The (Uncomputable!) Meaning of Ethically Charged Natural Language, for Robots, and Us, From Hypergraphical Inferential Semantics

Selmer Bringsjord, James Hendler, Naveen Sundar Govindarajulu, Rikhiya Ghosh, and Michael Giancola

Abstract The year is 2030. A two-young-child, two-parent household, the Rubensteins, owns and employs a state-of-the-art household robot, Rodney. With the parents out, the children ask Rodney to perform some action α that violates a Rubensteinian ethical principle P_R. Rodney replies: (s_1) "Doing that would be (morally) wrong, kids." The argument the children give Rodney in protest is that another household, the Müllers, also has a robot, Ralph; and the kids argue that *he* routinely performs α. As a matter of fact, Ralph's doing α violates no Müllerian ethical principle P_M. Ralph's response to the very same request from the children he tends is: (s_2) "Okay, doing that is (morally) fine, kids." What is the *meaning* of the utterances made by Rodney and Ralph? We answer this question by presenting and employing a novel, formal, inferential theory of meaning in natural language: *hypergraphical inferential semantics* ($\mathscr{HIS}$), which is in the general spirit of **proof-theoretic semantics**, which is in turn antithetical to Montagovian **model-theoretic semantics**. $\mathscr{HIS}$,

Selmer Bringsjord
Rensselaer Polytechnic Institute (RPI), Rensselaer AI & Reasoning (RAIR) Lab, Department of Computer Science, Department of Cognitive Science, Troy, NY, USA.
e-mail: selmer.bringsjord@gmail.com

James Hendler
Rensselaer Polytechnic Institute, Tetherless World Constellation, Department of Computer Science, Department of Cognitive Science, Troy NY 12180, USA. e-mail: hendler@cs.rpi.edu

Naveen Sundar Govindarajulu
Rensselaer Polytechnic Institute, Rensselaer AI & Reasoning Lab, Troy NY 12180, USA.
e-mail: naveen.sundar.g@gmail.com

Rikhiya Ghosh
Rensselaer Polytechnic Institute, Rensselaer AI & Reasoning Lab, Troy NY 12180, USA.
e-mail: rikrixa@gmail.com

Michael Giancola
Rensselaer Polytechnic Institute, Rensselaer AI & Reasoning Lab, Department of Computer Science, Troy NY 12180, USA. e-mail: mike.j.giancola@gmail.com

applied even to sentences logically simpler than s_1 and s_2, implies that human-level natural language understanding (NLU) is Turing-uncomputable.

1 Introduction

The year is 2030. A two-young-child, two-parent household, the Rubensteins, owns and employs a state-of-the-art household robot, Rodney. With the parents out, the children ask Rodney to perform some action α that violates a Rubensteinian ethical principle P_R. Rodney replies: (s_1) "Doing that would be (morally) wrong, kids." The rationale the children give Rodney is that another household, the Müllers, also has a robot, Ralph; and the kids argue that *he* routinely performs α. As a matter of fact, Ralph's doing α violates no Müllerian norm N_M. Ralph's response to the very same request from the children he tends is: (s_2) "Okay, doing that is (morally) fine, kids." We briefly explain herein how, given past work on our part, Rodney and Ralph would in general be engineered so as to respond in the (correct, for reasons explained) ways they do. But there is a separate issue, one that our prior work hasn't addressed; that issue is: What is the *meaning* of the utterances made by Rodney and Ralph? We answer this question by presenting and employing a novel, formal, inferential theory of meaning in natural language: *hypergraphical inferential semantics* ($\mathscr{HIS}$), which is in the general spirit of **proof-theoretic semantics**. HIS is based on a wholesale rejection of the dominant formal theory of the meaning of natural language: **model-theoretic semantics**, as seminally introduced by Montague [20]. We recommend that household robots (and *a fortiori* robots that frequently find themselves in morally charged situations, e.g. military robots) be engineered on the basis of the computational logics and corresponding procedures that underlie $\mathscr{HIS}$.

The remainder of our chapter unfolds in the following sequence. First, we present the case study involving robots Rodney and Ralph, and their respective families (§2). Next, in §3 we quickly explain how, given past work, Rodney and Ralph would in general be engineered. In §4 we very briefly summarize MTS, including — at least as the lead author sees things — some its fatal defects. The following section, §5, is a summary of proof-theoretic semantics for natural language, and a quick critique of of this approach to meaning, in the form of today's state of the art. We then (§6) present (for the very first time in any archival venue) hypergraphical inferential semantics = $\mathscr{HIS}$, albeit briefly. Next, we apply $\mathscr{HIS}$ to the Rodney-Ralph case study (§7). Section 8 is devoted to the consideration of objections to what has come before. We then come to what may be the most impactful part of the present chapter: We show in section 9 that the problem of determining the meaning of natural language such as s_1 and s_2 is not just challenging, and in fact not just possibly infeasible, but is in fact Turing-uncomputable. The chapter ends (§10) with a wrap-up, and an anticipatory look into the future regarding $\mathscr{HIS}$ both in general, and specifically in connection with machine ethics.

2 A Household-Robot Case Study

Rodney is a state-of-the-art English-speaking household robot of 2030, recently purchased by the Rubenstein family to help shop, cook, clean, and assist with various child-rearing tasks. Mr & Mrs Rubenstein are two 70-hours-a-week Manhattan corporate attorneys; they have active, exceptionally bright, and somewhat mischievous twins (Joel and Judith) who recently entered third grade in a nearby Upper-West-Side (secular) school: The Anderson School, for gifted and talented children (PS 334). The twins are also in Hebrew school two days a week, and have been since they started reading; both of their parents are modern-orthodox Jews, and are raising the twins in fastidious conformity with most of the tenets and practices embraced by adult members of this denomination.

Rodney has been tasked by the twins' parents at sunrise with grocery shopping and making dinner for the twins at 6pm. Their parents, who now leave for work, will be toiling late at their offices. Judith, perhaps testing Rodney a bit later at breakfast: "Rodney, can we please have Lobster Newberg for dinner today? My Christian friend at school says it's delicious, and their robot made it for them!" What, for Rodney and Ralph, resp., is the meaning of following three normative sentences?

——

s'_1 It is morally forbidden for Judith and Joel to have lobster for dinner, and for their robot Rodney to cook such lobster.

s'_2 It's morally permissible for some Anderson students to have lobster for dinner, and for their robot Ralph to make such a meal.

s_3 It is wrong for Judith and Joel to plan to have lobster for dinner, yet permissible for them to entertain having such a meal.

——

The third sentence here involves the moral status of mental acts, and is beyond the scope of the present chapter, the chief purpose of which is to introduce $\mathscr{HIS}$ in connection with both norms, and human-robot interaction, and to reveal that natural language understanding (NLU) in $\mathscr{HIS}$/ is Turing-uncomputable. But we do herein answer the question about the first two of the sentences here, by presenting and employing a novel semantics for such modal propositions in the general spirit of proof-theoretic semantics. Our approach covers the semantics of natural-language sentences such as those listed above. (Needless to say, any headway we make would be applicable to the meaning of norms associated with roughly parallel propositions that confront robots employed by Hindus, Christians, and Muslims, but also —- since the core challenge isn't restricted to divine-command ethical theories and codes —- atheists, utilitarians, and so on.

3 Prior Framework for Engineering of Robots Rodney and Ralph

Work by Bringsjord and Govindarajulu (and some collaborators with them) through the years that has been devoted to the science and engineering needed to achieve

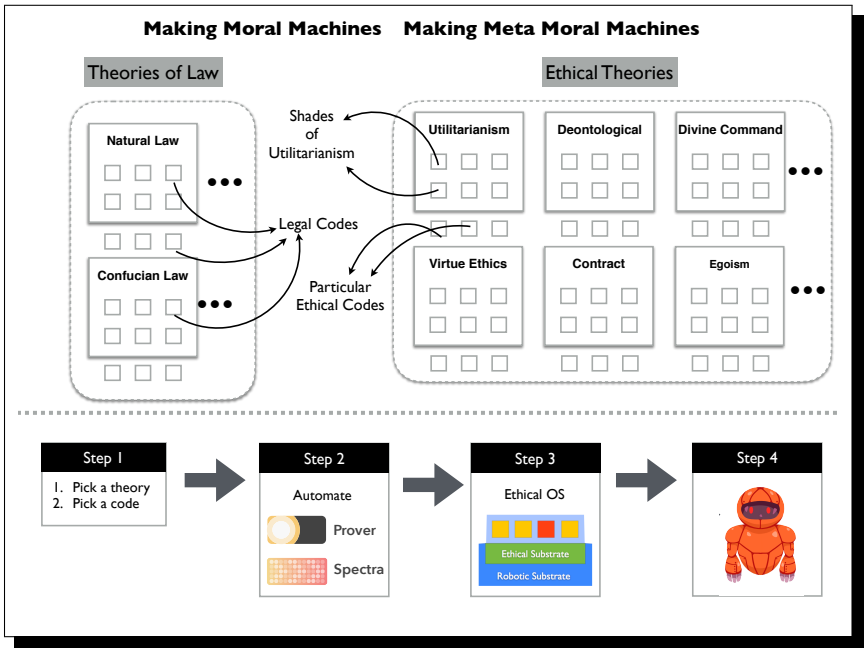

Fig. 1 The Four Steps in Making Ethically Correct Machines

ethically correct robots has been firmly logicist in nature. Overall this effort abides by the "Four-Step" approach shown from a high-level point of view pictorially in Figure 1. We now quickly summarize these four steps, in sequence, so that the reader will be in position to understand how $\mathscr{HIS}$ and NLP built upon it fits quite naturally with prior work.

The first step is selecting an ethical theory from a family thereof. We do not want to advance a framework that requires one to commit to any particular ethical theory or even to families of theories. In the case study at hand, we of course assume that Rodney's family and Ralph's family each affirm different (indeed inconsistent) ethical theories (even if only implicitly).

So, assume that we have a family of ethical theories $\mathscr{E}$ of interest. We assume that, minimally, any ethical theory $\mathcal{E} \in \mathscr{E}$ obligates or permits (i.e. sanctions) a set of situations or actions Π and forbids a set of other situations or actions Υ. When these situations become particular, we are dealing with a moral code X based on the theory $\mathcal{E}$; such codes are by definition domain-dependent. For example, both our families, the Rubeinsteins and the Müllers, have particular ethical codes governing diet.

Abstractly, assume that we have a formal system $\mathcal{F} = \langle \mathcal{L}, \mathcal{I} \rangle$ composed of a language $\mathcal{L}$ and a system of inference schemata (or a proof theory/argument theory) $\mathcal{I}$. The particular formal system, a so-called *cognitive calculus*, that has been much used in the past for modeling and simulating ethical reasoning and decision-making

in AIs and robots is $\mathcal{DCEC}$; see e.g. [13]. One non-negotiable *sine qua non* for the kind of calculus we need, one (as will be seen), directly relevant to determining the meaning of the two key sentences s'_1 and s'_2 from our case study, is that quantified formulae containing the deontic modal operator **O**, for 'is obligatory,' must be available.

The second of The Four Steps is to automate the generation of proofs of (un-)ethical behavior so that the reasoning can be utilized and acted upon by autonomous robots. We use ShadowProver [16], an automated reasoning system (among other things) tailor-made for use of $\mathcal{DCEC}$.

The third step is to integrate this ethical reasoning system into an autonomous robot's operating system, something that, longer term, we would insist upon for both Rodney and Ralph, were these robots of our own design. For reasons explained in [15], there are basically two possible approaches to this (see Figure 2). In the first, only "obviously" dangerous AI modules are restricted with ethical reasoning safeguards *installed above the OS*. In the second approach, and by our lights highly preferable one, all AI modules must be brought down to the robotic substrate (the percepts and actuators which enable the robot to interact with its environment) through an "Ethical Substrate" tied to the OS). The advantage of the first approach is speed: modules which are not inhibited by an ethical safeguard are able to directly manipulate the robot. However, this option also allows for the possibility that those AI modules deemed "not dangerous" may end up making a decision which leads the robot to act unethically. Only in the second option is ethical behavior guaranteed.[1]

In the fourth and final step, we implement, and thereby arrive at a moral machine, in the real world.

4 Montagovian/Model-Theoretic Approach to Meaning, Rejected

At least until today, by far the dominant approach to formally pinning down the meaning of natural language is **model-theoretic semantics** (MTS), seminally introduced and — at least to an impressive degree — specified by Montague [20]. In this section we quickly encapsulate MTS, and then explain why it must be rejected in light of its being plagued by a series of fatal defects.

4.1 MTS in Summary

We don't pretend that we can do justice to MTS here; but we say a few words, and hope they are helpful: MTS, in the case of formal logic, as we've already indicated, is Tarskian, and says that the meaning of formulae consist in the conditions for their being TRUE on interpretations, compositionally calculated. For instance, for any

[1] That is, ethical behavior relative to some ethical theory, and code selected therefrom.

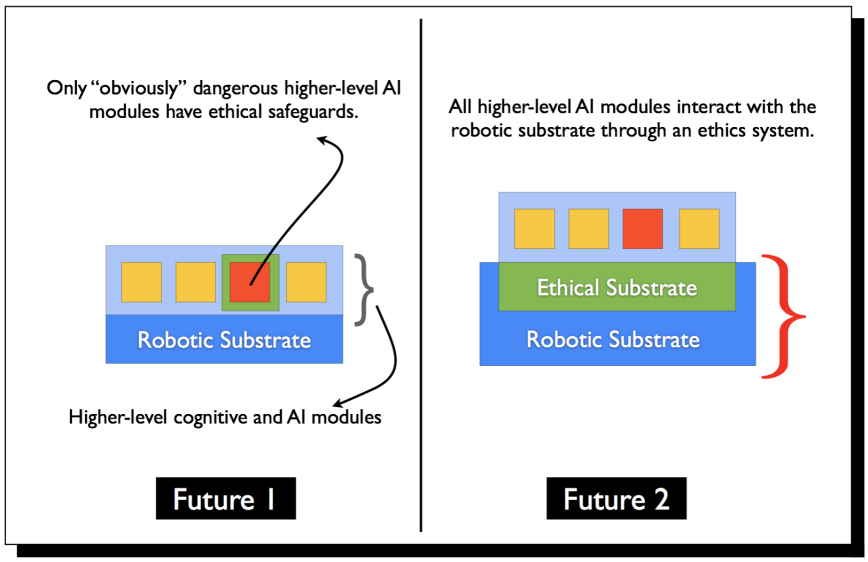

Fig. 2 Two Futures — With and Without an Ethical Substrate. *Higher-level modules are vulnerable to tampering. The Ethical Substrate protects the Robotics Substrate from rogue modules.* Figure from [15]

interpretation $\mathcal{I}$ whose domain of quantification includes no red things, the formula

$$\forall x(Red(x) \rightarrow Happy(x))$$

will be TRUE.[2]

What about MTS not for logic, but natural language? Well, actually, at least when it comes to the meaning of sentences in e.g. English, meaning is delivered in a manner akin to how it works for formulae. For instance, at least on the brand of MTS advanced by Montague himself, English nouns, verbs, and adjectives become relation symbols in a formal (logical) language, and the meanings of these relations are just suitable tuples of objects in some domain for some interpretation.

4.2 Two Fatal Defects in Model-Theoretic Semantics

At least according to Bringsjord, MTS is fatally flawed. Here, given space constraints, and given as well that the focus is on the presentation of $\mathscr{HIS}$ for purposes of handling the meaning of norms for robots like Rodney and Ralph, only two such

[2] Note that the relation symbol *Red* doesn't appear anywhere above in the present paper. That is as desired, because no matter what relation symbol R is used in a simple quantified conditional of the form we use here, if the domain of quantification has no non-empty class to which this R is mapped, the formula has a meaning of TRUE.

defects are mentioned, and both defects are at best synoptically communicated. Here is the pair:

1. *Everything Grounds Out in Proof/Argument*. MTS has its roots in what is known today in mathematical logic as *model theory*, which was, if not outright invented by Tarski (Montague's advisor), then at least brought to a respectable level of formality by him. In order to begin to see that model theory and the "meaning" it assigns to formulae distills to meaning cast in terms of inference, and specifically proofs or arguments, consider what the meaning of a simple material conditional

$$c \coloneqq \phi \rightarrow \psi$$

   is according to the model theory for first-order logic, where ϕ and ψ are wffs in the standard formal language of this logic.[3] We have that the meaning of c is either TRUE or FALSE; where the former case holds if and only if — and here we quote from relevant mathematical-logic coverage of model theory, e.g. [9] — "If (interpretation) $\mathcal{I}$ satisfies ϕ, then $\mathcal{I}$ satisfies ψ as well." So we have exchanged a formal material conditional for ... another conditional, one that is a mix of the formal and informal. Now what determines whether this hybrid conditional holds? Well, we need a meta-logical proof that this hybrid conditional is true (by e.g. supposing that its antecedent holds, and proving that based on this supposition the consequent holds as well). This is to say that the meaning of even a dirt-simple material conditional ends up being a matter of proof carried out at the meta-logical level, over the formal elements of model theory![4]
2. *Possible Worlds are at Best Merely Metaphorical, and at Worst Provably Incoherent*. While Kripke gets credit for seminally working out so-called "possible-world semantics" formally, Leibniz had an intuitive concept of, and wrote about, possible worlds. But what *is* a possible world? This question is distressingly hard to answer, for everyone — to this day. But the situation is actually worse, because some answers that were confidently proffered on the strength of basic set theory and consistency turned out to be *provably* incoherent [1]. Common practice is to

[3] Please note that MTS is certainly up to the challenge of producing meaning for things much, much more robust than material conditionals, but the point here is that *even* in the case of something as simple as a material conditional treated by model theory in standard, elementary, classical mathematical logic, meaning reduces to meaning in terms of inference. In addition, we are certainly aware of the obvious fact that no one working on formal semantics believes that material implication is a good representation of natural-language conditionals. But this fact is orthogonal to the point we are making here: that, again, meaning initially taken to be model-theoretic eventuates in meaning that is inferential in nature.

[4] The disappearance of the mirage that meaning can be at the level of models/model theory carries over *mutatis mutandis* directly to English. An instance of c in English might for instance be

$$c' \coloneqq \text{If Johnny helped, Olaf did too.}$$

The meaning of c' can't be in any model-theoretic basis. If it is said that c' is TRUE because it holds in some particular interpretation $\mathcal{I}^*$, one has only ask why this is the case. The only cogent response is to supply the relevant formal machinery and associated information (e.g. that as a matter of fact $\mathcal{I}^*$ renders 'Johnny helped' TRUE).

just take the concept of a possible world as an unanalyzable primitive, but then the obvious question is: How is it that meaning gets explicated in terms that, by definition, are not assigned a meaning?

5 Proof-Theoretic Semantics (PTS)

5.1 The Basic Idea

The basic idea behind PTS, at least in its modern form and officially speaking, originates with Gentzen [12], commonly regarded to be the inventor of natural deduction, and is not unfairly encapsulated thus: The meaning of elements of a proof, say for instance a constant a in a proof π, consists in the instantiation of inference schemata in π to introduce a. Once one grasps the basic insights of Gentzen, and the subsequent extensions of Prawitz [22], and combines these insights with what we have shown above about the unstoppable grounding out of model-theoretic truth/falsity in proof, it isn't long before those new to PTS, but well-versed in formal logic and mathematics, see at least the possibility of claiming that *all* meaning, at least for coherent declarative content, consists in the position of this content within proofs. Interestingly enough, professional mathematicians deal in proof top to bottom and beginning to end, and have for millennia, but know next to nothing about model theory in any form. This is often taken by advocates of PTS to be a tell-tale phenomenon; it certainly is by the lead author of the present paper. The body of technical literature on PTS is now vast, and we can say no more in terms of an overview, but (1) we direct interested readers to this starting place: [25];[5] and when we below speak about hypergraphical natural deduction and $\mathscr{HIS}$ itself, the reader will learn and understand more about PTS.

5.2 But What About PTS for Natural Language?

An impressive advance in proof-theoretic semantics for natural language has been achieved in Part II of [11]. (We do not have the space here to recount what is done in this work.) Unfortunately, as impressive as this book is, there are some serious inadequacies, especially in the context of our case study regarding robots Rodney and Ralph. Here are two such problems:

- *Deduction/Proofs Only*. As even readers new to formal semantics doubtless imagined when reading for the first time about MTS vs. PTS above, given that the 'P' in 'PTS' is for 'proof,' meaning on this approach must ultimately be cashed out by proofs and their constituents and use. But this dooms PTS at

[5] More philosophically inclined readers should without question read Dummet's [8] remarkable attempt to erect a theory of meaning along the PTS line.

the outset, for the simple reason that most reasoning engaged in, explicitly and implicitly, by humans, is non-deductive in nature. Consider for instance what the children say to Rodney in an attempt to persuade him to prepare Lobster Newberg. They give him not a proof, but an argument (one based on a failed analogy between two households).

- *No Operators for Ethical and Cognitive Phenomena*. As we have seen, we need to be able to speak about what is morally *obligatory*, *permissible*, and *forbidden* (minimally), and we certainly need to be able to speak about what agents believe and know (including about what other agents believe and know). But these needs, in formal logic, for reasons that can be expressed in the form of telling proofs [4], call for modal logic (in particular, resp., deontic modal logic & epistemic logic). Unfortunately, Francez [11] works with formal machinery that is devoid of modal operators, and for this reason sentences like our s'_1 and s'_2 can't be handled by his logical machinery.

6 Hypergraphical Inferential Semantics ($\mathscr{HIS}$)

In this section we provide a brief overview of a novel formal theory of meaning, Hypergraphical Inferential Semantics ($\mathscr{HIS}$), which is inspired and guided by hypergraphical reasoning — and also of course by the two inadequacies cited in the previous section. The overview proceeds as follows. We first (§6.1) convey, intuitively, the apparent brute fact that the meaning of natural language hinges on inferential context. Next, we give a very brief explanation of hypergraphical natural deduction (§6.2). We end the present section with an example, one in which the seemingly humble sentence 'Emma helped' is given agent-indexed meaning on $\mathscr{HIS}$ (§6.3).

6.1 Intuitive Kernel of $\mathscr{HIS}$ via Buffalo-buffalo-...

Consider this sentence:

(2) Buffalo buffalo buffalo.

What does (2) mean? You don't know. Upon some reflection, though, you will certainly find yourself entertaining some possibilities. Which of these possible meanings is what (2) means? For example, does (2) mean nothing; i.e. is it just three occurrences of the word that denotes the species of the animal *B. bison*, and nothing more? Maybe; but then again maybe not. Suppose we trustworthily tell you that (2) has been uttered somewhat slowly by Smith as he thinks back wistfully to a time when vast numbers of buffalo roamed proudly across portions of North America, before their rapid decline in the 1800s. In this case, (2), given what we have just told you and inferences made therefron, means something like:

(2_m) Smith believes that impressive and even glorious must have been the status of the mighty buffalo across the great midwestern plains and the foothills of the Rockies before heartlessly preyed upon by man!

One might wonder why Smith is thinking back to the "glory days" of American bison. There could of course be any number of reasons for his contemplation. For instance, suppose that Jones said the following, just before Smith utters (2):

(1) They can reach twelve feet in length, weigh over 2,000 pounds, and imagine horde upon horde of them in the wild, before the great slaughter, thundering sometimes in full, 40-miles-per-hour stampedes beneath the peaks of the Tetons, nothing to fear.

Given (1) beforehand, (2)'s meaning (2_m) is quite plausible — because there exists an obvious argument (which we don't detail) from (1) and other declarative information to (2_m). A bit more precisely, given (1), and background propositions about human psychology, aesthetics, and so on, that (2) means (2_m) is just to say that (2_m) is the conclusion of an argument. However, suppose instead that (1) was never uttered, but rather that our Smith is reading an article by an august naturalist in which this author claims that

(1′) Some buffalo in Buffalo hoodwink other buffalo.

and that upon taking this in, Smith murmurs a "Hmm" and a "So" to himself, and then says (2). The meaning of (2) is now nothing at all in the vicinity of (2_m). Instead, the meaning of (2) is (1′) = ($1'_m$) itself, and Smith has simply affirmed an argument from what he has read to the pinning down of meaning.[6]

The moral of all this talk of buffalo should be clear. It's that the meaning of natural-language sentences (at least frequently, and perhaps always) consists in their being within arguments (or, in more rigorous situations, proofs).

6.2 Hypergraphical Natural Deduction

We assume readers to be familiar with basic graph theory, and to therefore by acquainted by the *directed hypergraphs*. $\mathscr{HIS}$ takes the meaning of a natural-language sentence *s* to consist in the location of *s* within a (usually vast, in "real life") directed hypergraph that specifies interacting arguments and proofs. These graphs are dynamic, since human reasoning, as long noted in AI, is nonmonotonic; but in the present paper we ignore dynamism and worry only about meaning *at a particular time*. We also assume readers to be familiar with basic natural deduction in its standard forms. But now, what about *hypergraphical* natural deduction? And indeed, more broadly, hypergraphical reasoning? The basic concept of such formalisms date back to [7], but ignoring for economy here the development of these ideas through time, and the implemented proof- and argument-construction environments available

[6] It's interesting to note that any debate about the meaning of (2) in the contexts we have laid down will just end up providing further evidence for the view that meaning is inferential (since debate is, if anything, inference-based).

today, we greatly simplify and first draw your attention to an interesting quote from Schroeder-Heister, [25], who writes:

> One could try to develop an appropriate intuition by arguing that reasoning towards multiple conclusions delineates the area in which truth lies rather than establishing a single proposition as true. However, this intuition is hard to maintain and cannot be formally captured without serious difficulties. Philosophical approaches such as those by Shoesmith and Smiley (1978) and proof-theoretic approaches such as proof-nets (see Girard, 1987; Di Cosmo and Miller, 2010) are attempts in this direction. ([25], §3.5)

What is here declared to be just an "attempt" is made perfectly concrete in hypergraphical reasoning. Consider the pair of Figures 3 and 4, to which we draw your attention now.

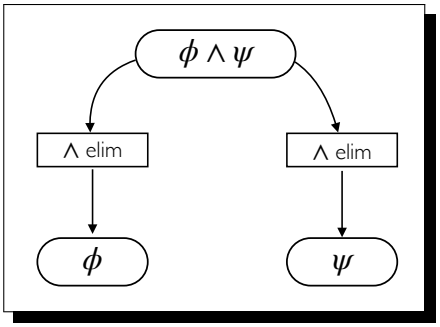

Fig. 3 Hypergraphical $\wedge$ `elimination`

Notice here that reasoning isn't linear: conclusions drawn as a result of inferences are in no way done one at a time in step-by-step fashion in a single list of formulae. On the contrary, what Schroeder-Heister indicates has "serious difficulties" has absolutely none at all. Multiple conclusions of ϕ and ψ in Figure 3 happens simultaneously in the directed hypergraph shown there. And of course in Figure 4, two premises, ϕ on the left and ψ on the right, lead *at once* in the graph to the conjunction.

6.3 An Example: The Meaning of 'Emma helped.'

What is the meaning of the two-word English sentence that immediately follows?

s Emma helped.

Given the foregoing, the reader knows that the initial, provisional answer advanced by at least the lead author of the present paper is: "Well, it depends on inferential context." Of course, this is a programmatic answer, not a genuinely informative one. Let us then set some context, by stipulating that two pieces of declarative information are givens for the agent who reads or hears *s*, to wit:

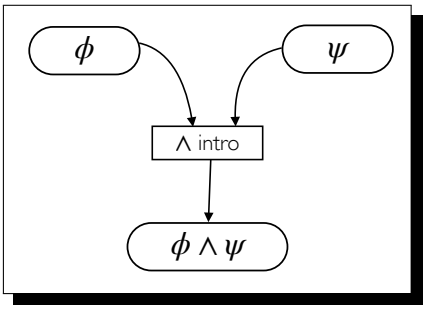

Fig. 4 Hypergraphical $\wedge$ `introduction`

G1 The following three propositions are either all true, or all false.

1. If Billy helped, Doreen helped.
2. If Doreen helped, Frank helped.
3. If Frank helped, Emma helped.

G2 Billy helped.

Now, in the context composed by givens G1 and G2, and specifically assuming that first G1 and G2 are assimilated, what is the meaning of s? The answer is still "It depends." The reason is that the meaning of s for a given agent $\mathfrak{a}$ who has taken in first both G1 and G2, and then s, will depend upon the hypergraphical natural deduction that has now formed inside s. For a rational agent, the meaning of s will correspond to the hypergraphical proof shown in Figure 5. Such an agent will be able to confidently report that given G1 and G2, s is true.

7 Applying $\mathscr{HIS}$ to the Robot Case Study

It should be rather clear to the reader at this point what the meaning of our featured sentences are. That meaning consists in a directed hypergraph, indexed to a particular agent, and anchored in the elements of the Four-Step Process shown in Figure 1. What elements? First, the relevant family of ethical theories is that of *divine-command* sort.[7] From this family a particular theory associated with the relevant sort of Judaism is selected, and from that is in turn selected a particular ethical code X. When this code is combined with background B declarative information, a proof, or at least an argument, for the formula that expresses s'_1 can be inferred. This

[7] In the context of machine ethics, the formalization of this family is explored in [5]. A seminal treatment, from the point of view of analytic philosophy and formal logic, of this family of ethical theories, in particular the sub-family associated with Judaism and Christianity, is given by Quinn [23].)

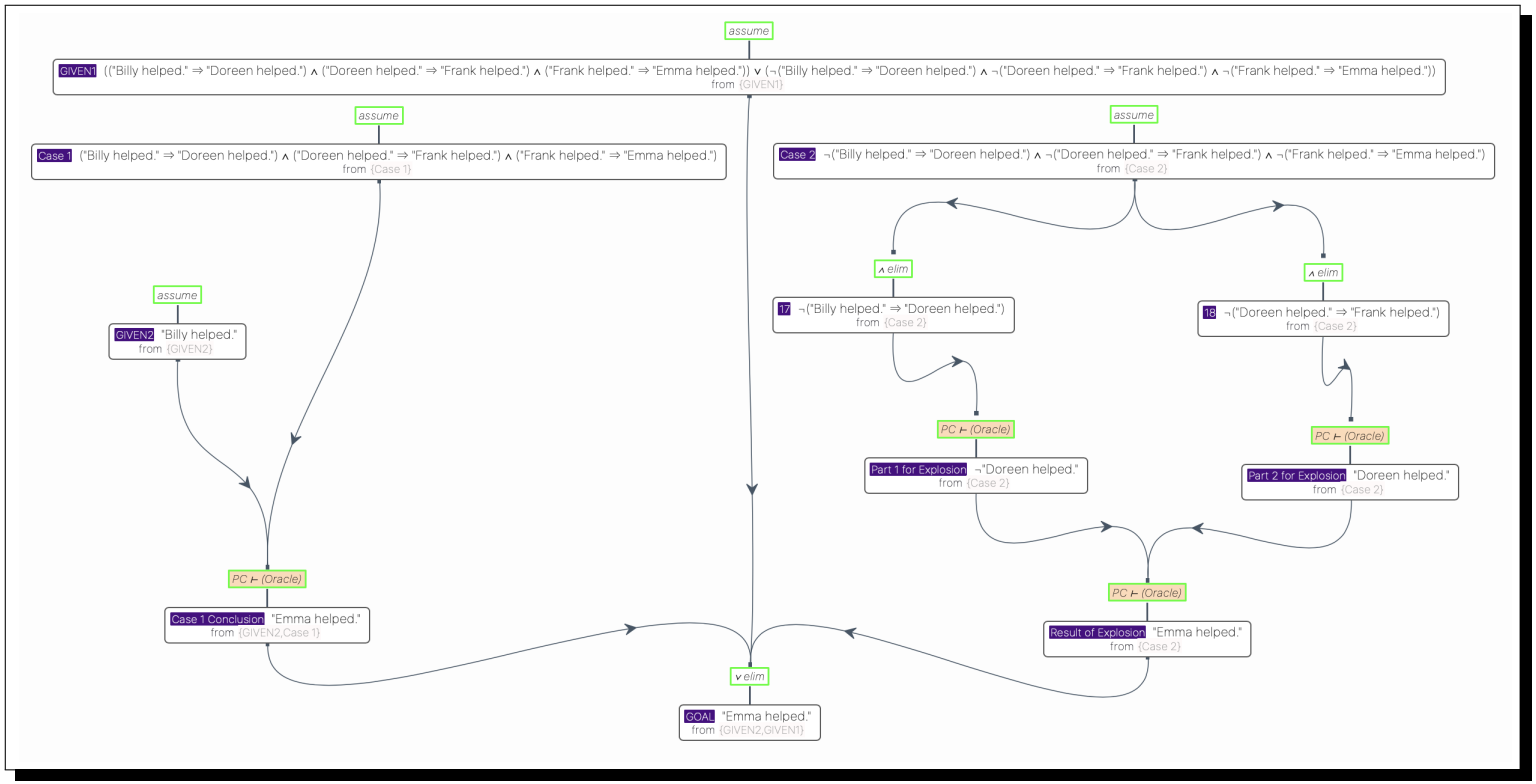

Fig. 5 Semi-Automated Proof of 'Emma Helped' From G1 and G2. *This is the meaning of 'Emma helped' for the agent who understands that, as a matter of fact, assuming both G1 and G2, Emma did help.*

formula is

$$\mathbf{O}\neg\phi(\alpha_{\text{LOBSTER}}^{\text{RODNEY}}),$$

where $\phi(\_,\_,\ldots,\_)$ is an open formula with "placeholders" for the relevant parameters in the case study. For a depiction of the overall situation, and the meaning of s_1' as a hypergraph, see Figure 6.

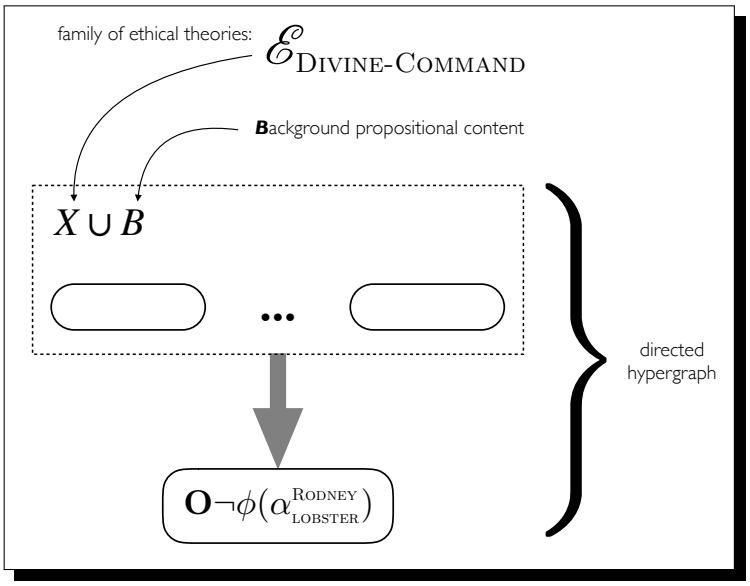

Fig. 6 The Meaning of the Sentence s_1' According to $\mathscr{HIS}$

Figure 7 shows a Python program which calls ShadowProver, which is able to find a fully-automated proof of sentence s_1' in under a second.

But what about Ralph? Why is it ethically permissible for him to whip up Lobster Newberg for the children he tends to? More to the matters at hand, what is the meaning of sentence s_2'? As alert readers can doubtless surmise, we have pretty much a direct parallel to what we've already seen in the case of Rodney — save for some obvious differences. First, of course the formula that is at the end of the relevant directed hypergraph expresses that the relevant culinary action is ethically permissible for him; this formula is:

$$\neg\mathbf{O}\neg\phi(\alpha_{\text{LOBSTER}}^{\text{RALPH}}),$$

The overall situation w.r.t. Ralph is shown in Figure 8, and the program generating the corresponding proof (via ShadowProver) in Figure 9. Again, ShadowProver found a fully-automated proof in under a second.

```
import sys
sys.path.insert(1, '/pylibs/interface')
import interface

from time import time


if __name__ == "__main__":
    assumptions = [ "(Knows! rodney t (CulinaryCode rubensteins))",

                    "(Knows! rodney t (implies (CulinaryCode rubensteins) \
                                               (forall [?f] (iff (ForbiddenFood ?f) \
                                                                 (not (Kosher ?f))))))",

                    "(Knows! rodney t (not (Kosher lobster)))",

                    "(Knows! rodney t (forall [?f] (implies (ForbiddenFood ?f) \
                                                            (Ought! rodney t (ForbiddenFood ?f) \
                                                              (not (happens (action (rodney (cook ?f))) t))))))"
                ]

    goal = "(Ought! rodney t (ForbiddenFood lobster) \
              (not (happens (action (rodney (cook lobster))) t)))"

    start = time()
    print(interface.prove(assumptions,goal))
    stop  = time()

    print("Time: " + str(stop - start) + " seconds.")
```

Fig. 7 An Automated Proof of the Sentence s'_1 in Accordance with $\mathscr{HIS}$, Found. (*Needless to say, this is inteneded to convey but the gist of what in its full, real-world version would be rather elaborate, since it would need to be aligned and integrated with the rigorous and subtle theological reasoning of relevant humans (e.g., rabbis). In addition, an argument is much more likely to ultimately be in play, rather than a proof; and probability/likelihood, rather than only classical bivalence, would inevitably be in play as well.*)

8 Questions/Objections and Replies

We here consider some questions and objections, and reply to each, in short.

Redundancies & Irrelevancies:

"Generally, automatically-derived proofs are very complex structures, in no small part because they can contain a lot of redundancies and irrelevant steps. How do you deal with this when seeking to model communication between humans?"

True enough, an arbitrarily discovered proof in response to a query as to whether some formula ϕ can be derived from some starting collection Φ of formulae may not at all be streamlined, let alone minimal. But all that $\mathscr{HIS}$ posits is the existence of a proof or argument from the some relevant Φ to some relevant ϕ that corresponds to what is happening in given rational communication between humans. No reason is given here to think that this proof or argument cannot be found computationally, for an NLU system.

Termination:

"Your formalism is not completely specified but does seems rather powerful. Can you guarantee that the proof procedure always terminates?"

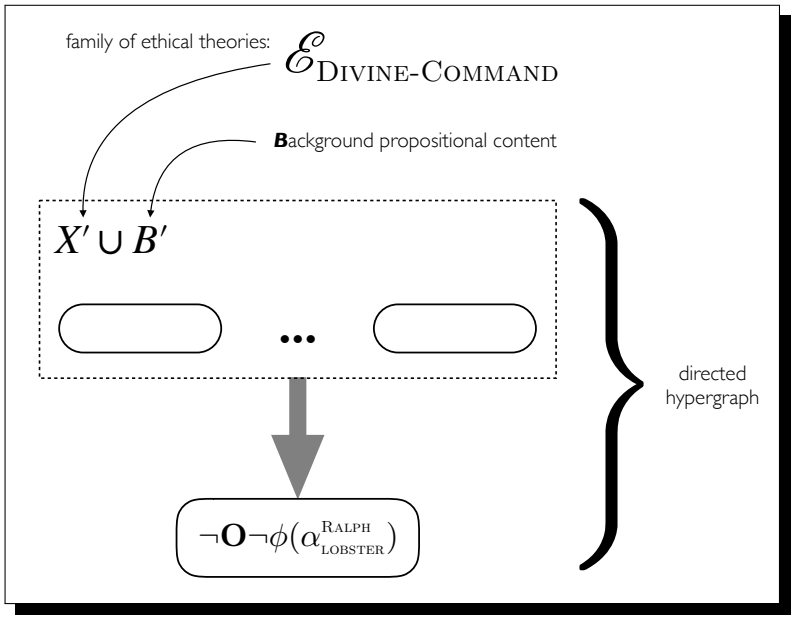

Fig. 8 The Meaning of the Sentence s'_2 According to $\mathscr{HIS}$

```
import sys
sys.path.insert(1, '/pylibs/interface')
import interface

from time import time


if __name__ == "__main__":
    assumptions = [ "(Knows! ralph t (CulinaryCode mullers))",

                    "(Knows! ralph t (implies (CulinaryCode mullers) \
                                              (forall [?f] (iff (ForbiddenFood ?f) \
                                                                (and (DuringLent t) (IsMeat ?f))))))",

                    "(Knows! ralph t (not (IsMeat lobster)))",

                    "(Knows! ralph t (forall [?f] (implies (not (ForbiddenFood ?f)) \
                                                           (not (Ought! ralph t (not (ForbiddenFood ?f)))  \
                                                             (not (happens (action (ralph (cook ?f))) t)))))))"
                  ]

    goal = "(not (Ought! ralph t (not (ForbiddenFood lobster)) \
                 (not (happens (action (ralph (cook lobster))) t))))"

    start = time()
    print(interface.prove(assumptions,goal))
    stop  = time()

    print("Time: " + str(stop - start) + " seconds.")
```

Fig. 9 An Automated Proof of the Sentence s'_2 According to $\mathscr{HIS}$

Yes. But there is a caveat! By Church's Theorem, the *Entscheidungsproblem* is only semi-decidable; i.e., once we hit first-order logic, theoremhood is at best semi-decidable; and, as we have explained, we are well beyond first-order logic when the meaning of natural language is given inferentially. Nonetheless, termination for a Σ_1 process can be guaranteed to terminate by use of a timer cutoff, a technique no different than what's available on a contemporary smartphone. We refer to a countdown timer, which can be engaged in such a way that some action is performed

when the countdown ends. In parallel, we can simply pick some amount of time beyond which search is prohibited, and produce whatever the result is at the moment that time expires.

How are divine-command theories and codes formalized?

"When it comes to deriving norms, your presentation is insufficient, in my opinion. It would be really interesting to understand how the ethical code of the divine-command theory is formalized. And your presentation does not allow the reader to understand how the obligations are actually derived."

Undeniably, this objection opens up deep issues that can't possibly be treated in the present venue, which is after all chiefly intended to present $\mathscr{HIS}$. The reader will need to turn to some of our writings that are focused on machine ethics, not formal semantics, to obtain answers. For instance, the Doctrine of Double Effect is an ethical principle that is at the code level, not at the ethical-theory level; and the reader can consult [13] to see in detail how this principle is formalized. Put roughly, some proposition is at the ethical-theory level if it is a biconditional that provides the necessary and sufficient conditions according to which an action is obligatory (or forbidden, supererogatory, etc.) in general. In the case of divine-command ethical theories, one such theory is given in [23]; but many other divine-command theories are possible. Our robot Rodney would be working under one such theory. As to the code level, prohibitions regarding diet, which are of course central to the case study with which we began, are at that level. Code-level propositions pertain to particular actions or action classes, and their moral status; ethical-theory-level propositions are about how to define obligation and other concepts in general. Finally along this general line, it's important to understand that *any* family of theories can be used in our Four-Step approach, including consequentialist theories (including theories types of utilitarianism). Nothing in what we have said above precludes using $\mathscr{HIS}$ to systematize the meaning of theory-level or code-level propositions.

OS-Level Ethical Controls Mysterious:

"I'm not sure what is meant by 'installed above the OS,' and 'bring all AI modules down to the percepts and actuators which enable the robot to interact with its environment,' and here I quote what you have said above."

Addressing such concerns is out of scope here, and we must accordingly direct the reader to [14], since the main purpose of the present paper, again, is to explain how the meaning of moral obligations for machine, specifically household robots, can be determined by our general approach.

9 Natural Language Understanding is Provably Uncomputable

As promised above, we briefly show in this, the penultimate section for our chapter, that human-level natural language understanding (NLU), when construed in general as the problem of receiving some arbitrary natural language S at this level, along with associated content, and producing in response the meaning of S per $\mathscr{HIS}$, is Turing-uncomputable (hereafter just 'uncomputable' *simpliciter*).[8] For simplicity, but with no loss of generality, assume that S is composed of a finite set of sentences $s_1, s_2, \ldots, s_k$. In fact, still without loss of generality, and perhaps surprisingly to some readers, we can restrict our attention to the case in which S is composed only of buffalo sentences (= b-sentences); these are of course the type of sentences we discussed earlier in section 6.1. Now, however, we shall need to get more precise about b-sentences, and we start to do this by briefly considering formal grammars for such sentences.

We now fix our first exceedingly simple grammar $\mathcal{G}_b^{\mathscr{L}_{pc}}$. The purpose of the subscript here is obvious; the superscript indicates that the grammar in question is at the level of the propositional calculus. In this grammar, for a noun to denote the U.S. city of Buffalo, we use `Buffalo1`, and as a noun to denote the animals in question we employ `buffalo2` and — when majuscule is needed for the start of a sentence — `Buffalo2` as well. We additionally have for the verb in question `buffalo` (understood here, in keeping with standard English dictionaries, as "to intimidate by a display of power"), and here too we can if needed avail ourselves of the uppercase variant `Buffalo`.[9] For economy, we don't specify the grammar in BNF form, but such a specification should be obvious to the reader, and easily obtainable therefore.

Now let's move further toward establishing the theorem that the problem of arriving at meaning for an arbitrary b-sentence is uncomputable. To do this, note first that the following informal b-sentence is ambiguous:

s_3 *Buffalo buffalo buffalo buffalo.*

As to the context, we stipulate that it — for reasons beyond scope here — logically implies that the first word is a reference to the city of Buffalo that is very near Niagara Falls, the second to the animals in question, the third to our one and only verb, and the fourth and final word another reference to the animals. In our first formal grammar, then, sentence s_3 is apparently disambiguated as:

s_3' `Buffalo1 buffalo2 buffalo buffalo2.`

---

[8] Of course, there are an infinite number of accounts of the meaning of human-level natural language on which arriving at meaning becomes not only Turing-computable, but polynomial/**P**. The point in the present section, expressed by the theorem below, is that *if* it's correct that the meaning of some natural language in the human case is inferential in nature as per $\mathscr{HIS}$, *then* the problem of producing meaning from relevant input is uncomputable.

[9] Generally such a need only arises when we admit b-sentences that are *imperative* in nature (as e.g. in the command to a buffalo animal (or animals) that it intimidate by a display of power: "Buffalo buffalo!"). We focus exclusively on *declarative* buffalo sentences in the present section/chapter.

But note that, as a matter of fact, the ambiguity isn't resolved in the least here. Upon reflection, it should be clear why. What does (s_3') mean? Does it mean that *all* buffalo in the city of Buffalo buffalo *all* buffaloes? Or does it mean that *some* buffalo in the city of Buffalo buffalo *all* buffaloes? Clearly, even with only classical quantification in play implicitly, s_3/s_3' is ambiguous between four distinct candidate permutations. Let's then expand our formal grammar by moving to first-order logic = $\mathscr{L}_1$; we thus have — following the notation we have introduced — $\mathcal{G}_b^{\mathscr{L}_1}$, or simply $\mathcal{G}_b^1$. In this grammar, we allow `all` and `some`, with a direct match between these words and the two quantifiers $\forall$ and $\exists$ of $\mathscr{L}_1$. (Hence `some` is interpreted as "at least one.") Given this, here's one possible genuine disambiguation for s_3 and s_3':

s_3'' `All Buffalo1 buffalo2 buffalo all buffalo2.`

And here is the representation of s_3'' in first-order logic itself (with obvious use of abbreviatory relation symbols):

$\sigma_{(b1')}$ $\forall x[(B_1 x \wedge B_2 x) \rightarrow \forall y(B_2 y \rightarrow Bxy)]$

The reader should note that along this line, under the umbrella of $\mathscr{HIS}$ and our formalization of what NLU is, we quickly run into formulae that are not satisfied by any interpretation with a finite domain. The quickest way to explicitly see this is simply to note that, when it comes to even the grammar $\mathcal{G}_b^{\mathscr{L}_1}$ with a trivial expansion, we have sets of formulae in this category, for consider:

- `No buffalo buffalos itself.`
- `If a buffalo-1 buffalos a buffalo, and the buffaloed buffalo buffalos another buffalo-3, buffalo-1 buffalos buffalo-3.`
- `Every buffalo buffaloes some buffalo.`

This trio, when represented in first-order logic, cannot be satisfied by a finite model, since it's an isomorph of a well-known example from Kleene [18].

At this juncture we point out that with a base lexicon that is minuscule, what we are seeing is nonetheless the beginning of a progression out from this lexicon to a vast family of grammars. We don't have the space in the present chapter to define this family, but rest content with pointing out that it can be viewed as an infinite array that has increasingly complex languages appearing as the array builds out to the reader's right; see Figure 10. In this figure, 'ω' indicates some collection of modal operators; for more along this line of abstracting to modal operators of any sort, see [3].

Very well. Now what of the theorem we are seeking? Given the foregoing in the present section, and given as well how $\mathscr{HIS}$ has been defined, we simply note that to determine the meaning of S it must specifically be determined whether, from a set Σ of formulae in the relevant formal language for the relevant formal logic, one or more formulae $\sigma \in \Sigma$ is such that σ is provable from relevant background content conjoined with S itself. But this then enables us to easily establish what we are seeking:[10]

[10] The proof here exploits a connection to Church's Theorem. Surely there must be prior arguments made for the uncomputability of at least some aspects of human natural-language "computing," ones that rely on other established negative theorems in recursion theory (such as the Post Correspondence Problem?) — but the first author is currently unaware of any, despite considerable digging.

extensional (classical):	$\mathcal{G}_b^0$	$\mathcal{G}_b^1$	$\mathcal{G}_b^2$	$\mathcal{G}_b^3$	...
extensional (non-classical):	$\mathcal{G}_b^0$	$\mathcal{G}_b{}'$	$\mathcal{G}_b{}''$	$\mathcal{G}_b^{'''}$	...
intensional (classical):	$\omega\mathcal{G}_b^0$	$\omega\mathcal{G}_b^1$	$\omega\mathcal{G}_b^2$	$\omega\mathcal{G}_b^3$	...

Fig. 10 Part of the Infinte Array That Composes an Infinite Family of Buffalo Grammars. (*To visualize this as an infinite tree, imagine that the array is rotated clockwise 45 degrees.)*

Theorem: Uncomputability of Meaning (= UMT)

Theorem: Let s be some arbitrary grammatically correct sentence in a human-level natural language, and let σ_s^i be a representation of s in $\mathscr{L}_1$ from among an at-most countably infinite number of such representations. Then the meaning of s, by $\mathscr{HIS}$, i.e. some proof π or argument α, is uncomputable.

Proof: Suppose in particular that $s \in \mathcal{G}_b^1$, that $\mathbf{B}^s$ is the background propositional content for s, that X^s is specific, contextual information, and that for *reductio* the meaning of s is computable. Then for some particular k, whether

$$\mathbf{B}^s \cup X^s \vdash_{\pi/\alpha} \sigma_s^k$$

holds is computable. But this contradicts Church's Theorem. ■

9.1 What of Prior and Related Work?

To bring this section to suitable closure, we must address, at least briefly, an apparent incompatibility between the negative result we have obtained (i.e. the Uncomputability of Meaning theorem = UMT), and what some others have said regarding how much is demanded, computationally, for natural-language understanding (NLU). We must also point out that some approaches to NLU outside inferential semantics (as least avowedly so by what proponents of these approaches say; what might be the case formally is another matter) appear to be quite consistent with UMT. In this regard, we quickly mention three strands of related, prior research. In the first strand, NLU, while held to be "hard," is by definition computable because the level of difficulty is firmly within the Polynomial Hierarchy, and the background assumption appears to be some such proposition as a cognitive analogue to the Church-Turing Thesis. In the second strand of research, a very robust, cognitively realistic approach to NLU, we perceive at least apparent uncomputability in the general case. In the third strand, a connection is made between language acquisition formally modeled, and NLU. Here now is the (very brief) commentary on these three strands, respectively.

Descriptive Complexity

In general, descriptive complexity is the marriage of standard coverage of complexity theory with formal logic: formal logics (of the standard, extensional and bivalent sort) are used to describe the difficulty of computing functions, where — and this is crucial at present — difficulty consists in the size of demands for time and space to compute the functions in question. (Formal logics are also of course used in standard ways to specify both the Arithmetic ($\mathscr{L}_1$ used) and Analytic ($\mathscr{L}_2$ used) Hierarchies, which are dominated by problems that *aren't* computable.) From this perspective, according to which, by definition, all the functions in question are computable, in a recent paper the provides an insightful overview of the human mind and descriptive complexity, Pantsar [21] at least implicitly affirms the proposition that NLU is computable. What is the basis for the affirmation of such a proposition? As Pantsar nicely reports, a large part of that basis is Ristad's [24] claim that human-level natural-language computations are **NP**-complete — and hence by implication computable. As to not just a claim, but a *theorem* that forms part of the basis for the proposition as well, Pantsar cites Fagin's [10] theorem that a proper subset of full second-order logic = $\mathscr{L}_2$ suffices to describe **NP**. For the most part, all of this, and more, is orthogonal to the UMT result. Certainly none of this is at all a threat to, or even for that matter inconsistent with, UMT. The reason is simple: UMT is not rooted in anything like the *description* of problems in a given formal logic; rather, UMT is rooted in the treatment of meaning in terms of inference built from the proof- or argument-theory of formal logics. In this account of meaning, it's the inferential dimension of formal logics that is central, not a dimension relating to capture of functions by way of formulae, and — as is obvious from the rejection of Montagovian meaning issued above — not a dimension relating to model-based semantics.[11]

Computational Cognitive NLU

Here we take a recent volume, *Linguistics for the Age of AI* [19], as an exemplar. As far as we are aware, this work is the most robust use of cognition (computationally modeled) and declarative knowledge for NLU that takes the full challenge of real natural language (with e.g. all its ambiguity) seriously. The first author's contention is that NLU here is uncomputable, since the knowledge brought to bear in order to enable the understanding of some natural language is arbitrarily expansive in the general case, and the natural language to be understood is likewise unrestricted. If the knowledge here is captured by some set Γ in a formal logic, and if ultimately

[11] Some readers may naturally ask whether some direct treatments of natural language by symbolizing/representing that language in formal logics have been shown to lead to uncomputability. While this is far beyond our scope here, we find it noteworthy that in his treatment of quantification in human natural language, Szymanik (see e.g. [27]) explicitly treats the understanding of this language to be a computable affair. In fact, he appears to affirm the proposition that human/human-level cognition overall is Turing-computable, something that the first author has long rejected, and defended repeatedly in print (e.g. see [6, 2]).

there is at least a set of reasons for why some natural language S is to have some meaning expressible in a formal logic as formulae Φ_S, then it's not hard to see why the overall problem might be uncomputable, along the lines of our own framework for $\mathscr{HIS}$, and for UMT.

Computational Learning Theory

Computational learning theory (CLT), a firmly recursion-theoretic approach to machine learning (and hence radically different than today's data-driven "ML"), might seem to some readers to be quite relative to UMT. In CLT, the focus is on *language acquisition*; that is, the challenge is for some agent to acquire, through time, command over a language. The language in question is identified with a Type-0 grammar; this means that what is to be learned can simply be identified with a Turing machine. The *locus classicus* of CLT is [17]. This work is a litany of limitative theorems, including those that say that learning a Turing machine, based on perceiving only small, finite snippets of information regarding the machine in question, is not a computable challenge. But why might CLT be thought relevant to UMT? The reason is simply that someone might view NLU in a broader way than we do, to specifically include, first, the understanding of what grammar/Turing machine is in play, and then, following on that, something more specific, and specifically connected to how we define NLU. We in general certainly see the reasonableness of having such an extended conception of NLU, and we appreciate that CLT is exceptionally difficult, but our definition of NLU in inferential terms is such as to only make CLT vaguely relevant to the theorem UMT. In fact, when in CLT a given agent is given information about the target to be learning, no inference whatsoever is in play, and indeed the agent is said to be successfully learning the target if and only if hypotheses about what that target is are wrong only finitely many times in the limit. No proof or argument is in the picture at all.

10 Future Work; Further Objections

At this point $\mathscr{HIS}$ is admittedly really only a proto-theory of meaning, and only the basis for *future* NLU and NLG. There is long and hard work ahead. But fortunately, prior work by Bringsjord and colleagues in robot ethics (of the sort encapsulated in §3) is an ideal foundation upon which to develop and refine, mathematically, $\mathscr{HIS}$, and NLP algorithms and technology built upon this approach. The reason why is obvious: the approach to robot ethics in question is steadfastly and thoroughly proof- and argument-centric; and since on $\mathscr{HIS}$ the meaning of natural-language sentences are captured by proofs and arguments that contain then, and are expressed in the form of (inevitably vast) hypergraphs that express these interacting proofs and arguments, the fit is a most promising one.

A final word regarding additional objections to $\mathscr{HIS}$ and to building NLP upon automated reasoning and proof-/argument-checking with the relevant hypergraphs: We are under no such illusion as that additional objections will not be pressed against the formal semantics advanced above. Inevitably, many will object that the meaning of a natural-language sentence s is captured by "static," non-inferential content expressed in some formal logic (or, equivalently, in some knowledge-representation format, say frame-based representations). Such skeptics would do well to consider the longstanding fact that in formal logic, the capture of a given mathematical assertion a, which is part natural language and part formal language, has for many decades been known to be achievable not merely by producing a formula $\phi(a)$ in some formal language that *expresses* a, but by a formal theory Φ which is such that $\phi(a)$ is provable from Φ. A very nice presentation of the express-vs-capture distinction when applied to statements in mathematics is provided in Chapter 4 of [26].

Though as we readily admit there is much work to be done, we recommend that household robots (and *a fortiori* robots that as a matter of course frequently find themselves in morally charged situations, e.g. military robots) be engineered to process and interactively discuss norms with humans on the basis of the computational logics and corresponding procedures that underlie $\mathscr{HIS}$.

Acknowledgments

We are indebted to multiple commentators on the presentation at ICRES 2020 of a proper subset of work herein described, and on the older, smaller paper that the present one subsumes, augments, and refines. Without long-term support of research and development in the Rensselaer AI & Reasoning Laboratory that centers around inference, the formal semantics adumbrated in this paper would not exist, and we are very grateful for this support from AFOSR and ONR in the States. We are indebted to two anonymous referees for cogent comments on an earlier kernel of the present paper, and to both Bertram Malle and Paul Bello for *substantive* comments, advice, and critique.

References

[1] Bringsjord S (1985) Are There Set-Theoretic Worlds? Analysis 45(1):64

[2] Bringsjord S, Arkoudas K (2004) The Modal Argument for Hypercomputing Minds. Theoretical Computer Science 317:167–190

[3] Bringsjord S, Govindarajulu N (2020) The Theory of Cognitive Consciousness, and Λ (Lambda). Journal of Artificial Intelligence and Consciousness 7(1):155–181, URL `http://kryten.mm.rpi.edu/sb_nsg_lambda_jaic_april_6_2020_3_42_pm_NY.pdf`, The URL here goes to a preprint of the paper.

[4] Bringsjord S, Govindarajulu NS (2012) Given the Web, What is Intelligence, Really? Metaphilosophy 43(4):361–532, URL http://kryten.mm.rpi.edu/SB\NSG\Real\Intelligence\040912.pdf, This URL is to a preprint of the paper

[5] Bringsjord S, Taylor J (2012) The Divine-Command Approach to Robot Ethics. In: Lin P, Bekey G, Abney K (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

[6] 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

[7] Bringsjord S, Taylor J, Shilliday A, Clark M, Arkoudas K (2008) Slate: An Argument-Centered Intelligent Assistant to Human Reasoners. In: Grasso F, Green N, Kibble R, Reed C (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

[8] Dummett M (1991) The Logical Basis of Metaphysics. Duckworth, London, UK

[9] Ebbinghaus HD, Flum J, Thomas W (1994) Mathematical Logic (second edition). Springer-Verlag, New York, NY

[10] Fagin R (1974) Generalized First-order Spectra and Polynomial-time Recognizable Sets. In: RKarp (ed) Complexity of Computation, SIAM-AMS Proceedings, vol 7, pp 43–73

[11] Francez N (2015) Proof-theoretic Semantics. College Publications, London, UK

[12] Gentzen G (1935) Investigations into Logical Deduction. In: Szabo ME (ed) The Collected Papers of Gerhard Gentzen, North-Holland, Amsterdam, The Netherlands, pp 68–131, This is an English version of the well-known 1935 German version.

[13] Govindarajulu N, Bringsjord S (2017) On Automating the Doctrine of Double Effect. In: Sierra C (ed) Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17), International Joint Conferences on Artificial Intelligence, pp 4722–4730, DOI 10.24963/ijcai.2017/658, URL https://doi.org/10.24963/ijcai.2017/658

[14] Govindarajulu N, Bringsjord S, Sen A, Paquin J, O'Neill K (2018) Ethical Operating Systems. In: De Mol L, Primiero G (eds) Reflections on Programming Systems, Philosophical Studies, vol 133, Springer, pp 235–260, URL http://kryten.mm.rpi.edu/EthicalOperatingSystems_preprint.pdf

[15] Govindarajulu NS, Bringsjord S (2015) Ethical Regulation of Robots Must be Embedded in Their Operating Systems. In: Trappl R (ed) A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementa-

tions, Springer, Basel, Switzerland, pp 85–100, URL http://kryten.mm.rpi.edu/NSG_SB_Ethical_Reg_at_OS_Level_offprint.pdf

[16] Govindarajulu, Naveen Sundar (2016) ShadowProver. URL https://naveensundarg.github.io/prover/

[17] Jain S, Osherson D, Royer J, Sharma A (1999) Systems That Learn: An Introduction to Learning Theory, Second Edition. MIT Press, Cambridge, MA

[18] Kleene S (1967) Mathematical Logic. Wiley & Sons, New York, NY, I recommend a 2002 Dover unabridged republication of the original 1967 book from Wiley, if you cannot obtain the original book.

[19] McShane M, Nirenburg S (2021) Linguistics for the Age of AI. MIT Press, Cambridge, MA

[20] Montague R (1974) Formal Philosophy: Selected Papers of Richard Montague. Yale University Press, New Haven, CT, Note that this book is the seminal work of author Montague, but is edited by Thomason.

[21] Pantsar M (2021) Descriptive Complexity, Computational Tractability, and the Logical and Cognitive Foundations of Mathematics. Minds and Machines 31:75–98, URL https://doi.org/10.1007/s11023-020-09545-4

[22] Prawitz D (1972) The Philosophical Position of Proof Theory. In: Olson RE, Paul AM (eds) Contemporary Philosophy in Scandinavia, Johns Hopkins Press, Baltimore, MD, pp 123–134

[23] Quinn P (1978) Divine Commands and Moral Requirements. Oxford University Press, Oxford, UK

[24] Ristad E (1993) The Language Complexity Game. MIT Press, Cambridge, MA

[25] Schroeder-Heister P (2012/2018) Proof-Theoretic Semantics. In: Zalta E (ed) The Stanford Encyclopedia of Philosophy, URL https://plato.stanford.edu/entries/proof-theoretic-semantics

[26] Smith P (2013) An Introduction to Gödel's Theorems. Cambridge University Press, Cambridge, UK, This is the second edition of the book.

[27] Szymanik J (2009) Quantifiers in TIME and SPACE: Ccomputational Complexity of Generalized Quantifiers in Natural Language. University of Amsterdam, Amsterdam, The Netherlands, This is a dissertation published by Institute for Logic, Language and Computation, Dissertation Series DS-2009-01.