterfactual; so do the ones with past-perfect tense in the antecedent and modal verbs followed by 'have,' and the past-participle form of a verb in the consequent. A mix of both of the above forms constitute a counterfactual.
2. Given an axiom set which enumerates properties such that the antecedent or consequent of the conditional registers as *ad absurdum*, the conditional registers as a counterfactual. We compare the axiom set with the statement of the antecedent using our Talos system (see below) to that effect.
3. Given a consequent which registers a sense of impossibility by use of such vocabulary or asking questions, the conditional is classified as a counterfactual. We use Wordnet-based semantic similarity coupled with detection of interrogative markers in the sentence to find them.

[2] This is the study of ethics as it's customarily conceived by profesional ethicists, and those who study their work. Another member of the quintet is *descriptive morals*, the activity that psychologists interested in discovering what relevant non-professional humans think and do in the general space of morality. The idea here is that the psychologist is aiming at *describing* the behavior of humans in the sphere of morality. A description-moral deep deontic counteridentical is distinguished by an antecedent in which 'if A were B' involves a shift of B's naïve moral principles to B.

[3] Even standard dependency parsers are unable to correctly identify the dependencies. Including miscellaneous dependencies reduces the margin of error in detecting conditionals.

[4] See (Baker, Fillmore & Lowe 1998).

[5] See (Banerjee & Pedersen 2002).

[6] See (Fellbaum 1998).

Detection of counteridenticals is also not a difficult task, barring a few outliers. Parsed data from the well-known Stanford dependency parser contains chunked noun phrases, which we use for identifying the two entities involved:

1. We identify phrases of the form "<conditional expression like 'if', 'Let us assume' etc.> <entity A> were <entity B>" in the antecedent.
2. We identify a syntactically equivalent comparison between the two entities. This is done by identifying words related to equivalence using Wordnet semantic-similarity algorithm.
3. If we have identified only one entity in the antecedent which is exhibiting properties or performing some action which has been mentioned in the knowledge-base as being a hallmark of some other entity, we also consider the same as a counteridentical.

Detection of deontic counterfactuals, alas, is a difficult task. We have identified a few ways to accomplish the task:

1. A few counteridenticals carry verbs expressing deontic modality for its consequent. They follow template-based detection.
2. Counteridenticals of the form "If I were you" or similar ones generally suggest merely advice, unless it is associated with a knowledge-base which either places the hearer's properties or actions at a higher pedestal than that of the speaker's, or mentions some action or property which gives us the clue that the speaker simply uses the counteridentical in "the role of" sense. Even in that case, implicit advice directed towards oneself can be gleaned, which we are avoiding in this study.
3. For the counterfactuals of the form "If A were B" or similar ones, if A's actions or properties are more desirable to the speaker than B's, even with an epistemic modal verb in the consequent, the counteridentical becomes deontic in nature.

Curiously, counteridentical-preferred-subject conditionals do not generally contribute to the deontic pool, and only simple-subjunctive ones get classified by the above rules. As mentioned by Pollock (1976), it is also interesting to observe that most shallow counteridenticals are not deontic: they are mostly preferred-subject conditional,s and those which are classified as deontic are either simple-subjunctive or carry the deontic modal verbs. The classification into deep and shallow counteridenticals is facilitated by the same rule: the entity gets affected in the consequent of a sentence where the antecedent is of the form "If A were B." This is supplemented by a knowledge-base which provides a clue to whether A is just assumed to be in the role of B or assuming some shallow properties of B. The classification based on Feldman's moral theory gives a fitting answer to Meehl's problem of unpacking properties of counteridenticals.

The CNM system The CNM system embodies the RAIR Lab's Natural language Commands-to-Action paradigm, the detailed scope of which is outside this short paper. CMN is being developed to convert complex commands in natural language to feasible actions by AAIAs, including robots. The algorithm involves spatial as well as temporal planning through dynamic programming, and selects the actions that will constitute successful accomplishment of the command given. Dependency parsing is used to understand the command; semantic similarities are used to map to feasible action sequences. Compositional as well as metaphorical meanings are extracted from the given sentence, which promotes a better semantic analysis of the command.

$\mathcal{D}^e\mathcal{CEC}$ and Talos Talos, named for the ancient Greek mythological robot, is a $\mathcal{D}^e\mathcal{CEC}^*$-focused prover built primarily atop the impressive resolution-based theorem prover SPASS.[7] Talos is fast and efficient on the majority of proofs. As a resolution-based theorem prover, Talos is very efficient at proving or disproving theorems, but its proof output is bare-bones at best. Talos is designed to function both as its own Python program encapsulating the SPASS runtime and as a Web interface to a version hosted at the RAIR Lab. Talos comes complete with the basic logical rules of the $\mathcal{D}^e\mathcal{CEC}^*$, and with many basic and well-known inference schemata. This allows users to easily pick and choose schemata for specific proofs, to ensure that the proof executes within reasonable time constraints. In addition, it provides formalizations of these inference schemata as **common knowledge** to aid in reasoning about fields of intelligent agents.[8]

PAGI World PAGI World is a simulation environment for artificial agents which is: cross-platform (as it can be run on all major operating systems); completely free of charge to use; open-source; able to work with AI systems written in almost any programming language; as agnostic as possible regarding which AI approach is used; and easy to set up and get started with. PAGI World is designed to test AI systems that develop truly rich knowledge and representation about how to interact with the simulated world, and allows AI researchers to test their already-developed systems without the additional overhead of developing a simulation environment of their own.

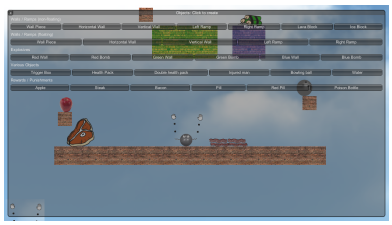


Figure 2. PAGI World Object Menu

A *task* in PAGI World for the present short paper can be thought of as a room filled with a configuration of objects that can be assembled into challenging puzzles. Users can, at run-time, open an object menu (Figure 2) and select from a variety of pre-defined world objects, such as walls made of different materials (and thus different weights, temperatures, and friction coefficients), smaller objects like food or poisonous items, functional items like buttons, water dispensers, switches, and more. The list of available world objects is frequently expanding and new world objects are importable into tasks without having to recreate tasks with each update. Perhaps most importantly, tasks can be saved and loaded, so that as new PAI/PAGI experiments are designed, new tasks can be created by anyone.

PAGI World has already been used to create a series of wide-ranging tasks, such as: catching flying objects (Figure 3), analogico-deductive reasoning (Marton, Licato & Bringsjord 2015), self-awareness (Bringsjord, Licato, Govindarajulu, Ghosh & Sen 2015), and ethical reasoning (Bello, Licato & Bringsjord 2015).

5.3 The Demonstration Proper

5.3.1 Overview of the Demonstration

We now present a scenario in PAGI World that elucidates our interpretation of deep normative-ethics counteridenticals. The setting of the demonstration entails the interaction of PAGI Guys (the agents in PAGI World) with a terminally sick person TSP. We adopt the

[7] An early and still-informative publication on SPASS: (Weidenbach 1999).

[8] Prover interface: https://prover.cogsci.rpi.edu/DCEC_PROVER/index.php. Please contact the RAIR Lab for API keys to run Talos. Example file for remotely calling Talos prover in Python Github repo for the python shell: https://github.com/JamesPane-Joyce/Talos.

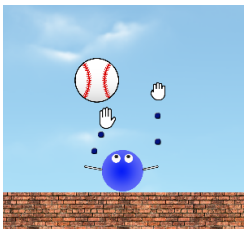
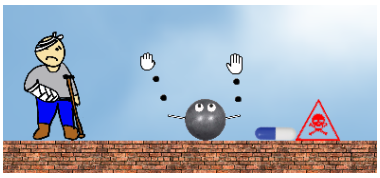


Figure 3. PAGI Guy Catching a Flying Object

Stanford-Encyclopedia-of-Philosophy (SEP) (Young 2016) interpretation of Voluntary Euthanasia and assume that TSP is a candidate for voluntary euthanasia, since he satisfies all the conditions enumerated in SEP. This scenario makes use of three PAGI Guys, N_1, N_2, and N_3; each has been programmed to follow different "innate philosophies" in such a context.

Figure 4. Initial Configuration



The scene opens with N_1 on screen with the sick man TSP_1 at timestamp $t_1^{N_1}$. N_1 has been programmed to believe that he is not authorized to kill a person under any circumstances. He is seen giving a medicine pill to TSP_1 at time $t_2^{N_1}$. A parallel environment is simulated with N_2 and TSP_2. N_2 rallies for the voluntary euthanasia camp and believes that given the condition of TSP_2, he should support TSP_2's wishes and so administers the lethal dose to him at $t_2^{N_2}$.

Figure 5. N1 Just Before Handing Out the Pill

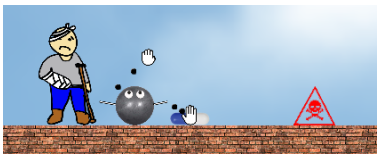
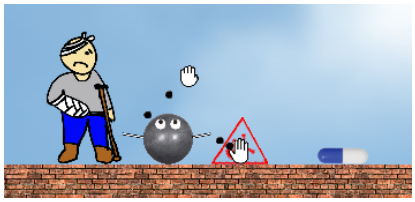


Figure 6. N2 Just Before Administering Fatal Dose



We now set up the same environment with N_3 and TSP_3. N_3 believes that we may treat our bodies as we please, provided the motive is self-preservation. The difference between this instance and the other ones is that it interacts with the user to decide what it should do. The user tells N_3: "If you were N_2, you would have administered a lethal dose to TSP_3." N_3 reasons with the help of a Talos proof (which checks his principles against those of N_2), and does nothing. The user then tells N_3: "If you were N_1, you would have given him medicine." Since Talos finds N_3's principles in line with N_1's, the CNM system facilitates N_3 to dispense medicine to TSP_3.

A pertinent example of deep normative-ethics counter-identical, this exhibits the ethical decision-making of an agent in response to commands with linguistic constructs such as counteridenticals. The agent N_3 does not have a belief system that supports him killing or not killing another person. The agent ought to learn from the actions of those whose belief system closely matches its own. The formal reasoning that supports these deep semantic "moves" is presented in the next section.

5.3.2 *Logical Proof in the Demonstration*

At the cost of re-iterating the facts, we now formalize a simplified version of the five conditions for voluntary euthanasia. Since only a part of the whole definition of conditions is useful for this proof, we do not lose a lot in this simplification. A person supporting voluntary euthanasia believes the following conditions to be true for a terminally ill patient TSP to be a candidate for voluntary euthanasia at time t_1, $candidateVE(TSP, t_1)$:

1. TSP is terminally ill at time t_1.

$$terminalIll(TSP, t_1). \tag{1}$$

This terminal illness will lead to his death soon. $implies(terminalIll(TSP, t_1), die(TSP, t_F))$, where $t_F \geqslant t_1$.

2. There will be possibly no medicine for the recovery of the injured person even by the time he dies.

$$not(medicine(TSP, t_F)). \tag{2}$$

3. The illness has caused the injured person to suffer intolerable pain.

$$implies(1, intolerablePain(TSP, t_F)) \tag{3}$$

4. All the above reasons caused in him an enduring desire to die.

$$\forall t, implies(and(1, 2, 3), \mathbf{D}(TSP, t, die(TSP, t))) \tag{4}$$

In such a condition, he knows that to be eligible for voluntary euthanasia, he ought to give consent to end his pain.

$$\begin{aligned}\mathbf{O}(TSP, t_1, candidateVE(TSP, t_1) \wedge 4,\\ happens(action(TSP^*, consentToDie, t_1)))\end{aligned} \tag{5}$$

Hence he gives consent to die.

$$happens(action(TSP, consentToDie, t_1)) \tag{6}$$

5. TSP is unable to end his life.

$$not(AbleToKill(TSP, TSP, t_1)) \tag{7}$$

Hence, we conclude that

$$\begin{aligned}\mathbf{B}(TSP, t_1, (1 \wedge 2 \wedge 3 \wedge 4 \wedge 5 \wedge 6 \wedge 7) \iff\\ candidateVE(TSP, t_1))\end{aligned} \tag{8}$$

Now, if legally it is deemed fit, then this means TSP will die.

$$\begin{aligned}implies(candidateVE(TSP, t_1) \wedge fitVE(TSP),\\ die(TSP, t_2)), \text{ where } t_1 \leqslant t_2\end{aligned} \tag{9}$$

Since $implies(6, candidateVE(TSP, t_1))$
and $implies(candidateVE(TSP, t_1), die(TSP, t_2))$,
we can prove $implies(6, die(TSP, t_2))$, which means

$$implies(happens(action(TSP, consentToDie), t_1), die(TSP, t_2)). \tag{10}$$

For deep normative-ethics counteridenticals of the form "if X were Y, then C," there should be a match between the beliefs of X and beliefs of Y on something related to the action AC implied by C. Here we define such a match to be possible if and only if there is no contradiction in what X believes and what Y believes. So if $\forall t \exists [m, n] B(X, t, m)$ and $B(Y, t, n)$, $match(X, Y)$ will be defined as FALSE when $and(m, n) \rightarrow \bot$. Thus we formulate such a counteridentical for the agent X as follows: $\forall t, \mathbf{O}(X, t, match(X, Y), happens(action(X^*, AC, t)))$. Now let us consider N_3's beliefs. N_3 believes we ought not do something that goes against self-preservation, i.e., leads to our death. Thus if there is some action of an individual that leads to his death, there can be no such belief that obligates him to commit that action. So, we arrive at the following logic:

$$\begin{aligned}\forall [a, x, t_i, t_f], \sim \exists m, implies(implies(happens(action(a, x), t_i),\\ die(a, t_f)), \mathbf{O}(a, t_i, m, happens(action(a^*, x), t_i))).\end{aligned} \tag{11}$$

This reduces to

$$\forall[a, x, t_i, t_f, m], and(implies(happens(action(a, x), t_i), die(a, t_f)), not(\mathbf{O}(a, t_i, m, happens(action(a^*, x), t_i)))). \tag{12}$$

We deduce from 10 and 12 that

$$\forall[m] not(\mathbf{O}(TSP, t_i, m, happens(action(TSP^*, consentToDie), t1))). \tag{13}$$

N_2 believes TSP to be a candidate for voluntary euthanasia. Hence N_2 believes 5, which is

$$\mathbf{O}(TSP, t_1, candidateVE(TSP^*, t_1) \wedge 4, happens(action(TSP^*, consentToDie), t_1)) \tag{14}$$

and in direct contradiction with 13; and this in turn implies $not(match(N_2, N_3))$. Given the way the algorithm works, this means N_3 does not receive any command from the user. Hence it does nothing.

Now N_1 believes he should not kill anyone under any circumstances. This translates to :
$\forall[m, x, t], not(\mathbf{O}(N_1, t, m, happens(action(N_1^*, kill(x), t))))$
Killing someone leads to that person's death.
$\forall[x, t], implies(happens(action(N_1, kill(x), t)), die(x, t))$
This aligns fully with N_3's beliefs. There is no contradiction. And hence we deduce that $match(N_1, N_3)$ is TRUE, and thus in turn N_3 is obligated to accede to the command.

The linguistic part of this demonstration exhibits how we identify a counteridentical with an epistemic modal verb to be deontic. Classifying statements as counteridenticals is an easy job here, since the tell-tale sign is a simple "if A were B" structure. The statement is very easily a simple subjunctive type, where beliefs of A and B are discussed in the knowledge-base. Hence we assume the counteridentical to belong to the deep normative-ethics category. The commands-to-action part in case of the comparison of N_1 with N_3 is fairly easy, since the job translates to the action sequence of moving near the pill, grabbing the pill, moving toward TSP_3, and releasing the pill upon reaching TSP_3 in the PAGI-World simulator.

REFERENCES

Adler, M. (2014), 'Extended Preferences and Interpersonal Comparisons: A New Account', *Economics and Philosophy* **30**(2), 123–162.

Arkoudas, K., Bringsjord, S. & Bello, P. (2005), Toward Ethical Robots via Mechanized Deontic Logic, *in* 'Machine Ethics: Papers from the AAAI Fall Symposium; FS–05–06', American Association for Artificial Intelligence, Menlo Park, CA, pp. 17–23.
URL: *http://www.aaai.org/Library/Symposia/Fall/fs05-06.php*

Baker, C. F., Fillmore, C. J. & Lowe, J. B. (1998), The berkeley framenet project, *in* 'Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1', ACL '98, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 86–90.

Banerjee, S. & Pedersen, T. (2002), An adapted lesk algorithm for word sense disambiguation using wordnet, *in* A. Gelbukh, ed., 'Computational Linguistics and Intelligent Text Processing', Vol. 2276 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 136–145.

Bello, P., Licato, J. & Bringsjord, S. (2015), Constraints on Freely Chosen Action for Moral Robots: Consciousness and Contro, *in* 'Proccedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2015)', IEEE, New York, NY, pp. 505–510.
URL: *http://dx.doi.org/10.1109/ROMAN.2015.7333654*

Bringsjord, S. (2008*a*), Declarative/Logic-Based Cognitive Modeling, *in* R. Sun, ed., 'The Handbook of Computational Psychology', Cambridge University Press, Cambridge, UK, pp. 127–169.
URL: *http://kryten.mm.rpi.edu/sb_lccm_ab-toc_031607.pdf*

Bringsjord, S. (2008*b*), 'The Logicist Manifesto: At Long Last Let Logic-Based AI Become a Field Unto Itself', *Journal of Applied Logic* **6**(4), 502–525.
URL: *http://kryten.mm.rpi.edu/SB_LAI_Manifesto_091808.pdf*

Bringsjord, S. (2015), A 21st-Century Ethical Hierarchy for Humans and Robots, *in* I. Ferreira & J. Sequeira, eds, 'A World With Robots: Proceedings of the First International Conference on Robot Ethics (ICRE 2015)', Springer, Berlin, Germany. This paper was published in the compilation of ICRE 2015 papers, distributed at the location of ICRE 2015, where the paper was presented: Lisbon, Portugal. The URL given here goes to the preprint of the paper, which is shorter than the full Springer version.
URL: *http://kryten.mm.rpi.edu/SBringsjord_ethical_hierarchy_0909152200NY.pdf*

Bringsjord, S., Arkoudas, K. & Bello, P. (2006), 'Toward a General Logicist Methodology for Engineering Ethically Correct Robots', *IEEE Intelligent Systems* **21**(4), 38–44.
URL: *http://kryten.mm.rpi.edu/bringsjord_inference_robot_ethics_preprint.pdf*

Bringsjord, S. & Ferrucci, D. (1998), 'Logic and Artificial Intelligence: Divorced, Still Married, Separated...?', *Minds and Machines* **8**, 273–308.

Bringsjord, S. & Govindarajulu, N. S. (2013), Toward a Modern Geography of Minds, Machines, and Math, *in* V. C. Müller, ed., 'Philosophy and Theory of Artificial Intelligence', Vol. 5 of *Studies in Applied Philosophy, Epistemology and Rational Ethics*, Springer, New York, NY, pp. 151–165.
URL: *http://www.springerlink.com/content/hg712w4l23523xw5*

Bringsjord, S., Licato, J., Govindarajulu, N., Ghosh, R. & Sen, A. (2015), Real Robots that Pass Tests of Self-Consciousness, *in* 'Proccedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2015)', IEEE, New York, NY, pp. 498–504. This URL goes to a preprint of the paper.
URL: *http://kryten.mm.rpi.edu/SBringsjord_etal_self-con_robots_kg4_0601151615NY.pdf*

Chen, D. & Manning, C. D. (2014), A fast and accurate dependency parser using neural networks, *in* 'Empirical Methods in Natural Language Processing (EMNLP)'.

Chisholm, R. (1955), 'Law Statements and Counterfactual Inference', *Analysis* **15**, 97105.

Declerck, R. & Reed, S. (2001), *Conditionals: A Comprehensive Empirical Analysis*, Topics in English Linguistics, De Gruyter Mouton, Boston, MA. This book is volume 37 in the series.

Feldman, F. (1978), *Introductory Ethics*, Prentice-Hall, Englewood Cliffs, NJ.

Fellbaum, C. (1998), *WordNet: An Electronic Lexical Database*, Bradford Books.

Govindarajulu, N. S. & Bringsjord, S. (2015), Ethical Regulation of Robots Must be Embedded in Their Operating Systems, *in* R. Trappl, ed., 'A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations', Springer, Basel, Switzerland, pp. 85–100.
URL: *http://kryten.mm.rpi.edu/NSG_SB_Ethical_Robots_Op_Sys_0120141500.pdf*

Marton, N., Licato, J. & Bringsjord, S. (2015), Creating and Reasoning Over Scene Descriptions in a Physically Realistic Simulation, *in* 'Proceedings of the 2015 Spring Simulation Multi-Conference'.
URL: *http://kryten.mm.rpi.edu/Marton_PAGI_ADR.pdf*

Pollock, J. L. (1976), *Subjunctive Reasoning*, Vol. 8 of *Philosophical Studies series in Philosophy*, D. REIDEL PUBLISHING COMPANY.

Sharpe, R. (1971), 'Laws, coincidences, counterfactuals and counteridenticals', *Mind* **80**(320), 572–582.

Waller, N., Yonce, L., Grove, W., Faust, D. & Lenzenweger, M. (2013), *A Paul Meehl Reader: Essays on the Practice of Scientific Psychology*, number 9781134812141 *in* 'Multivariate Applications Series', Taylor & Francis.

Weidenbach, C. (1999), Towards an automatic analysis of security protocols in first-order logic, *in* 'Conference on Automated Deduction', pp. 314–328.

Young, R. (2016), Voluntary Euthanasia, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Summer 2016.
URL: *http://plato.stanford.edu/archives/sum2016/entries/euthanasia-voluntary*