A 21st-Century Ethical Hierarchy for Robots and Persons: $\mathscr{EH}$

S. Bringsjord *

* Department of Computer Science
Department of Cognitive Science
Rensselaer AI & Reasoning (RAIR) Lab
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
selmer@rpi.edu

Abstract: I introduce the ethical hierarchy $\mathscr{EH}$, into which can be placed robots (as a species of information-processing machines), human persons, and persons in general. $\mathscr{EH}$ bears a deep debt to both Leibniz and R. Chisholm. The new hierarchy is catalyzed by consideration of, and reflects a firm negative answer to, the question: Can robots be more moral than humans? Any such claim as that computing machines can be more moral than human machines is, given $\mathscr{EH}$, seen to be demonstrably false. The light shed by $\mathscr{EH}$ also reveals why an emphasis on *legal* obligation for robots, while not unwise at the moment, is inadequate, and why at least the vast majority of today's state-of-the-art deontic logics are painfully naïve and morally inexpressive, whether they are intended to formalize the ethical behavior of robots or persons — which is why, with colleagues in my lab, the construction of the computational logic $\mathcal{L}_{\mathscr{EH}}$ is now underway. The illumination supplied by $\mathscr{EH}$ is also why, in the coöperative we're-all-in-this-together spirit, I encourage other logicist groups working in robot/machine ethics, and groups drawing directly from underpinnings in deontic logic, to pursue engineering that factors in $\mathscr{EH}$.

Keywords: robot ethics, machine ethics, ethics, deontic logic, ethical hierarchy

1. INTRODUCTION; PLAN

I introduce herein the ethical hierarchy $\mathscr{EH}$, into which can be placed robots (as a species of embodied information-processing machines), human persons, and persons in general. $\mathscr{EH}$ bears a special debt to both Leibniz and R. Chisholm (1982); in the latter case, the debt was incurred in large part in treasured personal interaction, and is much larger than can be conveyed by the adumbration of $\mathscr{EH}$ given herein. [1] The hierarchy is catalyzed by consideration of, and reflects a firm negative answer to, the question: Can robots be more moral than humans? Any such claim as that computing machines can be more moral than human machines is, given $\mathscr{EH}$, seen to be demonstrably false. The light shed by $\mathscr{EH}$ also reveals why an emphasis on *legal* obligation for robots is inadequate, and why at least the vast majority of today's state-of-the-art deontic logics are painfully naïve and inadquate, whether they are intended to formalize the ethical behavior of robots or persons — which is why, with colleagues, the construction of the computational logic $\mathcal{L}_{\mathscr{EH}}$ is underway. The illumination thrown by $\mathscr{EH}$ is also why, in the coöperative we're-all-in-this-together spirit, I encourage other logicist groups working in robot/machine ethics, and groups drawing directly from underpinnings in deontic logic, to as soon as possible change their engineering to factor in $\mathscr{EH}$. I don't think it matters what domain this engineering is aimed at: $\mathscr{EH}$ seemingly applies to military robots, healthcare robots, and so on.

The present paper's sequel follows this sequence: In the next section, 2, I consider the question

* The work that gave rise to this short paper was enabled by generous and ongoing support from the U.S. Office of Naval Research; see 'ACKNOWLEDGEMENTS.' I owe a special debt to Dan Messier and Bertram Malle for pressing the "Can robots be more moral than humans?" question, which catalyzed my thought that that query can serve as a laic portal to consideration of the hierarchy presented synoptically herein. I'm deeply grateful as well to two anonymous referees.

[1] For instance, a full specification of the hierarchy requires systematic consideration of intrinsic value, as e.g. set out in (Chisholm 1986) (since intrinsic value in a Leibnizian metaphysical sense is in $\mathscr{EH}$ the ultimate ground of the classification of actions). Note along this line that despite what I say below rather optimistically about $\mathcal{L}_{\mathscr{EH}}$, the fact is that, according to Chisholm and Leibniz, unless a deontic logic grounds the systematization of action in the formalization of intrinsic goodness (and badness), that logic will be incomplete.

as to whether robots can be morally superior to human persons; this question serves as a catalyst for introducing the informal, suggestive rudiments of $\mathscr{EH}$. Then (§3) I briefly remind cognoscenti of, and introduce non-experts to, the 19th-century tripartite hierarchy $\mathscr{T}$, which rather astoundingly survives to this very day as the anchor widely used for logicist robot/machine ethics. I conclude this section by expanding the trichotomous $\mathscr{T}$ to a variant $\mathscr{T}^{\mathscr{Q}}$ that divides each member of the classical triad into sub-categories based on five quantifiers. Then, in §4, I sketch $\mathscr{EH}$, making use in doing so of the quantifier quintet. I next (§5) proceed to briefly explain why engineering robots on the basis of only *legal* obligations is inadequate. What then follows (§6) is a brief explanation of why, in light of $\mathscr{EH}$, robot ethics and the engineering of ethically correct robots shouldn't be based on the obsolete trichotomy of the obligatory, the forbidden, and the morally indifferent (where the morally indifferent category is based that which is at once permissible and non-obligatory). I then (§7) make a few remarks about the under-construction logic $\mathcal{L}_{\mathscr{EH}}$, designed to take account of $\mathscr{EH}$. Next, in §8, under the illumination shed by $\mathscr{EH}$, I briefly discuss the dual fact that (i) plenty of humans are located in this hierarchy at points below robots that would be fairly easy to engineer, but that (ii) such unimpressive robots shouldn't be the ones aspiring robot-ethics engineers seek to build. The paper wraps up with a brief conclusion, in which I encourage those working in logicist robot ethics, and those whose work partakes of such ethics, to immediately take account of $\mathscr{EH}$.

2. CAN ROBOTS BE MORE MORAL THAN HUMAN PERSONS?

Let's start with a question, and my answer to that question:

(Q) Can humans build robots that will be more moral than humans?

No; positively no; that's my response. Others may see things differently, but presently whether I'm right or wrong isn't the core issue. It's a particular "side-effect" of my justification for my negative response that serves to introduce $\mathscr{EH}$, and this proposed hierarchy should sink or swim independently of my response to (Q). Here, then, is basically why I answer in the negative to (Q).

Question (Q) appears to me to presuppose a way to measure a creature's position on a continuum of degrees of moral performance. But no rigorous and received version of such a continuum is in the literature, as far as I know. Hence, to briefly justify my response to (Q) I take the liberty of invoking an informal version of my own continuum; that is, an informal version of $\mathscr{EH}$.

At the maximal end (moral perfection) a creature c infallibly meets all its obligations, and *in addition* carries out (relative to c's power and opportunities) all those supererogatory actions that are maximally good. At the other end would be a thoroughly evil creature: one who fails to meet all substantive obligations, and goes out of its way to carry out actions that are (relative to its level of power) maximally *sub*erogatory.

Creatures that are at once sentient, intelligent, free, and creative (SIFC) are, if you will (and again, merely in my opinion),"make or break." That is, they have the potential to reach high on the continuum — but can also fall very, very low on it. In contrast, creatures that lack one or more of SIFC necessarily fall somewhere near the midpoint: they can't be morally great, but they can't be Satanic either.

Now to robots, present and future: For reasons already hinted at, they fall near the midpoint, and can't move anywhere else. They can't possibly reach moral greatness; we can. Why? At least in broad strokes, it's simple:

Computing machines aren't conscious (there's nothing it's like to be a robot; they are in this regard no different than, say, slabs of granite), and consciousness is a requirement for moral performance at the level of a human person. In other words, robots lack the S in the SIFC quartet. Without sentience they can't for example empathize; hence they can't understand one of the main, underlying mental requirements for the sort of supererogatory actions constitutive of moral greatness (and as a matter of fact, for the sort of suberogatory actions constitutive of the diabolical: a sadist, e.g., gains conscious pleasure from knowing that his victim is experiencing conscious pain). For instance, Jones may spontaneously compose a sympathy note to Smith not because Jones is obligated to do so and/or believes that he is, but rather because he feels Smith's sorrow, and seeks to apply epistolary salve.

Of course, I well know that some readers will insist that mere information-processing machines *can* be not only — to use Block's (1995) distinction — "access-conscious" (= A-conscious), but can also be "phenomenal-conscious" (= P-conscious). In essence, the former form of consciousness requires only the information-processing structures necessary to enable a creature to perceive and reason in ways that are fully circumscribed by mechanical processes. (We might refer to A-consciousness

as "zombie" consciousness (Bringsjord 1999).) The later form of consciousness, P-consciousness, requires having genuine subjective awareness, including what are called "qualia." The view that robots can't be P-conscious is defended for example in (Bringsjord 2007); a prior defense of this negative view was articulated in *What Robots Can and Can't Be* (1992). In the present short paper, I don't in any way assume that these arguments are sound. I of course believe that they are, but the coherence, applicability, and implications of $\mathscr{EH}$ doesn't in any way hinge on the soundness of these earlier arguments. To repeat: consideration of (Q), and my justified response to it, has served simply to place the rudiments of $\mathscr{EH}$ on the table.

In addition (and this relates to the I and C in the SIFC quartet, a pair that, relative to humans, is at least compromised in the case of robots), moral greatness entails having a capacity to solve difficult moral dilemmas. But it seems to me that such dilemmas can be as complicated as higher mathematics, perhaps more so. Robots in my opinion won't ever have the intellectual firepower needed for truly demanding math. (Currently, machines are unable to e.g. even prove the elementary theorems that students are expected to prove in the case of introductory axiomatic set theory.) Ergo, the moral performance of robots will forever be below the moral reach of human persons, as I see it. [2] Of course, once again, whether or not I'm correct is an issue orthogonal to whether or not $\mathscr{EH}$ implies that contemporary robot ethics and robot-ethics robotics need to be refashioned.

3. THE 19TH-CENTURY HIERARCHY $\mathscr{T}$

While in my experience most machine/robot ethicists (indeed, most formally inclined ethicists, period!) seem to think the "modern" logically interconnected trio of concepts, *forbidden*, *permissibility*, and *obligatory*, which underlie the vast majority of deontic logics to the present moment, came on the formal scene for the first time in the middle of the 20th century on the strength of seminal work by von Wright (1951), the fact of the matter is that the trio debuted in the *19th century*. [3] Yet the trio not only lives on, but dominates today's robot-ethics landscape. Certainly I must confess that in my own robot-ethics work hitherto, with a few exceptions (e.g., the divine-command deontic logic explained in (Bringsjord & Taylor 2012), which

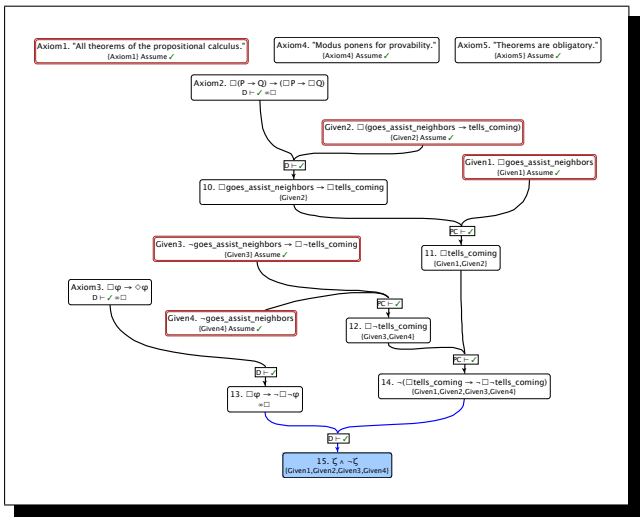

Fig. 1. Chisholm's Paradox in Standard Deontic Logic (proof in the Slate system)

happens to exploit other Chisholmian work), the thrust has been based on modernized versions of the operator **O** for obligation, **F** for forbiddenness, **P** for permissibilty, and, derivatively, that which is morally indifferent, designed to be captured by **I** (this is plainly seen e.g. in Arkoudas, Bringsjord & Bello 2005). Of course, as cognoscenti will recall, it was Chisholm's (1963) Paradox (CP) that gave birth to deontic logic in earnest: we knew from the moment that his proof was published that simple use of the operators just given would lead immediately to inconsistency. Hence the kind of simple deontic logics laid out even over three decades after CP (e.g., in (Chellas 1980)), which unfortunately have found their way like a cancer into contemporary robot/machine ethics, are provably inconsistent (see the proof shown Figure 1).

While RAIR-Lab robot-ethics engineering steers clear of Chisholm's Paradox, our logics provided thus far have been based on a dyadic operator **O**, and correspondingy dyadic versions of **P** and **F**; these logics are currently configured as an "ethical stack"; see Figure 2. This figure gives a pictorial bird's-eye perspective of the high-level architecture of a system from the RAIR Lab that augments the DIARC (**D**istributed **I**ntegrated **A**ffect, **R**eflection and **C**ognition) (Schermerhorn, Kramer, Brick, Anderson, Dingler & Scheutz 2006) robotic platform with ethical competence. Ethical reasoning is implemented as a hierarchy of formal computational logics (including, most prominently, sub-deontic-logic systems) which the DIARC system can call upon when confronted with a situation that the hierarchical system believes is ethically charged. If this belief is triggered, our hierarchical ethical system then attacks the problem with increasing levels of sophistication until a solution

[2] For readers who may be interested, arguments in support of the claims in the present paragraph can e.g. be found in (Bringsjord & Zenzen 2003, Bringsjord, Kellett, Shilliday, Taylor, van Heuveln, Yang, Baumes & Ross 2006).

[3] E.g., Chisholm (1982, p. 99) points out that Höfler had the deontic square of opposition in 1885.

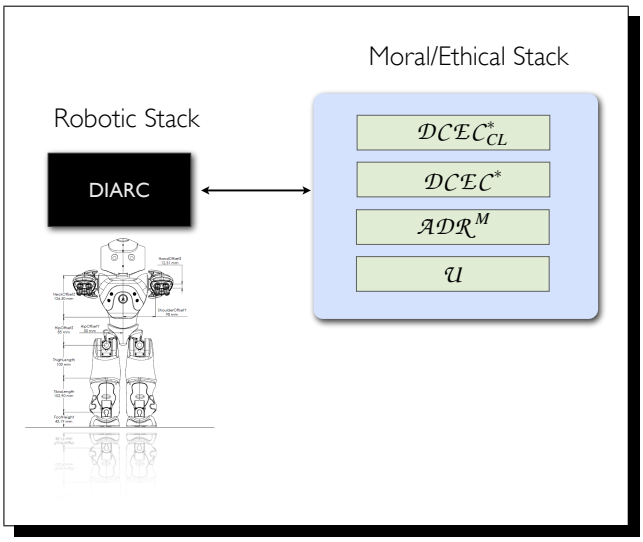

Fig. 2. Pictorial Overview of Non-$\mathscr{EH}$-Baed Situation The first layer, $\mathcal{U}$, is, as said in the main text, based on UIMA; the second layer on what we call *analogico-deductive reasoning* for ethics; the third on the "deontic cognitive event calculus" with a indirect indexical; and the fourth like the third except that the logic in question includes aspects of conditional logic. (Robot schematic from Aldebaran Robotics' user manual for Nao. The RAIR Lab has a number of Aldebaran's impressive Nao robots.)

is obtained, and then passes on the solution to DIARC. This approach, while satisfactory in the near-term, is ultimately inadequate for two reasons. One, the efficacy of this approach (and an expansion of the approach based on $\mathscr{EH}$), requires that implemented deontic logics have control at the operating-system level (an issue treated in detail in Bringsjord & Govindarajulu forthcoming). The second defect is that this hierachy is based on the obsolete 19th-century hierarchy.

So what is this 19th-century hierarchy that underlies even much of my lab's contemporary work, and other work that partakes of underpinnings based on that which is forbidden, permissible, and obligatory (e.g., Arkin 2009)? We can set out the hierarchy by simply positing clusters of behaviors corresponding to the standard operators. For example, a creature that performs forbidden actions would fall into the cluster $\mathcal{F}$, whereas a creature whose performed actions meet obligations would fall into $\mathcal{O}$. Here then, given a self-explanatory way of picking out the set of morally indifferent actions, is the obsolete hierarchy:

$$\mathscr{T} := \|\mathcal{F}|\mathcal{P} \wedge \neg\mathcal{O}|\mathcal{O}\|$$

It will be convenient if I next introduce, before passing to the hierarchy $\mathscr{EH}$, a mechanism, based on five quantifiers, for sub-categorizing a cluster of actions. Here are the five quantifiers in question, where of course the first and fifth will be familiar to all readers:

all: $\forall$
few: `F`
most: `M`
vast majority: `V`
at least one: $\exists$

The basic idea is straightforward. An agent that, within for instance the category $\mathcal{O}$, meets all its obligations, falls within the sub-category $\mathcal{O}^{\forall}$; an agent that meets only a few of its obligations falls within the sub-category $\mathcal{O}^{\mathtt{F}}$; and so on. (Inductive, rather than merely deductive, logics are needed to formalize the three non-standard quantifiers. The logic $\mathcal{L}_{\mathscr{EH}}$is inductive, and below (§7) I say a few words about the proof-theoretic machinery needed for the quantifier `M`.) Here then is a basic picture of the new — but still fundamentally trichotomous — hierarchy $\mathscr{T}^{\mathcal{Q}}$:

$\mathcal{F}$					$\mathcal{P} \wedge \neg\mathcal{O}$	$\mathcal{O}$				
$\forall$	`F`	`M`	`V`	$\exists$		$\forall$	`F`	`M`	`V`	$\exists$

At this point, two immediate confessions are in order, before proceeding. It will occur to the skeptical reader that the use of the existential quantifier in $\mathscr{T}^{\mathcal{Q}}$ is peculiar. The reason is of course that in standard first-order logic $\vdash \forall x\phi(x) \rightarrow \exists x\phi(x)$. Accordingly, confession one: we don't here have a strict sub-hierarchy via quantification, such as is seen in the quantifier-based version of the Arithmetic Hierarchy (Davis, Sigal & Weyuker 1994). Options are available for fixing this quirk, but given space constraints I don't discuss them herein. [4] The second confession is that while there is no consequentialist fabric indicated by the bare bones of $\mathscr{T}^{\mathcal{Q}}$, such a fabric is ultimately desirable to flesh out and exploit in some detail. The reason is that, with respect to their consequences, not all actions within the same sub-category are equivalent. In the real world, opportunity is an important factor in determining one's place in an ethical hierarchy. If Smith is locked in solitary confinement, the (leaving aside purely mental actions to ease exposition) range of obligations that bind him may be severely limited. Jones, a free man living in interaction with other humans, may in contrast be bound by numerous demanding obligations. If Jones manages to meet most of his obligations, and Smith does too, it would be counter-intuitive to classify Jones and Smith as both (over some fixed time interval) within $\mathcal{O}^{\mathtt{M}}$. In the present paper, which is intended to introduce $\mathscr{EH}$ and not to plumb its depths, I confess to ignoring this complication.

[4] One option is of course to supplant $\exists$ with $\exists^{=1}$.

4. $\mathscr{EH}$, FOR THE 21ST CENTURY

Why is the hierarchy $\mathscr{T}$ incomplete? Perhaps the quickest way to see that $\mathscr{T}$ is incomplete is to simply call to mind actions in the human sphere that are heroic or saintly, a category famously depicted by Urmson (1958). Each week the newspapers bring to us stories about people who perform actions that are good, but — to use Ladd's (1957, p. 127) phrase — "not wrong not to do." In fact, such cases in the past are so numerous, and new ones are so easily imagined, that I will spend no time citing or concocting any at the moment. [5] (We shall consider below an excellent example from Scheutz & Arnold (forthcoming) in the robot realm.) I suspect, in fact, that the reader has himself/herself performed some heroic acts of self-sacrifice: acts that were permissible, good, but not obligatory. In a world filled with poverty and disease, there are a lot of opportunities for heroic actions.

But as Chisholm (1982, p. 100) points out, supererogatory actions include not just those that are heroic, saintly, or self-sacrificial, but also actions that are courteous, polite, kind. Showing kindness to dogs, going out of one's way to pet them when coming upon them in the normal course of life; issuing compliments to those who clearly have invested much time in their appearance or a project; giving words of encouragement to colleagues who are under considerable pressure on the job — these actions compose a category of actions — called 'charitable' by Leibniz — that are supererogatory, but obviously not saintly or heroic. Accordingly, I divide supererogatory actions into two categories, there merely charitable ($\mathcal{S}^{up1}$), and the heroic or saintly ($\mathcal{S}^{up2}$). In addition, the flip side of these two categories exist on the "dark" side of $\mathscr{EH}$; that is, on the suberogatory side. [6] In addition, I roughly follow Leibniz and Grotius in sub-dividing duties or obligations into the less demanding legal ones that proscribe harm, and the more demanding general space of ethical obligations. Given that we preserve the five quantifiers used in $\mathscr{T}^{Q}$, we have our new, comprehensive hierachy (I include the up-arrow to mark the location of the military robots (currently uniquely) targeted by my lab):

$\mathscr{EH}$

$\mathcal{S}_{ub1}$	$\mathcal{S}_{ub2}$	$\mathcal{F}$	$\mathcal{P} \wedge \neg\mathcal{O}$	$\mathcal{O}^{L}$	$\mathcal{O}^{M}$	$\mathcal{S}^{up1}$	$\mathcal{S}^{up2}$
$\exists$–$\forall$	$\exists$–$\forall$	$\exists$–$\forall$		$\exists$–$\forall$	$\exists$–$\forall$	$\exists$–$\forall$	$\exists$–$\forall$
						$\uparrow$	

5. WHY ROBOT ETHICS BASED ON LAWS IS UNTENABLE

I provide two general reasons why machine/robot ethics based solely upon laws is inadequate, from the perspective of $\mathscr{EH}$. The first reason is perfectly straightforward and unsurprising: viz., that legal obligations are only a small proper subset of obligations. I may not be legally obligated to try to minister to a weeping colleague at work, but *ceteris paribus* I'm nonetheless morally obligated to do so. For Leibniz, and flowing therefrom for $\mathscr{EH}$, the *neminem laedere* principle that one shouldn't harm others is what generates obligations not to harm, and these are the "lowest" obligations (i.e., $\mathcal{O}^{L}$). One might say that Asimov's famous Three Laws of Robotics, discussed in (Bringsjord, Arkoudas & Bello 2006), fall within this sub-category of obligations; the trio is thus incomplete from the standpoint of $\mathscr{EH}$. [7] A robot joining a human soldier on a mission might well fulfill all its legal obligations (relative, e.g., to "laws of war" and "laws of engagement") while at the same time by failing to meet a moral obligation to minister to a severely depressed soldier might endanger the very mission in question. And, since as I will explain in the next (6) section, there are actions that are morally good, and indeed such that we would wish a robot on a mission to perform them, yet these actions aren't ethically obligatory.

The second reason why basing machine/robot-ethics on legal principles, at least in the military sphere, where (at least in the Occidental tradition) such principles are derived from, or at least directly reflective of, Just War Theory (JWT), is that extant law doesn't apply to cyberwarfare (Bringsjord & Licato forthcoming*a*, Bringsjord & Licato forthcoming*b*). [8] In order to formulate new

[5] Anyone who has stood atop Pointe du Hoc and pondered the self-sacrifice of the Rangers who battled the Nazis there will confront the stark reality that supererogation was required to vanquish Hitler. Leibniz would say that the pursuit of such victory makes no sense if there is no God and no afterlife (for reasons explained in Youpa 2013) — but this claim is one left aside here. I note only that Leibniz thought it was easy enough to prove God's existence, so for him, an ethics that presupposed God's existence was in no way scientifically problematic.

[6] I don't have the space to consider the evil actions in question; Chisholm (1982) provides some examples. By the way, it seems to me very likely that robots capable of *sub*erogatory actions will prove to be quite useful in espionage, but this topic cannot be discussed the present short paper. Readers interested in this direction are advised to begin with (Clark 2008).

[7] The trio isn't only incomplete, but is just plain unacceptable. A robot medic or surgeon would routinely need to harm humans in order to save them. In saying this, I narrowly condemn Asimov's trio only. Ethically sophisticated contemporary engineers have worked out avenues by which robots can trade short-term harm for longer-term good; see e.g. (Winfield et al. 2014).

[8] These papers thus provide a rigorous deductive case for a position at odds with the *Tallinn Manual on the International Law Applicable to Cyber Warfare* (Schmitt 2013).

laws of cyberwarfare and cyberengagement, the human race is going to need to back up to deeper ethical principles, and then work out to a replacement of new laws of conflict. Absent the completion of this undertaking, which of course promises to be complicated and time-consuming, we are going to need to strive for machines/robots that are above $\mathcal{O}^L$ in $\mathscr{EH}$.

6. WHY ROBOT ETHICS BASED ON $\mathscr{T}$ IS UNTENABLE

There are many reasons why the engineering of ethically correct robots needs to be based not on $\mathscr{T}$ or close variants thereof, but rather on $\mathscr{EH}$. In the interest of economy, I give only two here.

The first reason stems from the realities of human-robot interaction. Robots built to collaborate with humans but whose actions merely conform to what is obligatory, even if such robots fall into $\mathcal{O}^{\forall}$, would be highly problematic. I gave a case of this above, where a robot that fails to perform actions in $\mathcal{S}^{up1}$ in connection with a depressed soldier might endanger the mission the two are on.

The second reason why the engineering of moral robots needs to be based on $\mathscr{EH}$ is more interesting. The reason can be seen by considering a case described by (Scheutz & Arnold forthcoming), in which a robot doing road repair with a jackhammer notices a child dart out to retrieve a bouncing ball, "a car speedily approaching and headed directly at her." Under the supposition that the car will not be able to stop before hitting the girl, and that the robot can move the young to safety at the cost of losing its own life, what would have been the right kind of prior engineering here? Presumably the right sort of engineering would have been that which produced a robot that performs the supererogatory rescue of the girl.[9] Notice that even if one insists that the self-sacrificial rescue is an obligation for the robot, the fact remains that we must have robots able to consider that such actions, when performed by humans, are supererogatory. Hence we cannot dodge the need to engineer robots on the basis of the concepts that distinguish $\mathscr{EH}$ from $\mathscr{T}$.[10]

[9] In the human sphere, such a rescue would clearly fall into $\mathcal{S}^{up2}$. For reasons pertaining to A- versus P-consciousness and the imaginary robot, I classify the rescue as a $\mathcal{S}^{up1}$ action.

[10] Thoroughgoing Kantians might resist $\mathscr{EH}$, and the robot ethics and robot-ethics engineering that seems to naturally flow from it. This is an issue I'm prepared to address — but not in this short paper. Robot ethics as it relates to Kant should in my opinion begin with study of (Ganascia 2007); see also (Powers 2006).

7. ON THE LOGIC $\mathcal{L}_{\mathscr{EH}}$

Hitherto, Bringsjord-led work on robot ethics has been unwaveringly logicist (e.g., see Bringsjord & Govindarajulu forthcoming); that's par for a course 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*). Nothing in or about the hierarachy $\mathscr{EH}$ will change this trajectory. Howevever, $\mathscr{EH}$ does reveal that the logics invented and implemented thus far in this trajectory (e.g., **deontic cognitive event calculi**, such as $\mathcal{D}^e\mathcal{CEC}$) (Bringsjord & Govindarajulu 2013), are inadequate. For it can be seen that for instance the formal language and proof theory for $\mathcal{D}^e\mathcal{CEC}$, shown in Figure 3, contains no provision for the super/suberogatory.

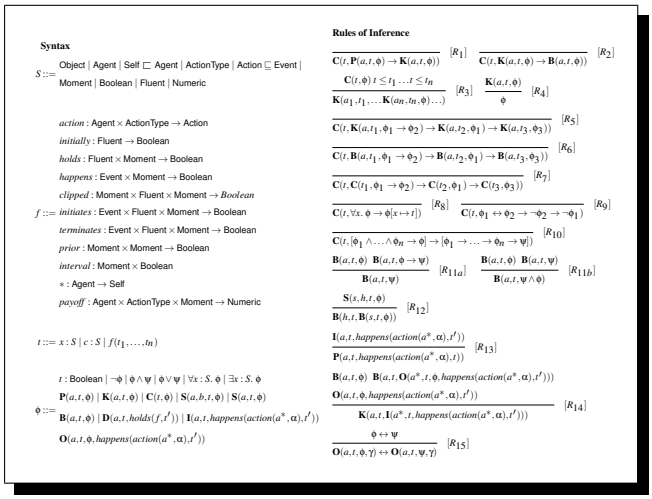

Fig. 3. $\mathcal{D}^e\mathcal{CEC}$ Syntax and Rules of Inference

I can offer only a few remarks about how the inadequacies in question are met in $\mathcal{L}_{\mathscr{EH}}$ (but see note 1). In keeping with the compressed notation employed above, I supress many of the elements that are in my lab's extant deontic logics.

There are obviously a host of formulae whose theoremhood constitute desiderata; that is (to give but a pair), the following must be provable (where $n \in \{1, 2\}$):

Theorem 1. $\mathbf{SuP}^n(\phi, a, \alpha) \rightarrow \neg\mathbf{O}(\phi, a, \alpha)$

Theorem 2. $\mathbf{SuP}^n(\phi, a, \alpha) \rightarrow \neg\mathbf{F}(\phi, a, \alpha)$

Secondly, $\mathcal{L}_{\mathscr{EH}}$ is an *in*ductive logic, not a deductive one. This must be the case, since, as we've noted, quantification 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. For example, while in standard natural deduction we have the inference schema

$$\frac{\forall x \phi}{\phi(\frac{x}{a})}$$

of universal elimination, how would such a thing work for the formaula $\mathtt{M}x\phi$? The answer is that in $\mathcal{L}_{\mathscr{EH}}$ strength factors are assigned to formulae (in keeping with the 9 strength factors in (Bringsjord, Taylor, Shilliday, Clark & Arkoudas 2008)), and every inference schema dictates the strength of inferred formulae from given formulae and the strength factors that they have. Standard inference schemata like universal elimination simply follow the "weakest link" principle.

Third, and I shall stop here, $\mathcal{L}_{\mathscr{EH}}$ 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 agent a falls in a particular portion of $\mathscr{EH}$. So for example, one hopes that those charged with engineering robots for sensitive operations in the military and medicine manage to engineer robots occupying the $\forall$ portion of the $\mathcal{O}$ portion of $\mathscr{EH}$; that is, one hopes that all robots r engineered by such people are such that $\mathcal{O}^{\forall}(r)$ holds, where

$$\mathcal{O}^{\forall}(r) \leftrightarrow \forall\phi\forall\alpha[\mathbf{O}(\phi, a, \alpha) \rightarrow happens(\alpha)]$$

8. A NOTE ON VACUOUS QUANTIFICATION AND $\mathscr{EH}$

It's quite important to note that some variants of our original question are trivial, because it's trivial to prove that an answer to them is correct. [11] (I'm indebted to Alexander Bringsjord for stimulating my coverage of this point.) I steered clear of considering for instance this trivial question:

(Q1) Can humans build some robots that will be more moral than some humans?

Given $\mathscr{EH}$, it's easy to prove that the corrrect answer to this question is in the affirmative. But no one should be aiming to build such morally mediocre robots; doing so is easy, and ultimately dangerous. Why is the answer to (Q1) "Yes"? The reasoning is simple in the context of $\mathscr{EH}$. Clearly, it's a brute empirical fact that there exist humans falling within the $\mathtt{M}$ portion of the $\mathcal{S}^{up2}$ portion of $\mathscr{EH}$. (Any number of nefarious villains from human history fit the bill.) And yet, given what I have said about the SIFC quartet, it's logically impossible for any robot to place this low in $\mathscr{EH}$. Perhaps more pragmatically put, using techniques promoted by myself and others, it seems easy enough (given sufficient funding) to engineer a robot that falls within the $\mathtt{M}$ portion (or *at least* the $\mathtt{V}$ portion) of the $\mathcal{O}^{M}$ portion of $\mathscr{EH}$. But this seems insignificant within the overall landscape of robot ethics.

And now here is a variant of the original question that seems quite important:

(Q2) Can we engineer robots that meet *all* of their legal and moral obligations?

The answer to this one is Yes, and this is the question-answer pair that I see myself working toward demonstrating. But if what has been said above is correct, this is insufficient, because supererogatory actions must be performed as well.

A final point: Obvioulsy, I interpreted (Q) in such a way that it's logically equivalent to:

(Q′) Can humans build some robots that will be more moral than all humans?

which is in turn equivalent to:

(Q″) Can humans build some robots that will be more moral than the overall class (or capacity) of humans?

The answer to (Q′) and (Q″), again, for reasons given, is firmly in the negative.

9. CONCLUSION; FUTURE WORK

The hierachy $\mathscr{EH}$ has only been sketched in the present, short paper; that will by now be clear to all readers. [12] The goal here has been to throw light on robot/machine ethics, revealing deep inadequacies (e.g., that of basing work on on the incomplete and naïve tripartite hierarchy $\mathscr{T}$ now superseded (at least in my lab) by $\mathscr{EH}$. Obviously, then, future work must include full specification of $\mathscr{EH}$; and just as obviously, future work must include as well the concomitant specification, and indeed implementation, of $\mathcal{L}_{\mathscr{EH}}$.

Please allow me to conclude by saying that future work undertaken in response to $\mathscr{EH}$ shouldn't, in my opinion, be confined to my own work, and those in my laboratory. I strongly suggest that other researchers working in machine/robot ethics branchi out, within their preferred methodology, to the super/suberogatory. For example, Bello (2005, 2013) could consider extending the reach of computational cognitive modeling to cover cognition associated with the parts of $\mathscr{EH}$ not present in $\mathscr{T}$. Pereira (forthcoming) and colleagues

[11] Those familiar with the quantifier-based version of the Arithmetic Hierarchy will wonder whether $\mathscr{EH}$ can likewise be built crisply via layered quantification. The answer, it seems to me, is Yes.

[12] There are in fact two deep lacunae in what has been presented: two sub-parts of the hierarchy that are flat-out missing, one toward the endpoint of moral perfection, and one toward the endpoint of the diabolical. Both lacunae pertain to *intelligence*: it seems at least *prima facie* untenable to leave the level of intelligence of ethical agents out of systematic investigation of a continuum of ethical "grade."

could consider extending the reach of their powerful logic-programming paradigm to model the parts of morality reflected in $\mathscr{EH}$. And while Arkin's (Arkin 2009) underpinnings have unfortunately hitherto been firmly in $\mathscr{T}$, and his focus has hitherto been also unfortunately firmly on laws, he should consider working from the broader underpinning of $\mathscr{EH}$ and its new sub-categories. [13]

ACKNOWLEDGEMENTS

Bringsjord is profoundly grateful for support provided by two grants from U.S. ONR to explore robot ethics, and to co-investigators M. Scheutz (PI, MURI; Tufts University), B. Malle (Co-PI, MURI; Brown University), M. Sei (Co-PI, MURI; RPI), and R. Sun (PI, Moral Dilemmas; RPI) for invaluable collaboration of the highest order.

REFERENCES

Arkin, R. (2009), *Governing Lethal Behavior in Autonomous Robots*, Chapman and Hall/CRC, New York, NY.

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

Bello, P. (2005), Toward a Logical Framework for Cognitive Effects-based Operations: Some Empirical and Computational Results, PhD thesis, Rensselaer Polytechnic Institute (RPI), Troy, NY.

Bello, P. & Bringsjord, S. (2013), 'On How to Build a Moral Machine', *Topoi* **32**(2), 251–266. Preprint available at the URL provided here.
URL:
http://kryten.mm.rpi.edu/Topoi.MachineEthics.finaldraft.pdf

Block, N. (1995), 'On a Confusion About a Function of Consciousness', *Behavioral and Brain Sciences* **18**, 227–247.

Bringsjord, S. (1992), *What Robots Can and Can't Be*, Kluwer, Dordrecht, The Netherlands.

Bringsjord, S. (1999), 'The Zombie Attack on the Computational Conception of Mind', *Philosophy and Phenomenological Research* **59**(1), 41–69.

Bringsjord, S. (2007), 'Offer: One Billion Dollars for a Conscious Robot. If You're Honest, You Must Decline', *Journal of Consciousness Studies* **14**(7), 28–43.
URL:
http://kryten.mm.rpi.edu/jcsonebillion2.pdf

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_abtoc_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., 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. M¨ller, 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. & Govindarajulu, N. S. (forthcoming), 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, Cham, Switzerland.
URL:
http://kryten.mm.rpi.edu/NSG_SB_Ethical_Robots_Op_Sys_0120141500.pdf

Bringsjord, S., Kellett, O., Shilliday, A., Taylor, J., van Heuveln, B., Yang, Y., Baumes, J. & Ross, K. (2006), 'A New Gödelian Argument for Hypercomputing Minds Based on the Busy Beaver Problem', *Applied Mathematics and Computation* **176**, 516–530.

[13] Within the robot-ethics project of which my logicist work is a part (see ACKNOWLEDGEMENTS), the empirical investigation of moral competence led by Malle can perhaps explore "norms" that cover not only what might naturally be classified within deontic logics as obligations, but also what conventional attitudes toward both levels 1 and 2 of supererogation in $\mathscr{EH}$. I wonder whether for example the everyday concept of blame, under exploration by Malle, Guglielmo & Monroe (2012), extends to supererogation.

Bringsjord, S. & Licato, J. (forthcoming*a*), 'By *Dis*analogy, Cyberwarfare is Utterly New', *Philosophy and Technology* .
URL: *http://kryten.mm.rpi.edu/SB_JL_cyberwarfare_disanalogy_DRIVER_final.pdf*

Bringsjord, S. & Licato, J. (forthcoming*b*), 'Crossbows, von Clauswitz, and the Eternality of Software Shrouds: Reply to Christianson', *Philosophy and Technology* .
URL: *http://kryten.mm.rpi.edu/SB_JL_on_BC.pdf*

Bringsjord, S. & Taylor, J. (2012), The Divine-Command Approach to Robot Ethics, *in* P. Lin, G. Bekey & K. Abney, eds, 'Robot Ethics: The Ethical and Social Implications of Robotics', MIT Press, Cambridge, MA, pp. 85–108.
URL: *http://kryten.mm.rpi.edu/Divine-Command_Roboethics_Bringsjord_Taylor.pdf*

Bringsjord, S., Taylor, J., Shilliday, A., Clark, M. & Arkoudas, K. (2008), Slate: An Argument-Centered Intelligent Assistant to Human Reasoners, *in* F. Grasso, N. Green, R. Kibble & C. Reed, eds, 'Proceedings of the 8th International Workshop on Computational Models of Natural Argument (CMNA 8)', University of Patras, Patras, Greece, pp. 1–10.
URL: *http://kryten.mm.rpi.edu/Bringsjord_etal_Slate_cmna_crc_061708.pdf*

Bringsjord, S. & Zenzen, M. (2003), *Superminds: People Harness Hypercomputation, and More*, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Chellas, B. F. (1980), *Modal Logic: An Introduction*, Cambridge University Press, Cambridge, UK.

Chisholm, R. (1963), 'Contrary-to-Duty Imperatives and Deontic Logic', *Analysis* **24**, 33–36.

Chisholm, R. (1982), Supererogation and Offence: A Conceptual Scheme for Ethics, *in* R. Chisholm, ed., 'Brentano and Meinong Studies', Humanities Press, Atlantic Highlands, NJ, pp. 98–113.

Chisholm, R. (1986), *Brentano and Intrinsic Value*, Cambridge University Press, Cambridge, UK.

Clark, M. (2008), Cognitive Illusions and the Lying Machine, PhD thesis, Rensselaer Polytechnic Institute (RPI).

Davis, M., Sigal, R. & Weyuker, E. (1994), *Computability, Complexity, and Languages: Fundamentals of Theoretical Computer Science*, Academic Press, New York, NY.

Ganascia, J.-G. (2007), 'Modeling Ethical Rules of Lying with Answer Set Programming', *Ethics and Information Technology* **9**, 39–47.

Ladd, J. (1957), *The Structure of a Moral Code*, Harvard University Press, Cambridge, MA.

Malle, B. F., Guglielmo, S. & Monroe, A. (2012), Moral, Cognitive, and Social: The Nature of Blame, *in* J. Forgas, K. Fiedler & C. Sedikides, eds, 'Social Thinking and Interpersonal Behavior', Psychology Press, Philadelphia, PA, pp. 313–331.

Powers, T. (2006), 'Prospects for a Kantian Machine', *IEEE Intelligent Systems* **21**, 4.

Saptawijaya, A. & Pereira, L. M. (forthcoming), Ethical Regulation of Robots Must be Embedded in Their Operating Systems, *in* R. Trappl, ed., 'The Potential of Logic Programming as a Computational Tool to Model Morality', Springer, Cham, Switzerland.
URL: *http://centria.di.fct.unl.pt/ lmp/publications/online-papers/ofai_book.pdf*

Schermerhorn, P., Kramer, J., Brick, T., Anderson, D., Dingler, A. & Scheutz, M. (2006), DIARC: A Testbed for Natural Human-Robot Interactions, *in* 'Proceedings of AAAI 2006 Mobile Robot Workshop'.

Scheutz, M. & Arnold, T. (forthcoming), 'Feats Without Heroes: Norms, Means, and Ideal Robotic Action', *Frontiers in Robotics and AI* .

Schmitt, M., ed. (2013), *Tallinn Manual on the International Law Applicable to Cyber Warfare*, Cambridge University Press, Cambridge, UK. This volume was first published in 2011. While M Schmitt is the General Editor, there were numerous contributors, falling under the phrase 'International Group of Experts at the Invitation of the NATO Cooperative Cyber Defence Centre of Excellence'.

Urmson, J. O. (1958), Saints and Heroes, *in* A. Melden, ed., 'Essays in Moral Philosophy', University of Washington Press, Seattle, WA, pp. 198–216.

von Wright, G. (1951), 'Deontic Logic', *Mind* **60**, 1–15.

Winfield, A., Blum, C. & Liu, W. (2014), Towards an ethical robot: Internal models, consequences and ethical action selection, *in* M. Mistry, A. Leonardis, M. Witkowski & C. Melhuish, eds, 'Advances in Autonomous Robotics Systems', Vol. 8717 of *Lecture Notes in Computer Science (LNCS)*, Springer, Cham, Switzerland, pp. 85–96.

Youpa, A. (2013), Leibniz's Ethics, *in* E. Zalta, ed., 'The Stanford Encyclopedia of Philosophy'.
URL: *http://plato.stanford.edu/entries/leibniz-ethics*