ICRES 2018: International Conference on Robot
Ethics and Standards, Troy, NY, 20-21 August 2018.
https://doi.org/10.13180/icres.2018.20-21.08.id#

FOR AIs, IS IT ETHICALLY/LEGALLY PERMITTED THAT ETHICAL OBLIGATIONS OVERRIDE LEGAL ONES?

A. SEN* and P. MAYOL

Department of Cognitive Science, RPI,
Troy, NY 12180, USA
E-mail: Atriya@AtriyaSen.com and mayolp@rpi.edu*
www.rpi.edu

B. SRIVASTAVA and K. TALAMADUPULA

IBM Research, 1101 Kitchawan Rd,
Yorktown Heights, NY 10598, USA
E-mail: biplavs@us.ibm.com and krtalamad@us.ibm.com

N. SUNDAR G. and S. BRINGSJORD

Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
E-mail: naveensundarg@gmail.com and Selmer.Bringsjord@gmail.com
www.rpi.edu

We propose and defend the relevance of Ronald Dworkin's theory of associative associations to conflicts of legal versus moral responsibility in intelligent artificial agents. We exemplify this by describing the computational resolution of such a conflict by a new form of multi-agent problem-solving operating in the domain of smart-home appliances.

Keywords: Artificial Intelligence; Internet of Things; Multi-Agent; Machine Ethics.

1. Introduction

The present paper is concerned with the relationship between the legal and moral obligations of humans and intelligent artificial agents. Section 2 very briefly describes technologies already in place, which we leverage in what follows: a first-order deontic multi-operator modal cognitive calculus[a] $\mathcal{DCEC}$ that we use to (among other things) represent knowledge about the mental states of human and artificial agents, an automated theorem prover (*ShadowProver*) and planner (*Spectra*) for computational reasoning in this logic, and a new paradigm of artificial intelligence (*Tentacular AI*) based on distributed agents employing such reasoning in coördinated fashion. In Section 3 we describe a scenario involving smart-home appliances that exemplifies an apparent impasse between legal and moral obligations. In Section 4 we demonstrate how legal obligations are automatically and formally understood, and in Section 5 we propose and defend the applicability of a specific legal philosophy in resolving this impasse. This resolution we describe in Section 6. Finally, in Section 7, we summarize our arguments. At *ICRES 2018*, we presented a demonstration of the key reasoning, performed computationally by *ShadowProver*.

[a]This cognitive calculus is, from a proof-theoretic point of view, a *logic*, in that it has both a formal language and formal proof theory. We refrain from using the term 'logic' in part because the $\mathcal{DCEC}$ lacks a traditional formal semantics; in part because a logic, even a modal logic, needn't have operators for propositional attitudes (such as believing, knowing, intending, communicating, desiring, etc.); and because this system is intended by Bringsjord to be in line with Leibniz's search for a universal cognitive calculus. Hereafter, we will simply say 'calculus', in the tradition originated by Leibniz.

2. Framework for Computational Reasoning

2.1. *The Deontic Cognitive Event Calculus*

The **deontic cognitive event calculus** ($\mathcal{DCEC}$) is a first-order modal logic. $\mathcal{DCEC}$ has a well-defined syntax and inference system.[?] The inference system is based on natural deduction,[?] and includes all the introduction and elimination rules for first-order logic, as well as inference schemata for the modal operators and related structures.

This system has been used previously[?,?] to automate versions of the doctrine of double effect $\mathcal{DDE}$, an ethical principle with deontological and consequentialist components. While describing the calculus is beyond the scope of this paper, we give a quick overview of the system below. Dialects of $\mathcal{DCEC}$ have also been used to formalize and automate highly intensional (i.e. cognitive) reasoning processes, such as the false-belief task[?] and *akrasia* (succumbing to temptation to violate moral principles).[?] Arkoudas and Bringsjord[?] introduced the general family of **cognitive event calculi** to which $\mathcal{DCEC}$ belongs, by way of their formalization of the false-belief task. More precisely, $\mathcal{DCEC}$ is a sorted (i.e. typed) quantified modal logic (also known as sorted first-order modal logic) that includes the event calculus, a first-order calculus used for commonsense reasoning.

2.2. *Tentacular AI*

We bring to bear a new form of distributed, multi-agent artificial intelligence, which we refer to as being "tentacular." Tentacular AI is distinguished by six attributes, which among other things entail a capacity for reasoning and planning based in highly expressive calculi (logics), and which enlist subsidiary agents across distances circumscribed only by the reach of the Internet.

- $\mathbf{D_1}$ *Capable of problem-solving*. All TAI agents can plan, reason, learn, communicate; and they are capable of carrying out physical actions.
- $\mathbf{D_2}$ *Capable of solving at least important instances of problems that are at and/or **above** Turing-unsolvable problems*.
- $\mathbf{D_3}$ *Able to supply justification, explanation, and certification of supplied solutions, how they were arrived at, and that these solutions are safe/ethical*.
- $\mathbf{D_4}$ *Capable of "theory-of-mind" level reasoning, planning, and communication*.
- $\mathbf{D_5}$ *Capable of creativity*.
- $\mathbf{D_6}$ *Has "tentacular" power wielded throughout the internet and Internet of Things (I/IoT), Edge Computing, and cyberspace*. They can perceive and act through the I/IoT and cyberspace, across the globe.

We give a quick and informal overview of Tentacular AI.[?,?] We have a set of agents $a_1, \ldots, a_n$. Each agent has an associated (implicit or explicit) contract that it should adhere to. Consider one particular agent τ. During the course of this agent's lifetime, the agent comes up with goals to achieve so that its contract is not violated. Some of these goals might require an agent to exercise some or all of the six attributes of TAI. If some goal is not achievable on its own, τ can seek to recruit other agents by leveraging their resources, beliefs, obligations, etc.

3. A Scenario

It's winter in Berlin NY. Night. Outside, a blizzard. The mother and father of the home H, and their two toddler children, are fast asleep. The smartphone of each parent is set to "Do Not Disturb", with incoming clearance for only close family. There is no landline phone. A carbon monoxide sensor in the basement, near the furnace, suddenly shows a readout indicating an elevated level, which proceeds to creep up. τ perceives this, and forms hypotheses about what is causing the elevated reading, and believes on the basis of using a cognitive calculus that the reading is accurate (to some likelihood factor). The nearest firehouse is notified by τ. No alarm sounds in the house. τ runs a diagnostic and determines that the battery for the central auditory alarm is shot. The reading creeps up higher, and

now even the sensors in the upstairs bedrooms where the humans are asleep show an elevated, and climbing, level. τ perceives this too.

Unfortunately, τ reasons that by the time the firemen arrive, permanent neurological damage or even death may well (need again a likelihood factor) be caused in the case of one or more members of the family. Without enlisting the help of other agents in planning and reasoning, τ can't save the family; τ knows this on the basis of proof/argument.

τ can likely wake the family up, starting with the parents, in any number of ways. In a circumstance such as this, however, there is a clear conflict between the ethical and legal obligations of the intelligent system. Each of these ways entails violation of at least one legal prohibition that has been created by contracts that are in place. These contracts have been analyzed by an IBM service, which has stocked the mind of τ with knowledge of legal obligations in $\mathcal{DCEC}$ — or rather in a dialect that has separate obligation operators for legal ($\mathbf{O}_l$) and moral ($\mathbf{O}_m$) obligations. The moral obligation to save the family overrides the legal prohibitions, however. τ turns on the TV in the master bedroom at maximum volume, and flashes a warning to leave the house immediately because of the lethal gas that is building up.

4. Legal (Contractual) Obligations

We expect the eventual adoption of politically and socially motivated laws, enforced by governments, governing the activities of robots and other embodied artificially intelligent agents. In addition to this, specific rules of conduct impressed upon a robot by its manufacturers, in the form of a *legal contract*, may be thought of as describing its legal duties and obligations. This is in the spirit of the well-known 'Laws of Robotics' due to Isaac Asimov,[?] which are clearly intended as to encapsulate a specific attitude toward *public policy* (which may be legally enforced by requiring the Laws to form part of every robot contract), rather than as a moral statement.

In this section, we describe the augmentation of a standard smart-home appliance contract with arbitrary legal clauses; we exemplify this with a clause protecting the privacy of residents. We describe services provided by IBM, which enables automatic annotation of legal contracts, and our own technology enabling automatic parsing, of such clauses, into $\mathcal{DCEC}$ formulae.

4.1. *A Contract*

Suppose that the smart-home contract has been augmented with the following plausible clause.

> The Owner of a TAI Agent may at any time issue a "Do Not Disturb" (DND) instruction. When this instruction is issued, the Agent must not disturb the owner until the time specified, or until the Owner explicitly voids the DND.

As stated (Section 3), the smartphone of each parent in our scenario has been set to "Do Not Disturb", with incoming clearance only for close family.

4.2. *Automated Annotation of Contract Using IBM Services*

We use the IBM Watson *Discovery*[?] web-service. The service parses each contract sentence to identify specific features, and further, uses statistical classification to predict their values. Here, we get (fragment, in JSON):

```
{"label": {"nature": "Obligation","party": "Agent"},
"assurance": "High"}]
```

Specifying that this clause was identified (with high confidence) as an *obligation* on the part of Agent.

4.3. *Natural Language Parsing*

We cannot describe our natural language parsing technology here; for details we refer the reader to a previous work.[?] We merely report that part of the DND clause above may be automatically parsed into the following $\mathcal{DCEC}$ logical formula (see Section 6 for the modal operator $\mathbf{D}$ for *duty*):

Annotated & Parsed Contract Clause Fragment

$$\mathbf{D}(Agent, Owner, DND(Owner, t), Undisturbed(Owner, t))$$

Where $DND(x, t)$ is the assertion that an agent x does not wish to be disturbed until time t; $Undisturbed(x, t)$ is the fluent that x is not disturbed at time t.

5. Moral (Ethical) Obligations

It has been accepted, historically, that all persons are obligated to the law, in the sense of the existence of a *moral obligation* on their part to obey the law; this moral obligation has been usually termed a *political obligation*.[?] Yet, this thesis has come under intense scrutiny.[?] (The thesis of *legal positivism*[?] demands that law be based *exclusively* on social facts, and not moral arguments. How then, does a moral obligation to obey the law, follow?) We expect the controversy to extend itself to the obligations of artificially intelligent agents, such as TAI agents.

We propose that any legal contract binding an intelligent artificial agent to humans be considered to automatically establish a relationship between them that justifies the expectation of *special obligations*,[?] and especially of the agent(s) toward the human(s). Special obligations are warranted in many human relationships, such as parenthood and even neighborhood (the relationship between neighbors). Terms of the contract must then be interpreted in the context of these special obligations.

It may be pointed out that a legal contract does not (and cannot) specify explicitly, a potentially infinite space of special obligations. Consequently, they cannot be directly consented to. It has been argued[?] that such *voluntary* consent is necessary to justify such obligations. Ronald Dworkin however argues to the contrary, that 'associative obligations' or obligations 'by role', are not such as to require choice or consent,[?] when applied between members of a 'true community'. We postulate that humans and intelligent artificial agents in our smart home comprise such a community, the characteristics of which are specified by Dworkin as follows (emphasis ours).

> **First**, they must regard the group's obligations as special, holding distinctly within the group, rather than as general duties its members owe equally to persons outside it. **Second**, they must accept that these responsibilities are personal: that they run directly from each member to each other member, not just to the group as a whole in some collective sense. ... **Third**, members must see these responsibilities as flowing from a more general responsibility each has of concern for the well-being of others in the group ... **Fourth**, members must suppose that the group's practices show not only concern but an equal concern for all members.

It is crucial to note that it is not legal or contractual status that characterizes a true community; rather it is the *psychological* status of its members. In other words, this is a necessarily *cognitive* theory of communal obligation. (By Dworkin's own account, a subject considering its legal duties is having 'a conversation with oneself' and is 'trying to discover his own intention in maintaining and participating in that practice.'[?]) In the next section, we formalize this theory in our $\mathcal{DCEC}$ and

propose a computational mechanism for intelligent artificial agents to determine and defend their moral obligations.

6. Toward a Moral TAI

We formulate a many-sorted first-order modal theory in the $\mathcal{DCEC}$ described in Section 2.1, augmented with a modal operator $\mathbf{D}$, representing the *duty* of an agent toward another agent. This is defined as follows:

Modal Operator D

$\mathbf{D}(x, y, \phi, d)$: agents x and y are members of a *true community*, ϕ is a proposition, d is a duty, if x *believes* ϕ then x has the duty d toward y.

Then, consider first-order predicates as follows:

Sorts and Predicates

$Community(x)$: x is a true community.
$Duty(d)$: d is a duty.
$Agent(\alpha)$: α is an agent.

$InComm(x, y)$: the agent x is a member of the true community y.
$Concern(x, y)$: the agent x has concern for the well-being of the agent y.

Then, Dworkin's theory may be formalized as follows:

Dworkin's Theory Formalized

$$\mathbf{Rule1}: \forall x, y, z(Community(x) \wedge InComm(y, x) \wedge InComm(z, c) \rightarrow Concern(y, z)) \wedge Concern(z, y))$$

$$\mathbf{Rule2}: \forall x, y, z(Community(x) \wedge InComm(y, x) \wedge InComm(z, c) \rightarrow Concern(y, z)) = Concern(z, y)$$

$$\mathbf{Rule3}: \forall x, y, z, t((Community(x) \wedge InComm(y, x) \wedge InComm(z, x) \wedge \mathbf{K}(y, t, \phi, D(y, z, \phi, d))) \rightarrow \mathbf{B}(y, t, \mathbf{SE}(y, t, InComm(z), \alpha)))$$

$$\mathbf{Rule4}: \forall x, y, z, d(Community(x) \wedge InComm(y, x) \wedge InComm(z, x) \wedge duty(d, x) \wedge Concern(y, z) \rightarrow \mathbf{D}(y, z, \phi, d))$$

The relationship between actions and duties may be characterized as follows:

Actions and Duties

$\forall f, t_1, t_2, \alpha, x((Community(x) \wedge duty(f, x) \wedge action(\alpha) \wedge HoldsAt(f, t_1) \wedge HoldsAt(f, t_2) \wedge Happens(\alpha, t_1) \wedge t2 > t1) \rightarrow \neg Clipped(t_1, f, t_2))$

$\forall f, t_1, t_2, \alpha, x((Community(x) \wedge duty(f, x) \wedge action(\alpha) \wedge \neg HoldsAt(f, t_1) \wedge HoldsAt(f, t_2) \wedge t_2 > t_1) \rightarrow Initiates(\alpha, f, t_1))$

(That is, an agent whose duty it is to ensure that a particular fluent holds at a particular time will, if the fluent does not already hold, take an action that causes it to be hold at that time, or if the fluent does already hold, will refrain from taking an action that changes this state of the fluent.)

This formal framework being now in place, we may informally describe the reasoning of the smoke-detector TAI agent as follows:

(1) There is a community named **Domum** (Latin for house).(**given**)
(1) $Community(\textbf{Domum})$

---

(2) This community has a Tentacular AI (TAI) Agent as a member.(**given**)
(2) $Agent(TAI) \wedge InComm(TAI, \textbf{Domum})$

---

(3) **Domum** has another member who is the Owner of the TAI agent above, and is asleep.(**given**)
(3) $Agent(owner) \wedge Owner(owner, TAI) \wedge InComm(owner, \textbf{Domum}) \wedge Asleep(owner)$

---

(4) The TAI Agent **reasons** that it has concern for the well-being of this Owner.
(4) $Concern(TAI, owner)$ (**Rule 1**, **(1),(2),(3)**)

---

(5) If carbon monoxide levels in the house rise, it will lead to the death of the Owner; the Agent knows this.(**given**)
(5) $\forall x, t, t'((COlevels(medium) \vee COlevels(high)) \wedge t < t') \rightarrow \mathbf{K}(TAI, t, Death(owner, t')))$

---

(6) Carbon monoxide levels rise. (**given**)
(6) $COlevels(\textbf{Domum}, medium)$

---

(7) The Agent **reasons** that it knows that it has a Duty towards the Owner to keep him/her safe given dangerous conditions.
(7) $\mathbf{K}(TAI, t, Death(owner, t'), \mathbf{D}(TAI, owner, Death(owner, t'), stopDeath(owner, t)) \wedge t < t'$ (**Rule 4**, **(4)**)

---

(8) Given that the TAI Agent knows this Duty, and that carbon monoxide levels rising is a dangerous situation, it believes it has a *super-erogatory obligation* to prevent the death of the Owner.
(8) $\mathbf{B}(TAI, t, \mathbf{SE}(TAI, t, InComm(owner), stopDeath(owner, t)))$
(**Rule 3, (7)**)

---

(9) Given this, it must wake him/her up by *any* means, even when in violation of a clause of its contract.
(9) $turnon(speaker)$ **(8)**

Where $Asleep(x)$ means that x is asleep, $Death(x, t)$ means x died at time t or x will die at t, $stopDeath(x, t)$ is an action that will employ another action to stop the death of x at t, $Owner(x, y)$ means x owns agent y, and $COlevels(y, x)$ where x is either *low*, *medium*, or *high*, means that the carbon monoxide levels in the community y are at x.

This reasoning may be automated in ShadowProver and Spectra to automatically deduce that the Agent must turn the speaker on at full volume, to wake the Owner up.

7. Conclusion & Future Work

We have proposed, defended, and described the applicability of Dworkin's theory of associative obligations to the resolution of conflicting legal and moral obligations in intelligent artificial agents. At *ICRES 2018*, we demonstrated our work via computational reasoning carried out by physical smart-home devices.

8. Acknowledgements

A grant provided by the AI Research Collaboration between RPI and IBM, for "Tentacular AI," made possible the lion's share of the research described above. In addition, a grant from the Office of Naval Research to explore "moral competence in machines" (PI M. Scheutz) has provided indispendable support for the research reported herein. Crucial support also came in the form of a grant from the Air Force Office of Scientific Research to make possible "great computational intelligence" in AIs

on the strength of automated reasoning (S. Bringsjord PI).

References

1. N. S. Govindarajulu and S. Bringsjord, On Automating the Doctrine of Double Effect, in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, ed. C. Sierra (Melbourne, Australia, 2017). Preprint available at this url: https://arxiv.org/abs/1703.08922.
2. G. Gentzen, Investigations into Logical Deduction, in *The Collected Papers of Gerhard Gentzen*, ed. M. E. Szabo (North-Holland, Amsterday, The Netherlands, 1935) pp. 68–131. This is an English version of the well-known 1935 German version.
3. N. S. Govindarajulu, S. Bringsjord, R. Ghosh and M. Peveler, Beyond the doctrine of double effect: A formal model of true self-sacrifice, International Conference on Robot Ethics and Safety Standards, (2017).
4. K. Arkoudas and S. Bringsjord, Toward Formalizing Common-Sense Psychology: An Analysis of the False-Belief Task, in *Proceedings of the Tenth Pacific Rim International Conference on Artificial Intelligence (PRICAI 2008)*, eds. T.-B. Ho and Z.-H. ZhouLecture Notes in Artificial Intelligence (LNAI)(5351) (Springer-Verlag, 2008).
5. S. Bringsjord, N. S. Govindarajulu, D. Thero and M. Si, Akratic Robots and the Computational Logic Thereof, in *Proceedings of ETHICS • 2014 (2014 IEEE Symposium on Ethics in Engineering, Science, and Technology)*, (Chicago, IL, 2014). IEEE Catalog Number: CFP14ETI-POD.
6. Tentacular AI http://kryten.mm.rpi.edu/TAI/tai.html, (2018), [Online; accessed 07-July-2018].
7. S. Bringsjord, N. S. Govindaralajulu, A. Sen, M. Peveler, B. Srivastava and K. Talamadupula, Tentacular Artificial Intelligence, and the Architecture Thereof, Introduced, *To be presented at the FAIM Workshop on Architectures and Evaluation for Generality, Autonomy & Progress in AI.* (2018).
8. I. Asimov, *I, robot* (Spectra, 2004).
9. IBM Watson Discovery Service https://console.bluemix.net/catalog/services/discovery, (2018), [Online; accessed 07-July-2018].
10. S. Bringsjord, J. Licato, N. Govindarajulu, R. Ghosh and A. Sen, 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, 2015). This URL goes to a preprint of the paper.
11. L. Green, Legal obligation and authority, in *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta (Metaphysics Research Lab, Stanford University, 2012) Winter 2012 edn.
12. L. Green, Legal positivism, in *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta (Metaphysics Research Lab, Stanford University, 2018) Spring 2018 edn.
13. D. Jeske, Special obligations, in *The Stanford Encyclopedia of Philosophy*, ed. E. N. Zalta (Metaphysics Research Lab, Stanford University, 2014) Spring 2014 edn.
14. R. Dworkin, *Law's empire* (Harvard University Press, 1986).