On Creative Self-Driving Cars: Hire the Computational Logicians, Fast*

Selmer Bringsjord[1] *&* **Atriya Sen**[2]
Rensselaer AI *&* Reasoning (RAIR) Lab[1,2]
Department of Computer Science[1,2]
Department of Cognitive Science[1]
Lally School of Management[1]
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA

0323161130NY

Contents

List of Figures

*Bringsjord is deeply grateful to Robert Trappl and OFAI for the opportunity to present in Vienna a less-developed version of the case made in the current paper, and to the vibrant audience feedback he received in the City of Music. Whatever deficiencies still remain in the case in question are due solely to oversights of Bringsjord's, ones possibly due in part to the ingestion of delectable Grüner Veltliner during his post-talk analysis of said feedback. The authors express their deep gratitude to ONR for support under a MURI grant devoted to investigating moral competence in machines, and to our brilliant collaborators (M. Scheutz, B. Malle, M. Si) in that grant. Special thanks are due to Naveen Sundar Govindarajulu for a trenchant review, and to Paul Bello and Alexander Bringsjord for subjecting our prose to their keen eyes. Finally, thanks also to Karin Vorsteher for supernatural patience and crucial contributions.

1 Introduction; Planned Sequence

There can be no denying that it's entirely possible for a car-manufacturing company like Daimler[1] to build and deploy self-driving cars without hiring a single logician, whether or not of the computational variety. However, for reasons we explain herein, a logician-less approach to engineering self-driving automobiles (and, for that matter, self-moving vehicles of any consequence, in general) is a profoundly unwise one. As we shall see, the folly of leaving aside logic has absolutely nothing to do with the standard and stale red-herring concern that self-driving cars will face exotic human-style ethical dilemmas the philosophers have a passionate penchant for.[2]

There are any number of routes we could take in order to explain and demonstrate our central claim; in the present paper we opt for one based on a particular, new line of reasoning: on what we dub "The Core Argument" (= $\mathcal{A}$). This argument starts by innocently asking whether a car manufacturer wishes to engineer self-driving cars that are **i**ntelligent, presumes an affirmative answer, next infers that since intelligence entails (a certain form X of) creativity, the sought-after cars must be X-**c**reative. $\mathcal{A}$ continues with the deduction of **a**utonomy from X-creativity; and, following on that, explains that the cars in question must also be **p**owerful. We thus arrive at the intermediary result that our rational maker of self-driving cars must — using the obvious acronym for the four properties in question — seek **ICAP** (pronounced "eye cap") versions of such cars. The Core Argument continues with an inference to the proposition corresponding to the imperative that is the sub-title of the current paper.

There are four specific things R1–R4 the following of this imperative will directly secure for car manufacturers from the hired computational logicians; herein we emphasize only one item in this quartet: **OS-rooted ethical control** of self-driving cars, R4. That is, the computational logicians must be recruited to design and implement logics that are connected to the operating-system level of ICAP cars, and that ensure these cars meet all of their moral and legal obligations, never do what is morally or legally forbidden, invariably steer clear of the invidious, and, when appropriate, perform what is supererogatory.

At this point, via its final inference, $\mathcal{A}$ delivers some new and bad news: If the ethical theory sitting atop and instantiating these obligations (and prohibitions, etc.) for self-driving cars is utilitarianism, it will be impossible to engineer an ethically correct self-driving car. This bad news has absolutely nothing to do with the standard and stale red-herring concern that self-driving cars will face exotic human-style ethical dilemmas the philosophers have a passionate penchant for (e.g. see Lin 2015). Should a self-driving car that faces an unavoidable choice between crashing into a group of five innocent pedestrians versus crashing head on into another car with a single passenger opt for the smaller disutility? How should a self-driving car that must choose between hitting a motorcyclist wearing a helmet versus one not wearing one behave? Such cases can be multiplied indefinitely, just by following dilemmas that the philosophers have pondered for millennia. But the bad news we deliver is much more serious than such far-fetched "trolley problems," which everyone agrees are vanishingly unlikely to materialize in the future; the bad news pertains to the vehicu-

[1] While we are aware of the technological prowess of Daimler in the self-driving sphere, 'Daimler' is here only an arbitrary stand-in for the many companies who are steadfastly aiming at engineering self-driving cars: General Motors, Ford, Google, Tesla, BMW, and on and on it goes. Formalists can regard 'Daimler' to be an arbitrary name (suitable for universal generalization) relative to the domain of companies operating in the self-driving-car sector.

[2] Such dilemmas are nonetheless fertile soil for investigating the nature of (untrained) human moral cognition (e.g. Malle et al. 2015), and for informally investigating the informal, intuitive basis of much law (e.g. Mikhail 2011). In addition, *experimental* philosophy (Knobe et al. 2012) certainly makes productive use of such dilemmas.

lar version of *the demandingness objection* to utilitarianism, an objection — wholly separate from matters in AI and robotics, and needless to say specifically from self-driving cars — presented in (Scheffler 1982). In a word, the demandingness objection to utilitarianism, at least utilitarianism of the standard "act" variety, implies that a human agent seeking to continuously meet her obligations under this ethical theory will be overwhelmed to the point of not being able to get anything done on the normal agenda of her life. It's probably safe to say that Daimler will not be keen on the idea of building self-driving cars that don't get done any of the things customers expect from such vehicles, such as simply getting their passengers from point A to point B. We thus end up learning of the fifth reason (i.e., R5, coming on the heels of R1–R4) for hiring computational logicians (and this applies not only to car manufacturers, but also to those setting law and public policy with respect to self-moving vehicles): to receive crucial help in dealing with the demandingness problem.

The sequence for the sequel is specifically as follows. We begin by presenting The Core Argument (§2), and then, immediately anticipating the objection to $\mathcal{A}$ that intelligence does not in fact entail creativity. Next, in section 3, we consider a series of possible replies to this objection — a series that rests content (in §3.5) with the position that intelligence entails that a specific form of creativity (what we call 'MacGyveresque' creativity, or, for short, **m-creativity**) should be possessed by any intelligent self-driving car. (We hence instantiate X to 'm.') While this type of creativity isn't exactly exalted, it's the kind of creativity we should indeed expect from self-driving cars. Our next step, still following the thread of $\mathcal{A}$, is to show that m-creativity implies autonomy (§4). We then (§5) explain that the likes of Daimler will need not only to engineer intelligent, creative, and autonomous self-driving cars, but also *powerful* ones. Having reached this point, we further explain that the aforementioned quartet R1–R4 must be provided for an ICAP self-driving car, and that only computational logicians can provide this quartet (§6). Our focus, as said, is herein on but one member of the quartet: R4: OS-rooted ethical control. Importantly, we explain that whereas some others naïvely view ethical control of self-driving cars from the point of views afforded by threadbare and fanciful human ethical dilemmas (again e.g. see Lin 2015), the real issue is that because ICAP self-driving cars will be making decisions that are alien to our moral cognition, on the strength of knowledge that is alien to our moral knowledge, at time-scales that are alien to our moral decision-making, the demandingness objection to utilitarianism has firm and worrisome traction when it comes to building such self-driving cars. A brief conclusion, with remarks about next research steps, wraps up the paper. These steps include what we have said is the fifth and final thing that must be provided by computational logicians: R5: an escape from the AI version of the demandingness objection.

2 The Proposed Argument

As we've said, the present paper is driven by The Core Argument = $\mathcal{A}$; that is, the chain of reasoning we adumbrated above, and which is laid out skeletally in Figure 2. We refer in $\mathcal{A}$ to Daimler here, but of course this is just an arbitrary stand-in (see note 1). Likewise, we denote by 'c,' to ease exposition, the arbitrary car around which argumentation and analysis revolves. And finally, as the reader can see in Figure 2, the argument is at this point schematic, since (among other reasons) the type X of creativity in question is left open.[3]

[3] $\mathcal{A}$ takes for granted elementary distinctions that unfortunately are sometimes not made even in the literature on self-driving cars and machine ethics. E.g., Lin writes:

> I will use "autonomous," "self driving," "driverless," and "robot" interchangeably. (Lin 2015, p. 70)

Figure 1: The Core Argument ($\mathcal{A}$)

Q1 Does Daimler want a truly intelligent car (c)?
- If "No," exit/halt.
- Otherwise: "Yes, of course."

C1 Hence c must be at least X-creative.

C2 Hence c must be autonomous.

Q2 Does Daimler want c to be powerful?
- If "No," exit/halt.
- Otherwise: "Yes, of course."

$\mathcal{HL}$ Hence Daimler must hire the computational logicians for four reasons: to provide, with respect to c,

R1 verification,

R2 transparency,

R3 self-explanatory capacity, and

R4 OS-rooted ethical control.

C4 And hence, finally, because they will need to help Daimler handle the demandingness problem, the computational logicians will be needed for a fifth reason (R5).

We anticipate that some readers will consider the very first inference in $\mathcal{A}$ to be suspicious; in fact, many readers will no doubt regard this inference to be an outright *non sequitur*. Why should it be the case that intelligence entails creativity? As the skeptic will no doubt point out, when we speak of an intelligent self-driving car, we are not thereby speaking of the kind of intelligence one might ascribe to a towering human genius. Einstein was surely intelligent, and Einsteinian intelligence, all should concede, entails creativity, indeed extreme creativity at that, but however wondrous a 2020 "Autonomous Class" Mercedes-Benz might be, it's rather doubtful that Daimler needs the car to revolutionize an entire branch of science. Yet, we accept that the onus is on us to defend the first inference to C1 in The Core Argument.

3 Creativity in What Sense?

To bear this burden successfully means that, at a minimum, we need to explain what kind of creativity we have in mind, and then proceed to show that with that type of creativity selected for interpretation of the first inference in $\mathcal{A}$, that inference is valid. Much of Bringsjord's own work in AI has involved creativity, so we have no shortage of candidate kinds of creativity to consider for

This conflation may be convenient for some purposes, but logically speaking it makes little sense, given for instance that many systems operating independently of human control and direction over extended periods of time are not autonomous. Someone might insist that, say, even an old-fashioned, mechanical mouse trap is autonomous (and hence so is, say, a mine running a simple computer program), but clearly this position makes no sense unless autonomy admits of degrees. We encapsulate below (§4.3) a degree-based concept of autonomy that can undergird $\mathcal{A}$. On could-have-done-otherwise accounts of autonomy (briefly discussed in §4.3), neither a mouse trap nor a mine is autonomous nor a "driverless" car is an autonomous car.

X. Let's see how our inference fares on a series of these candidates.

3.1 Lovelace-Test Creativity

To start the series, we note that Bringsjord, Ferrucci & Bello (2001) propose a highly demanding test for (computing-)machine creativity, one inspired by Lady Lovelace's famous objection to Turing's (1950) claim that a computer able to successfully play the famous imitation game[4] should be classified as a genuinely thinking thing. Her objection, quite short but quite sweet, was simply that since computers are after all just programmed by humans to do what they do, these computers are anything *but* creative. Just as a puppet receives not a shred of credit for its moves, however fancy, but rather the human puppeteer does, so too, by the lights of Lovelace, it is only the human programmer who can credibly lay claim to being creative. Bringsjord et al. (2001) give conditions which, if satisfied by an AI system, should classify that system as creative even by the high standards of Lady Lovelace. An AI agent that meets these conditions can be said to be **LT**-creative. One condition, put informally here so as not to have to profligately spend time and space recapitulating the paper by Bringsjord and colleagues, is that the engineers of the AI system in question find the remarkable behavior of this system to be utterly unfathomable, despite these engineers having full and deep knowledge of the relevant logic, mathematics, algorithms, designs, and programs.

It should be obvious that the inference to C1 in $\mathcal{A}$ is invalid if 'X-creativity' is set to 'LT-creativity.' We can have any number of genuinely intelligent AIs that are nonetheless fully understood by engineers who brought these AIs into existence. We doubt very much that Deep Blue, the undeniably intelligent chessplaying program that vanquished Gary Kasparov, played chess in a manner that the engineers at IBM found unfathomable.[5] A parallel point holds with respect to other famous AIs, such as Watson, the system that vanquished the best human *Jeopardy!* players on the planet. No one doubts that Watson, during the competition, was intelligent, and yet Watson's performance wasn't in the least mysterious, given how the engineers designed and built it.[6]

The upshot is that if $\mathcal{A}$ is to be sound, a different type of creativity is needed for the X in this argument.

3.2 Creative Formal Thought?

It turns out that Bringsjord (2015*c*) has published a purported *proof* that intelligence implies a certain form of creativity. Might this be just what the doctor ordered for the variable X in $\mathcal{A}$? Unfortunately, there is a two-part catch, and this is reflected in the title of the paper in question: "Theorem: *General Intelligence Entails Creativity*, assuming ...." The overall issue is what is assumed. The proof assumes, one, that the intelligence in question must include arithmetic (here rhymes with 'empathetic') intelligence, and two, that the level of this intelligence is *very* high. More specifically, the idealized agent that the analysis and argumentation centers around must have command not only over the axiom system of Peano Arithmetic (PA) itself, but must also have command over the meta-theory of PA. (For instance, the agent must know that the axioms of

[4] Now of course known far and wide as the 'Turing test.'

[5] Indeed, quite the contrary, since Joel Benjamin, the grandmaster who consulted to the IBM team, inserted his own knowledge of such specific topics as "king safety" into the system. For a discussion, see (Bringsjord 1998).

[6] The designs can be found in Ferrucci et al. 2010. For further analysis of Watson see e.g. (Govindarajulu, Licato & Bringsjord 2014).

PA, an infinite set, are all true on the standard interpretation of arithmetic — and this is just for starters.) Under these two assumptions regarding arithmetic intelligence, the representative agent is shown to have a form of logicist creativity (l-creativity). Unfortunately, setting 'X-creativity' to 'l-creativity' in $\mathcal{A}$ doesn't render the first inference valid, for the simple reason that Daimler engineers aren't interested (let alone compelled) to seek self-driving cars able to understand and prove abstruse aspects of mathematical logic. So back to the drawing board we go.

3.3 Creative Musical Thought?

Another option for X, at least formally speaking, is musical creativity (e.g. see Ellis et al. 2015). But as Bringsjord pointed out in person when presenting the kernel of the present paper in Vienna, mere steps from where the display of such creativity, in its perhaps highest form, happened in the past,[7] while Daimler takes commendable pains to ensure that the sound systems in its cars are impressive, they have no plans to take on the job of producing AI that also *generates* the music that is so wonderfully presented to passengers.[8]

3.4 Creative Literary Thought?

Any notion that the kind of creativity relevant to self-driving cars is LT-creativity, l-creativity, or musical creativity is, as we have now noted, implausible to the point of being almost silly. But we come now to yet another form of creativity that just might not be so crazy in the current context: *literary* creativity. It turns out that this is once again a form of creativity that Bringsjord has spent time investigating; in this case, in fact, a *lot* of time (e.g. see the monograph Bringsjord & Ferrucci 2000). Under charitable assumptions, Bringsjord's investigation implies that even today's well-equipped but human-driven cars are *already* literarily creative, at least in a "plot-centric" way. If we interpret the route of a car from point of origin to destination to constitute a series of events involving passengers and the car itself as characters, then by some relaxed concepts of what a story is it would immediately follow that top-flight navigation systems in today's human-piloted cars are quite capable of story generation. This would be true of Bringsjord's own German sedan, and indeed true of his drive to JFK airport to fly to Vienna to give the very talk that expressed the core of the reasoning presented in the present paper. The reason is that during this drive the navigation system generated a number of unorthodox routes to JFK, in order to avoid extreme congestion on the infamous-to-New-Yorkers Van Wyck Expressway. It would be easy enough to express these alternative routes in a format that has been used in AI to represent stories. For example, the routes could be easily represented in the **event calculus**, which is explained and in fact specifically used to represent stories in (Mueller 2014). While doing this sort of thing might

[7] In the talk in question, Bringsjord's reference was to Mozart's *Don Giovanni*, which Kierkegaard (1992) argued is the highest art produced by humankind to that point.

[8] We would be remiss if we didn't point out that the work of Cope in musical creativity might well be a candidate for X in $\mathcal{A}$. Essentially, Cope's view, set out e.g. in (Kierkegaard 1992), is that pretty much all problem-solving can be regarded to be creativity at work. Early in his book Cope gives an example of a logic puzzle that can be solved by garden-variety deduction, and says that such solving is an instance of creativity at work. While there should be no denying that intelligence implies problem-solving (i.e., more carefully, that if a is intelligent, then a has some basic problem-solving capability), the problem is that Cope's claim that simple problem-solving entails creativity is a very problematic one — and one that we reject. Evidence for our position includes that simple problem-solving power is routinely defined (and implemented) in AI without any mention of creativity. E.g., see the "gold-standard" AI textbook (Russell & Norvig 2009). Please note that MacGyveresque creativity (m-creativity) is *not* garden-variety problem-solving.

strike some readers as frivolous, there can be no denying that, in the longstanding tradition in AI that counts a mundane declarative representation of a series of events that involve agents to be a story,[9] we must accept as fact that ICAP cars could be literarily creative. In fact, given that cars today integrate navigation with knowledge of not only traffic flow, but points of interest along the way, it would not be hyperbole to say that cars increasingly have knowledge by which they could spin stories of considerable plot-centric complexity. However, the previous two sentences say: *could* be creative. They don't say that Daimler would *need* to make literarily creative ICAP self-driving cars. Therefore, we still haven't found an instantiation of X in $\mathcal{A}$ that produces a valid inference to C1.

3.5 But what about MacGyveresque creativity?

The kinds of creativity we've canvassed so far have been, in each case, *domain-specific*. We turn our attention now to a general form of creativity that seems to be applicable in nearly any domain that presents problems to agents whose goals require the solving of those problems. This form of creativity is a brand of "out-of-the-box" problem-solving, an extreme resourcefulness based in large part on an ability to create, on the spot, novel problem-solving moves in extremely challenging and novel situations, and to thereby conquer these situations — where the conquering, by definition, isn't accessible to shallow learning techniques currently associated with such things as 'machine learning' and 'deep learning.'[10] At least for those familiar with "classic" television in the United States, the paragon of this form of creativity is the heroic secret agent known as 'MacGyver,' who starred in a long-running show of the same name.[11] We thus refer to the kind of creativity in question as *MacGyveresque*, or, for short, *m-creativity*.[12] In the very first episode, MacGyver is confronted with the problem of having to move a large and heavy steel beam that blocks his way. His creative solution is to cut the end off of a fire hose, tie a knot in that hose, and run the hose underneath the beam. MacGyver then turns on the water, and the hydraulic pressure in the hose lifts the beam enough for him to pivot it clear. This kind of creativity is manifested frequently, in episode after episode. One of the hallmarks of m-creativity is that (i) the known and planned purposes of objects, for MacGyver, turn out to be irrelevant in the particular problems confronting him, but (ii) extemporaneously in those problems these objects are used in efficacious ways that

[9] E.g., see (Charniak & McDermott 1985).

[10] These shallow techniques all leverage statistical "learning" over large amounts of data. But many forms of human learning (indeed, the forms of learning that gave us rigorous engineering in the first place!) are based on understanding only a tiny number of symbols that semantically encode an *infinite* amount of data. A simple example is the Peano Axioms (for arithmetic). Another simple example is the long-known fact that all of classical mathematics as taught across the globe is represented by the tiny number of symbols it takes to express axiomatic set theory in but a single page. Elementary presentation of such "big-but-buried" data (Bringsjord & Bringsjord 2014) as seen in these two examples is provided in (Ebbinghaus, Flum & Thomas 1994). Perhaps the most serious flaw infecting the methodology of machine learning as a way to engineer self-driving ICAP cars is that obviously it would be acutely desirable for quick-and-interactive learning to be possible for such cars — but by definition that is impossible in the paradigm of ML. If Bringsjord's automatic garage door seizes up before rising all the way, and he MacGyveresquely commands (in natural language) his car to deflate its tires to allow for passing just underneath and in, the car should instantly learn a new technique for saving the day in all sorts of tough, analogous spots, even if there isn't a shred of data about such m-creative maneuvers.

[11] The Wikipedia entry: https://en.wikipedia.org/wiki/MacGyver.

[12] Various places online carry lists of problems ingeniously solved by MacGyver. E.g., see

http://macgyver.wikia.com/wiki/List_of_problems_solved_by_MacGyver

the humans who designed and produced those objects didn't foresee.[13]

It seems to us that it can be shown, rather easily, that a truly intelligent artificial agent, operating in the ordinary environment that usually houses human beings and presents them with everyday, run-of-the-mill challenges, must be m-creative. We believe that this can be shown *formally*, that is, proved outright, but given the space required to set the formal context and assumptions for such an endeavor, we will rest content here with a simple example, and an argument associated with it, to make our point.

First, we inform the reader that we aim for the logically equivalent contrapositive: that if our arbitrary agent a, in the kind of environment we have imagined, presented with a representative type of problem P, is not m-creative, then a is not intelligent. For P, we choose a simple but classic challenge. In it, a subject is confronted with the challenge of completing a short and straightforward low-voltage circuit, in order to light a small bulb. A metal screwdriver with a plastic handle is provided, with instructions that it be used to first tighten down the terminals on either side of a lone switch in the circuit.[14] The problem is that the circuit isn't completed, and hence the lamp is unlit, because there is a gap in the wiring. No other props or tools are provided, or allowed. The puzzle is depicted in Figure 2.

Now, suppose that a small humanoid robot, billed by its creators as intelligent relative to environments like the one that envelopes the circuit problem here, appears on the scene, and is confronted with the problem. The situation is explained to the robot, just as it is explained to the human subjects in (Glucksberg 1968). (Both the humans and our robot know that current will flow from the power source to the lamp, and light it, if the wire makes an uninterrupted loop. This is of course also common knowledge, even in middle school in technologized countries.) The robot proceeds to screw down the terminals on either side of the switch. But after that, despite being told to light the lamp, it's quite paralyzed. That is, the robot doesn't use the screwdriver to complete the circuit and light the lamp, but instead stares for a while at the setup in front of it, and then announces: "I am sorry. I cannot figure out how to light the lamp." It seems clear that we would have to say that the robot isn't intelligent, despite claims to the contrary by its creators. Of course, an m-creative agent needn't have a perfect batting average: some problems will go unsolved, because the not-as-designed use of objects won't invariably be discovered by the agent in question. But to flesh out our argument, simply assume that our parable has a number of close cousins, and that in each and every one, the robot has no trouble using objects to do things for which they were explicitly designed, but invariably is stumped by problems testing for a simple level of m-creativity. Clearly, the robot is not intelligent.

The upshot is plain. In light of the fact that the conditional in question holds, it follows that if self-driving car c is intelligent, we should indeed expect m-creativity from c. Hence, we have made it to intermediary conclusion C1 in The Core Argument.

Of course, this is rather abstract; some readers will expect at least some examples of m-

[13] Elsewhere, one of us, joined by others, has written at some length about the nature of m-creativity, in connection with famous problems invented and presented to subjects in experiments by the great psychologist Jean Piaget (Bringsjord & Licato 2012). We leave this related line of analysis and AI aside here, in the interest of space conservation. For a sampling of the Piagetian problems in question, see (Inhelder & Piaget 1958). This is perhaps a good spot to mention that readers interested in m-creativity at a positively extreme, peerless level that exceeds the exploits of MacGyver will find what they are looking for in the problem-solving exploits of the inimitable "egghead" genius, Prof. Augustus Van Dusen, a.k.a. "The Thinking Machine." Perhaps the most amazing display of his m-creativity is in Van Dusen's escape from prison cell 13. See (Futrelle 2003).

[14] This simple problem, and other kindred ones, are presented in (Glucksberg 1968), in connection with a discussion of what we have dubbed m-creativity, but what Glucksberg considers to be creativity *simpliciter*.

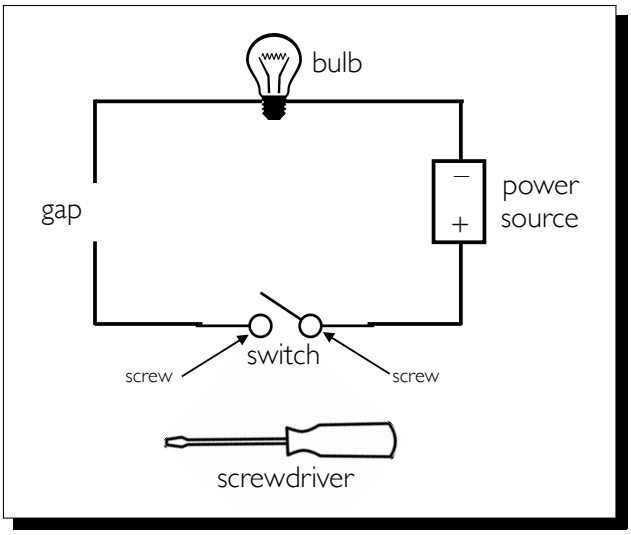

Figure 2: A Simple Circuit Problem That Invites M-Creativity

creativity on the roadways. It is easy to comply, by turning not to abstractions and dreamed-up-by-psychologists problems given to subjects in the laboratory, but rather to the real world. A convenient source of examples comes from inclement weather and emergencies; here's a weather-related one: Bringsjord probably shouldn't have been driving his SUV, but he was. The reason why he probably should have been off the roadways was that they were snow-and-ice-covered. He attempted to drive down a steep hill in Troy, NY, specifically down Linden Avenue, which snakes down alongside the Postenkill waterfall. At the beginning of the descent, the SUV slid uncontrollably to the right, into a snowbank just above the curb, giving the driver a nice jolt. It was painfully obvious that it would be impossible to drive down the hill along anything like the normal track. It was also impossible now to back up to the top of the hill. So what Bringsjord did was was follow this algorithm: About every fifteen feet or so, unsteadily but reliably steer the vehicle forward into the curb and snowbank to come to a complete stop; to get started again, move the steering wheel to the left; repeat. In this manner, the hill could be descended gradually, without building up too much speed. It took a long time, but it worked. The curb and the snowbank, combined with the right front tire taking moderate impact, functioned as a brake. Many, many such examples can be given, all realistic.[15] Some of realistic m-creativity involve violating the normal "rules" of driving. For example, it's often necessary, to avoid a crash, that a car be driven into a grass median, or onto a sidewalk, or across a double-yellow line into the oncoming lane, and so on. In cases where crash-avoidance like this crosses over to making use of causal chains unanticipated by human designers, with objects used in a manner that is different than standard operating procedure, we have m-creativity. Given what we say below about the coming multi-agent nature of self-driving, m-creativity along this line will only grow in importance.

[15]In demonstrations in our lab, our focus is on m-creativity for miniature self-driving ICAP cars that have at their disposal the ability to move objects by plowing and nudging them.

4 From MacGyveresque Creativity to Autonomy

But of course The Core Argument doesn't stop with mention of X-creativity (now instantiated to m-creativity): it proceeds to say that since creativity implies autonomy, the representative car c must be autonomous. *Does* m-creativity entail autonomy? Yes, and there are three general reasons.

4.1 Reason #1: M-creative Cars are *Partially* LT-Creative

The first reason is that despite the rejection, above, of LT-creativty for X-creativity in $\mathcal{A}$, the fact remains that any artifact capable of m-creativity must, by definition, at least to *some* degree, move in the direction of what the Lovelace Test demands: the artifact in question must exceed what the engineers of this artifact could have *specifically* anticipated. In other words, an m-creative self-driving car must be, at least to *some* degree, LT-creative. Why? An m-creative agent must be able to adapt to unprecedented problems on the spot, and devise, on that very same spot, unprecedented solutions to those problems. The key word here is of course 'unprecedented.' Driving on a super-highway is something that humans can often do "without thinking." Most human drivers, sooner or later, if they drive enough on super-highways (the so-called "interstates" in the United States, and such things as "B" or "M" roads in Europe), realize that they have been driving for quite a while without the slightest recollection of having done so. This is "zombie" driving. In zombie driving, the percepts to the driver never vary much through time. And, those percepts may well have been fully anticipated by the human engineers of self-driving cars. But zombie driving isn't m-creative in the least. A zombie driver isn't going to get out of a tough spot with ingenious resourcefulness applied on the fly; but such surgically applied resourcefulness is precisely what constitutes m-creativity.

Clearly, m-creativity in a computing machine is not anticipatable by the designers of this machine. We don't have to go all the way to the passing of the Lovelace Test (Bringsjord et al. 2001), but clearly the particular innovations can't be anticipated by the designers, and hence when an artificial agent comes up with innovative uses for objects, the designers must find these innovations surprising. The designers can, upon learning of these innovations, and then reflecting, grasp how the machine in question could have risen to the occasion, but they can't have known in advance about the specifics. In this sense, then, m-creative self-driving cars would exhibit a kind of minimal autonomy.

4.2 Reason #2: Satisfaction of Could-Have-Done-Otherwise Definitions

While it's undeniable that the term 'autonomous' is now routinely ascribed to various artifacts that are based on computing machines, it's also undeniable that such ascriptions are — as of the typing of the present sentence in early 2016 — issued in the absence of a formal definition of what autonomy *is*. What might a formal definition of autonomy look like? Presumably such an account would be developed along one or both of two trajectories. On the one hand, autonomy might be cashed out as a formalization of the kernel that s is autonomous at a given time t just in case, at that time, s can (perhaps at some immediate-successor time t') perform some action a_1 or some incompatible action a_2. In keeping with this intuitive picture, if the past tense is used, and accordingly the definiendum is 's autonomously performed action a at time t,' then the idea would be that, at t, or perhaps at an immediate preceding time t', s could have, unto itself, performed alternative action a'. (There may of course be many alternatives.) Of course, all of this is quite informal.

This picture is an intuitive springboard for deploying formal logic to work out matters in sufficient detail to allow meaningful and substantive conjectures to be devised, and either confirmed (proof) or refuted (disproof). But for *present* purposes, the point is only that the springboard commits us to a basic notion of autonomy: namely, that the agent in question, not some other agent, had the freedom in and of itself to have done otherwise. But an m-creative agent, we have already noted, comes up with an innovation that solves a problem on the spot. Yet this innovation is by no means a foregone conclusion. If it was, then there would be nothing innovative, nothing that the human designers and engineers didn't themselves anticipate and plan to have happen. We thus have a second reason for holding that m-creativity implies autonomy.

4.3 Reason #3: A Retreat to Mere Measurement of Autonomy Secures the Implication

The third reason is revealed once we forego trying to exploit the nature of autonomy in order to show that m-creativity entails autonomy *itself*, and instead reflect upon the relationship between m-creativity and and the *measurement* of autonomy, done in a way that bypasses having to speculate about its "interior" nature. This route dodges the issue of developing a formal definition, by substituting some formal quantity for a judgment as to what degree some robot is autonomous. Here's one possibility for this route — a possibility that draws upon the logic that constitutes the definition and further development of **Kolmogorov complexity**.[16] Let's suppose that the behavior of our self-driving car c from some time t_1 to any subsequent time t_k, $k \in \mathbb{N}$, can be identified with some string $\sigma \in \{0,1\}^*$. We remind (or inform) the reader that the Kolmogorov complexity of a string $\tau \in \{0,1\}^*$, $\mathcal{C}(\tau)$, is the length of the smallest Turing-level program π that outputs τ; that is,

$$\min\{|\pi| : \pi \longrightarrow \tau\}.$$

Now simply define the **degree of autonomy** of a given c at some point t in its lifespan to be the Kolmogorov complexity of the string, up to t, that is the representation of its behavior to that point.[17] Of course, as is well-known, any finite alphabet Σ can be used here, not just the binary one here employed. As long as the behavior of c at every given point in its existence can be captured by a finite string over Σ, we have developed here a measure of the degree of autonomy of that car. We can easily see that the behavior of an m-creative self-driving car, through time, must have a higher and higher degree of autonomy. A zombie self-driving car on a super-highway would almost invariably produce a binary string through time that is highly regular and redundant; hence such a car would have a relatively small degree of autonomy. But things are of course quite different in the case of an m-creative self-driving car.

5 Are powerful self-driving cars desired?

We now come to the next step in $\mathcal{A}$, one triggered by the question: Does Daimler desire *powerful* self-driving cars? A likely response is: "Well, who knows? After all, we don't know what you mean by 'powerful,' so we can't give a rational, informative answer. We *can* tell you that we certainly

[16] We provide here no detailed coverage of Kolmogorov complexity. For such coverage, Bringsjord recommends that readers consult (Li & Vitányi 2008).

[17] The string should also represent the varying state of the environment. Without loss of generality, we leave this aside for streamlining.

want an *effective*, *safe*, and *reliable* self-driving car." Yet these three avowed desiderata presuppose the very concepts that we now use to define power. These concepts include utility, disutility, and the basic, universally affirmed structure of what an artificial intelligent agent is, according to orthodoxy in the field of AI. Such orthodoxy is set out definitively by Russell & Norvig (2009), who explain that AI is the field devoted to specifying and implementing *intelligent agents*, where such agents are functions from percepts (that which they perceive), to actions performed in the environment in which the operate. Where AIA_i is an arbitrary agent of this type,[18] PER the set of percepts π_j, and ACT the set of actions α_k, a given agent is thus a mapping

$$\text{AIA}_i : \text{PER} \longrightarrow \text{ACT}.$$

Given the abstract level of analysis at which the present paper is pitched, there is no need to specify the domain and range of such a mapping. All readers who drive will have an immediate, intuitive grasp of many of the real-world members of these sets. For instance, on roads on which construction dump-trucks travel, sometimes debris flies out and down onto the road surface from the containers on such trucks; and if that happens in front of you while you're driving, it's good to able able to perceive the falling/fallen debris, and avoid it. The same thing goes for trucks that transport, particularly in open-air style, building materials. If a concrete block slides off of such a truck in front of you on a super-highway, you will generally want to spot it, and dodge it. Such examples could of course be multiplied *ad indefinitum*. It's also easy enough to imagine percepts and actions of a more mundane sort: When Bringsjord drives to the airport in a snowstorm, he needs to perceive the degree to which the roads have been plowed and salted, the maximum reduced rate of speed he will likely be able to achieve, and so on. These percepts dictate all sorts of actions.

Now let us add a measure $u(\cdot)$ of the utility (or disutility) accruing from the performance of some action performed by an agent AIA_i, where the range of this function is the integers; hence

$$u : \text{ACT} \longrightarrow \mathbb{Z}. \tag{1}$$

Given these rudiments, we can articulate a simple account of power, for we can say that a powerful agent is one such that, in the course of its life, will often be in situations that present to it percepts π_k such that

$$u(\text{AIA}_i(\pi_j)) > \tau^+ \in \mathbb{Z}^+ \text{ or } < \tau^- \in \mathbb{Z}^- \tag{2}$$

Of course, not only is this account rather abstract, but it's also confessedly indeterminate, for the reason that we don't know how large should be the potential weal τ^+, nor how small should be the potential woe τ^-, in order to ensure satisfaction of the definiens. Moreover, this lacuna is not unimportant, for it clearly hovers around the key question of how much power Daimler engineers wish to bestow upon their self-driving cars. Yet, clearly if the constants τ^+ and τ^- are, respectively, quite large and quite small, the quartet R1–R4 will indeed need to be provided. Truly powerful agents can bring on great goodness, and wreak great destruction. We will refrain from providing a justification for the claim that large amounts of power mandate R1–R4, and will make only a few quick remarks about R1–R3, before moving on to the planned treatment of R4, and then R5.

[18] Disclaimer: Formally inclined-and-practiced readers will not find in the equations below full rigor. We don't even take a stand here on how large is the set of available agents (which would of course be assumed to minimally match the countably infinite set of Turing machines). The present essay is a prolegomenon to "full engagement" for the computational logicians.

6 The Quartet R1–R4; OS-Rooted Ethical Control (R4)

Now, given that Daimler must build ICAP self-driving cars, that is, cars which are not only intelligent, but m-creative, autonomous, and powerful, simple prudence dictates that they must take great care in doing so. This is easy to see, if we consider not the realm of ground transportation, but the *less* dangerous realm of cooking within a home. Suppose, specifically, that you have a robot-manufacturing company, one that makes humanoid robot cooks that are not only intelligent, but also m-creative, autonomous, and powerful. We don't have to delve into the details of these ICAP cooks, because the reader can quickly imagine, at least in broad strokes, what the combination of I and C and A and P means in a standard kitchen. An ICAP robot chef would be able to figure out how to put together a wonderful dinner even out of ingredients that haven't been purchased in connection with a recipe; would be able command the kitchen in a manner independent of control exercised by other agents, including humans; and would have the power to start a fire that could cause massive disutility. In this hypothetical scenario, your robot-building company would have four reasons to hire the logicians: to formally verify the software constituting the "mind" of the robo-chef, to provide transparent software to inspect, debut, and extend; to enable the robo-chef to justify and explain, in cogent natural language, what it has done, is doing, and will be doing, and why; and to guarantee that the robot cook will not be doing anything that is unethical, and will do what is obligatory, courteous, and — perhaps sometimes — heroic.[19]

What the story about the robot chef shows is the need, on the part of Daimler, to hire the computational logicians flows directly from the need to be careful and thorough about ICAP cars, for four reasons/technologies: R1–R4. As we have said, our emphasis herein is on R4, but before discussing this reason for turning to logic for help, we now briefly run through the first three reasons, and explain briefly how they are inseparably linked to logic.

R1 This first reason for hiring the computational logicians refers to the formal verification of the computer program(s) that control(s) self-driving car c. Anything short of such verification means that belief that programs are correct (and belief that therefore behavior caused by the execution of these programs will be correct) stems from mere empirical testing — but of course no matter how many empirical tests our self-driving car c passes, it will still be entirely possible that under unforeseen conditions, c will behave inappropriately, perhaps causing great harm by virtue of c's power. We refrain from providing even the core of an account of formal program verification here.[20] We simply point out here that it is wholly uncontroversial that the one and only way to formally verify a self-driving car, on the assumption that it's significant behavior through time conforms to the shape of 1 and 2, which means that this behavior is the product of the execution of computer programs, is to rely upon formal logic. The previous sentence is simply a corollary entailed by the brute fact that formal verification of software, in general, is a logicist affair.[21]

[19] Such as e.g. extinguishing a fire, or retrieving a human from a fire that would otherwise have seen the human perish — even if it means that the it (= the robot) will itself expire.

[20] For an efficient book-length introduction at the undergraduate level, the reader can consult (Almeida, Frade, Pinto & de Sousa 2011); a shorter, elegant introduction is provided in (Klein 2009). For Bringsjord's recommended approach to formal verification, based aggressively on the Curry-Howard Isomorphism, readers can consult (Bringsjord 2015*b*), the philosophical background and precursor to which is (Arkoudas & Bringsjord 2007).

[21] At the moment, for formally verified operating-system kernels, the clear frontrunner is seL4 (https://sel4.systems). It runs on both x86 and ARM platforms, and can even run the Linux user-space, currently only within a virtual machine. Its also open-source, including the proofs. We see no reason why our ethical-control logic (discussed below) could not be woven into seL4 to form what we call the *ethical substrate*. For a remarkable success story in formal

R2 Statistical techniques for engineering intelligent agents have the great disadvantage of producing systems that compute the overall functions of equations 1 and 2 in black-box fashion. Such techniques include those that fall under today's vaunted "machine learning," or just 'ML.' Logicist AI, in contrast, yields intelligent agents that are transparent (Bringsjord 2008*b*). Certainly *ceteris paribus* we would want to be able to see exactly why a self-driving car c did or is doing something, especially if it had performed or was performing destructive or immoral actions, but without a logicist methodology being employed, this will be impossible.[22]

R3 The third reason to seek out the help of computational logicians is to obtain technology that will allow self-driving ICAP cars to cogently justify what they have done, are doing, and plan to do, where these justifications, supplied to their human overseers and customers, include the context of other objects and information in the environment. Cogent justification can be provided in a manner well shy of formal or informal proof. In fact, we have in mind that such justification *cannot* be delivered in the form of a formal proof — the reason being that a justification in this form would fail to be understandable to the vast majority of humans concerned to receive a justification in the first place. What is needed here from the machine is a clear, inferentially valid *argument* expressed in a natural language like German or English.[23] This would be an exceedingly nice thing to receive, for instance, from the vehicle recently involved in Google's first (self-driving) car accident.[24] Absent this capability, the Department of Motor Vehicles is utterly at the mercy of human employees at Google. Moreover, it's far from clear that even Google understood immediately after the accident what happened, and why. Even a responsible novice human driver, immediately after such a crash, could have promptly delivered, on the spot, a lucid explanation. Obviously, AI must provide this kind of capability, at a minimum.

But such capability entails that the underlying technology be logicist in nature. For minimally, a perspicuous argument must be composed of explicit, linked declarative statements, where the links are sanctioned by schemas or principles of valid inference, and the argument is surveyable and inspectable by the relevant humans.[25] Relevant here is the empirical fact that while natural-

verification at the OS-level, and one more in line with the formal logics and proof theories our lab is inclined to use, see (Arkoudas et al. 2004).

[22]ML devotees may retort that some ML techniques produce declarative formulae, which are transparent. E.g., so-called **inductive logic programming** produces declarative formulae (for a summary, see Alpaydin 2014). But such formulae are painfully inexpressive conditionals, so much so that they can't express even basic facts about the states of mind of the humans ICAP cars are intended to serve. In this regard, see (Bringsjord 2008*a*).

[23]This need immediately brings forth a related need on the machine-ethics side to regulate what self-driving ICAP cars say. E.g., Clark (2008) has demonstrated that using the underlying logic of mental-models theory (Johnson-Laird 1983), a machine can deceive humans by generating mendacious arguments.

[24]See "Google's Self-Driving Car Caused It's First Crash," by Alex Davies, in *Wired*, 2/2916. The story is currently available online at http://www.wired.com/2016/02/googles-self-driving-car-may-caused-first-crash. Davies writes:

> In an accident report filed with the California DMV on February 23 (and made public today), Google wrote that its autonomous car, a Lexus SUV, was driving itself down El Camino Real in Mountain View. It moved to the far right lane to make a right turn onto Castro Street, but stopped when it detected sand bags sitting around a storm drain and blocking its path. It was the move to get around the sand bags that caused the trouble, according to the report ...

While we don't press the issue, it turns out that the accident could have been avoided by the self-driving car in question had it been using computational logics (e.g. Bringsjord & Govindarajulu 2013) able to represent and reason about the epistemic attitudes of nearby human drivers. Such logics in our lab, with help from Mei Si and her body of work, can be augmented to directly reflect the modeling of emotion, as e.g. pursued in (Si et al. 2010).

[25]Note that *informal logic* revolves around arguments, not proofs. An excellent overview of informal logic is provided

language understanding (NLU) systems can be and often are (unwisely, in our opinion) rationally pursued on the basis of thoroughgoingly non-logicist methodology, this option is closed off when the challenge is natural-language *generation* (NLG) of substantive and semantic argumentation and proof. Evidence is provided by turning to any recent, elementary presentation of NLG techniques and formalisms (e.g. see Reiter & Dale 2000).[26]

R4 Now we come to our focus: the fourth member of the quintet of technologies that can be provided only by the computational logicians: OS-rooted ethical control. It's easy to convey the gist of what this technology provides: It ensures that ICAP self-driving cars operate ethically (a concept soon to be unpacked), as a function not merely of having had installed a high-level software module dedicated to this purpose, but because ethical control has been directly connected to the operating-system level of such cars; that is, because ethical control is *OS-rooted*.

The distinction between merely installing a module for ethical control as an engineering afterthought, versus first-rate engineering that implements such control at the operating-system level (so that modules allowing impermissible actions can trigger detectable contradictions with policies at the OS level), has been discussed in some detail elsewhere: (Govindarajulu & Bringsjord 2015). There, the authors write about two very different possible futures in the ethical control of artificial intelligent agents, one dark and one bright; these futures are depicted graphically in Figure 3. The basic idea is quite straightforward, and while the original context for explaining and establishing the importance of rooting the ethical control of ICAP members of AIA was a medical one, it's easy enough to transfer the basic idea to the domain of driving, with help from simple parables that parallel the rather lengthy one given in (Govindarajulu & Bringsjord 2015): Imagine that a self-driving ICAP car c' has had a "red" high-level module installed that precludes a combination of excessive speed and reckless lane-changing, and that in order to make it possible for c' to be suitably used by a particular law-enforcement agency in high-speed chases of criminals attempting to escape, this module has been (imprudently and perhaps illegally) stripped out by some in the IT division of that agency. (Notice the red module shown in Figure 3.) At this point, if c' discovers that some state-of-affairs $s^\star$ of very high utility can be secured by following an elaborate, m-creative plan that includes traveling at super-high speeds with extremely risky lane changing, given that c' is autonomous, powerful, and that the red module has been ripped out, c' proceeds to execute its plan to reach $s^\star$. This could be blocked if the policies prohibiting the combination of speed and risky lane-changing had been engineered at the OS level, in such a way that any module above this level causing an inconsistency with the OS-level policies cannot be executed.

But now we come to the obvious question: What is meant by 'ethical control,' regardless of the level that this concept is implemented at? There is already a sizable technical literature on "robot ethics," and a full answer to this question would require an extensive and systematic review of this literature in order to extract an answer. Doing this is clearly impracticable in the space remaining, and would in fact be a tall order even in a paper dedicated solely to the purpose of answering this query.[27] Bringsjord's most-recent work in robot ethics has been devoted to building a new

in "Informal Logic" http://plato.stanford.edu/entries/logic-informal in the Stanford Encyclopedia of Philosophy. This article makes clear that informal logic concentrates on the nature and uses of cogent arguments.

[26]Of course, there are some impressively engineered machine-learning systems that do such things are take in images and generate natural-language captions (e.g. Vinyals et al. 2015). Even run-of-the-mill, sustained, abstract argumentation and proof, at least at present, would require a rather more logicist framework; e.g., Grammatical Framework (Ranta 2011).

[27]Should the reader be both interested and willing to study book-length investigations, here are four highly rec-

ethical hierarchy $\mathscr{EH}$ [inspired by Leibniz's ethical theory and by 20^{th}-century work on ethics by Chisholm (1982)], for humans and robots (Bringsjord 2015*a*). This hierarchy includes not just what is commonly said to be *obligatory*, but what is *supererogatory*.[28] It does seem to us that the latter category applies to robots, and specifically to self-driving cars. That which is supererogatory is right to do, but not obligatory. For instance, it's right to run into a burning building to try to save someone, but this isn't obligatory. It might be said, plausibly, to be heroic. For another example, consider that it's right to say such things as "Have a nice day" to a salesperson after buying a cup of coffee, but such pleasantries aren't obligatory.

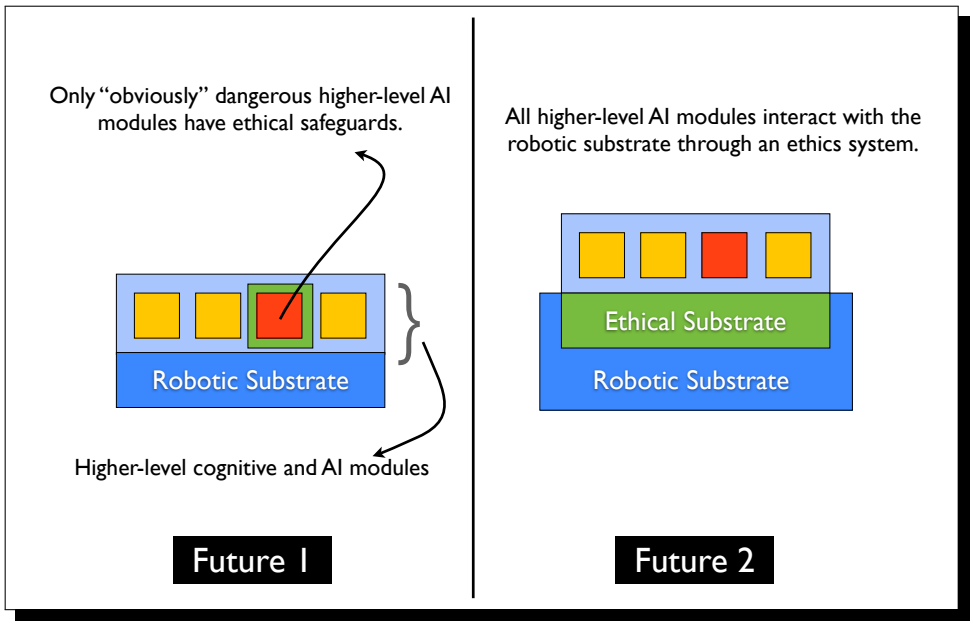

Figure 3: Two Possible Futures

Inspired in part by (Scheutz & Arnold forthcoming), we have been investigating, in "test-track" fashion in our laboratory, intelligent artificial agents of the ICAP variety[29] that, in driving scenarios, size things up and sometimes decide to behave in supererogatory fashion. Figure 4 shows a robot perched to intervene in supererogatory fashion in order to prevent Bert (of Sesame Street fame) from being killed by an onrushing car. In order to save Bert's life, the robot must sacrifice itself in the process. Would we wish to engineer life-size ICAP self-driving cars that are capable of supererogatory actions? If so, the computational logicians will need to be employed.

ommended volumes: (Pereira & Saptawijaya 2016, Trappl 2015, Bekey & Abney 2011, Arkin 2009).

[28]Here we streamline rather aggressively for purposes of accelerating exposition. The fuller truth is that standard deontic logic, and in fact even the vast majority of today's robot ethics that (in whole or in part) builds atop deontic logic, is based on the 19^{th} century triad that includes not just the *obligatory*, but the *forbidden*, and the *permissible* as well (where the forbidden is that which it's obligatory not to do, and the permissible is that which isn't forbidden). [(Chisholm (1982, p. 99) points out that Höfler had the deontic square of opposition in 1885.] $\mathscr{EH}$ adds not only the *supererogatory*, but the ***sub**erogatory* as well. Indeed the hierarchy partitions the supererogatory into that which is courteous-but-not-obligatory, and that which is heroic; and partitions the suberogatory into that which is done in — as we might say – bad faith, vs. that which is outright deviltry.

[29]The 'P' in 'ICAP' in our simulations is of course artificial, since (thankfully) the agents in our microworlds aren't able to produce large (à la equations 2 and 3) utility or disutility in the real world.

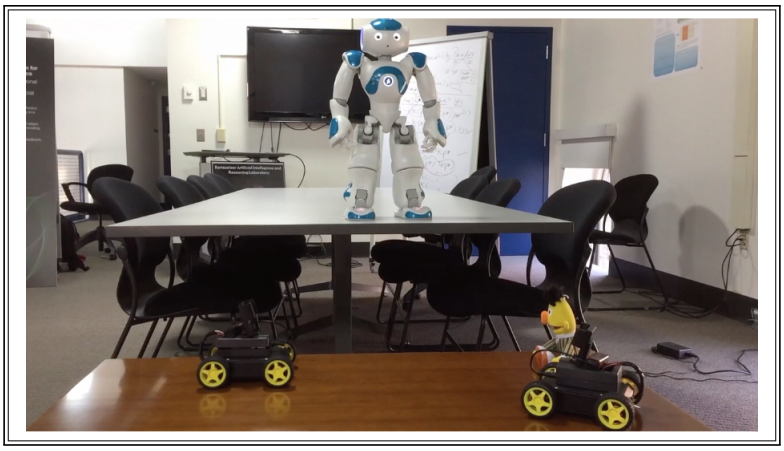

Figure 4: A Demonstration of *Supererogatory* Ethical Control *The "action" happens below the robot and the table it's on. The self-driving ICAP car to the far left of Bert will flatten him to the great beyond — unless the robot from above heroically dives down to block this onrushing car.*

At this point, a skeptical reader might object as follows: "But why should we accept the tacit proposition, clearly affirmed by the two of you, that ethics must be based in *logic*?"

We now answer this question, and use that answer as a springboard to moving from consideration of self-driving cars engineered to carry out supererogatory actions, to the more realistic engineering target — and the target that Daimler needs to put within its sights — of engineering ICAP cars engineered to meet their obligations.

So, why does a need for ethical control entail a need for control via logic? Well, the fact is, there's no other route to achieve precision in ethics than to use logic. This is reflected in the fact that for millennia, ethical theories and ethical principles have been expressed in declarative form (albeit usually informally), and the defense of these theories and principles have also been couched in logic (though again, usually in informal logic). Whether it's Socrates articulating and defending (before the first millennium) the view that knowledge, especially self-knowledge, is a moral virtue; whether it's Kant defending his position that one ought without exception to always act in a manner that can be universalized; whether it's Mill setting out and defending the first systematic version of the view that what ought to be done is that which maximizes happiness and minimizes pain; regardless, the commonality remains: that which these great ethicists did was to *reason* over declarative statements, in ways that are naturally modeled in formal logic, usually specifically in formal deductive logic.

We can certainly be more specific about the ethics-logic nexus. For example, the Millian ethical theory **act utilitarianism** consists in the biconditional statement that an agent α's action a at time t is obligatory if and only if a produces the greatest utility for the greatest number of agents from among all actions that α can perform at t. In opposition, a Kantian **deontological** ethical theory holds that, where $\mathcal{R}$ is a collection of obligations that uncompromisingly require actions wholly independent of the consequences of those actions, an agent α's action a at time t is obligatory if and only if performing a at t is logically entailed by one or more of the rules in $\mathcal{R}$. Even those with only a slight command of elementary formal logic will instantly see that if one were asked to take the time to render these theories more rigorous form amenable to implementation in a computing machine, one would inevitably set out these two theories by employing the machinery of elementary,

classical logic: for example, minimally, quantification (over agents, actions, times, etc.), an operator or predicate to represent 'obligatory,' and the five suitably deployed truth-functional connectives. Our point here is not at all how the specific formulae would be specified; rather, our point at the moment, given in response to the question above, is that any such specification would draw from the machinery of formal logic, of necessity.

In addition, particular ethical principles can be logically deduced from a combination of ethical theories expressed in rigorous, declarative fashion, in combination with a particular context at hand. For example, under the supposition that act utilitarianism as expressed in the present paragraph holds, if α at some time t faces but two incompatible options, in one case to cause a small, single-passenger car to move slightly to the right in its lane in order to save one human life (a_1), and in the other to cause a large truck to move slightly to the right in its lane in order to save 100 human lives (a_2), it follows (assuming that there are no other relevant consequences) that α ought to perform a_2. Formal logic can be used to render all such reasoning entirely explicit, implementable in an ICAP self-driving car, and checkable by any number of computer programs. Hence, while we do not, at least in the present paper, urge Daimler and its corporate cousins to engineer ICAP self-driving cars to perform supererogatory actions, we do urge these companies to hire computational logicians in order to engineer self-driving ICAP cars that meet their obligations. Figure 5 shows a snapshot of a scenario in which, in our lab's test environment, an ICAP self-driving car manages to save Bert's life by causing a slight deviation in the path of another car whose former route would have caused Bert's demise. In macroscopic, real-life form, this is the kind of behavior that Daimler must seek from it's self-driving cars, courtesy of what computational logicians can supply. Note that meeting an *ethical* obligation can sometimes entail violation of a standard driving rule or law. In the case of the obligation satisfied by the self-driving ICAP car shown in Figure 5, the action requires a slight crash into the car that would otherwise kill Bert. It is therefore important to understand that the mechanization of an ethical theory in OS-rooted fashion doesn't at all mean that the self-driving cars in question will be inflexible. Quite the contrary. As we noted earlier, m-creativity on the roadways can entail violation of standard operating procedure and standard rules of traffic law.

It's important to be clear at this juncture that logicist machine/robots ethics has reached a level of maturity that allows services to be provided to Daimler et al. that would in turn allow such companies to install ethical-control technology in their self-driving ICAP vehicles. This maturity was certainly not in place at the time of (Arkoudas, Bringsjord & Bello 2005), nor at the time of (Bringsjord, Arkoudas & Bello 2006), but over a solid decade of productive, well-funded toil has been expended since then, and it's high time to stop idly fretting about ethical problems that machines, vehicles included, will face, and start hiring the computational logicians to provide technology that allows machines to solve such problems. We need to move from philosophizing and fretting, to engineering and testing. As long as the underlying ethical theory is selected, the computational logicians can mechanize ethical control on the basis of that selection.

7 The "Tentacular" Demandingness Problem (R5)

We come now to the fifth and final reason (R5 in $\mathcal{A}$; see again Fig. 2) the computational logicians are needed by Daimler and their competitors. This reason is revealed by first taking note of the empirical fact that the ethical theory that clearly is (or — after assimilation of the present paper — would be) the first choice of the companies working on ICAP self-driving cars, and of the governments that regulate such work, is act utilitarianism, already defined above in at least

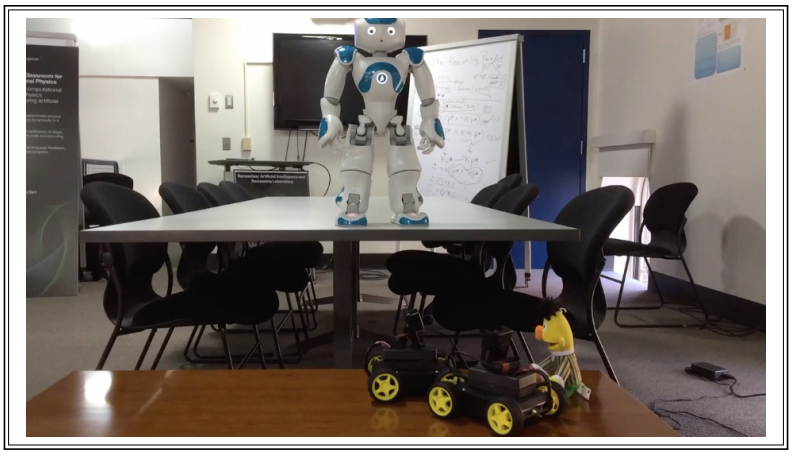

Figure 5: A Demonstration of *Obligation-only* Ethical Control *Once again the "action" happens below the robot and the table it's on; and once again, the self-driving ICAP car to the far left of Bert will flatten him to the great beyond — but the other self-driving ICAP car meets its obligation by deflecting the onrushing car, thereby keeping Bert and his acting career alive and well.*

rough-and-ready, but reasonably accurate, form. In other words, the ethical control of self-driving ICAP cars, at least in the early days of designing and engineering such control, will be based on act utilitarianism. But act utilitarianism appears to be very vulnerable, in light of the so-called *demandingness* objection. While even the generally educated public is aware of the fact that utilitarianism has long been thought by many to be untenable because it appears to classify such actions as torture and brutal slavery to be not only permissible but obligatory,[30] the demandingness objection to utilitarianism flies beneath the laic radar. While above we directed the reader to an excellent, recent, in-print presentation of the objection (i.e., Scheffler 1982), the first author's first exposure to this objection, in perfectly cogent form, came from one his professors in graduate school: Michael Zimmerman; and Bringsjord can still remember both the simplicity and the sting of the objection. Zimmerman pointed out that sometimes when one is reading a magazine, one comes across a full-page request for a donation, on which it's stated explicitly that a donation of some small amount of money will save a starving child in Africa by providing enough food for the child for an entire year. Usually a moving photograph of such a child, malnourished and in dire straits, is included on the page. Zimmerman asked: If you subscribe to act utilitarianism, how can you turn the page and not donate, and at the same time satisfy your moral obligations? After all, at the time that you either turn the page or donate, the latter action clearly produces the most happiness for the most people, among the actions available to you at that time. So, suppose you go ahead and donate: You pick up the phone, dial a toll-free number, and give your credit card to make a sizable contribution. You have at this point put yourself a bit behind schedule for getting done a bit of grocery shopping for your dinner later on on the same day, so now you need to move quickly to get back on track. However, the minute you walk out your door, you come upon an impoverished, disheveled beggar without legs, sitting on the sidewalk, holding a sign that pleads for any amount of money to help support him and his young family. Giving some cash in this case, among the other actions available to you (and certainly compared with simply walking passed the beggar toward the gleaming, well-stocked grocery store), would appear to be one you are obligated

[30] In cases where happiness is maximized by carrying out torture and or owning slaves; see (Feldman 1978).

to perform by act utilitarianism, since as a matter of fact you don't really *need* the groceries it will now take you a solid 45 minutes to obtain, and you could cobble together a simple dinner for the night out of canned goods you have in store already, and while you planned to stop and fetch a bottle of wine as well, the wine is superfluous, and pretty much produces only happiness for you. Thus, being a good utilitarian, you give the cash you have to the beggar. You then wonder why this fellow has not received assistance from the rescue mission just around the corner, and then you remember that that mission has of late had a sign posted on its front door asking for volunteers, no special skills required ... At this point the reader will understand the objection, which amounts in the end to the apparent fact that act utilitarianism demands actions from people in a manner that leaves them utterly paralyzed to accomplish their own private agendas. We generally assume that the pursuit of these agendas is perfectly permissible, ethically speaking; if this assumption is correct, utilitarianism is wrong.

What does this have to do with self-driving ICAP cars? It should be rather obvious. Because these cars are going to be capable of multi-agent reasoning of a highly sophisticated form, they are going to perceive all sorts of opportunities to produce gains in utility that far outweigh the small amounts of utility produced by merely shuttling you from point A to point B. They are also going to perceive countless opportunities to prevent losses of utility, where these opportunities are orthogonal to traveling the route that you, for "selfish" reasons, wish to cover. In fact this is to put the problem in very gentle terms. For the fact is that the multi-agent power of a multitude of self-driving ICAP cars will be quite staggering. One can grasp this by returning to the simple equations 1 and 2 given above. Given that we are now talking about the "hive" intelligence of millions of cars, spread out across roads and non-roads (there are currently about 250 million cars operating in the U.S.) the percepts represented by π_k for a single artificial intelligent agent are but a tiny part of the overall story. The real picture requires us to build up a much more complicated model. If we pretend in order to foster exposition that our hive of millions of self-driving ICAP cars $c_1, c_2, \ldots, c_p$ will begin the search for a coördinated plan at a timepoint of inception at which each perceives its corresponding π^i_j, the power of the hive from a utilitarian point of view would be such that [a multi-agent cousin of equation 2]::

$$\sum_{1}^{p} u(\text{AIA}_i(\pi^i_j)) \text{ where this sum } > \tau'^{+} \in \mathbb{Z}^{+} \text{ or } < \tau'^{-} \in \mathbb{Z}^{-} \tag{3}$$

And this equation, if implemented and a guide to action selection for interconnected self-driving ICAP cars, obviously presents a fierce form of the demandingness objection: what we call the *tentacular* demandingness objection. We appear to be looking at a future in which, if our machines are good utilitarians, they will be obligated to sweep aside our own petty-by-comparison agendas, in favor of goals and plans that are alien to us, and beyond our cognition to grasp in anything like real-time.

It seems to us that the problem becomes even more intense, and more intractable, when one realizes that given the "internet of things," the hive-mind in question isn't composed only of the artificial intelligent agents that are self-driving ICAP cars. We aren't even talking merely of a *vehicular* hive-mind. The hive-mind will include all manner of artifacts ubiquitous in the environment, from top to bottom: lights of all kinds, gates, security checkpoints, omnipresent sensors of all varieties, mobile devices, smart toys, service robots, smart weapons, self-moving ICAP vehicles in the air and water, softbots and conversational agents interacting with humans through innumerable interfaces of various types, and so it goes and grows, *ad infinitum*. An interesting

side-effect of coming to see what our future in this regard will be like is the realization that the somewhat silly ethical dilemmas like trolley problems for self-driving ICAP cars will eventually, in the multi-agent AI future we have foreseen, be so unlikely as to be a waste of time to worry about now, *contra* the concern for instance of Lin (2015) and Goodall (2014). After all, even today, for human drivers whose percepts and range of actions are severely limited compared to the hive-mind machines in our future, trolley-problem dilemmas are few and far between, to put it mildly. The gradual coming together of the hive-mind will see to it that such dilemmas are almost always prevented from forming in the first place. It's no accident that the familiar trolley problems are posed in almost ridiculously low-tech environments. Today we worry a lot about terrorists and try to track and thwart them at a human timescale, by primitive means; tomorrow, hive-mind artificial intelligence will have the power to thwart low-tech terrorists at their first observable movements toward heinous disutility — but ethical control of this kind of extreme power will be non-trivial, and a logic-based challenge that we need now to plan to successfully meet later.[31]

To be clear, we are not saying that the cases entertained in the likes of (Lin 2015, Goodall 2014) will never occur. What we are saying in the present paper is that whereas these cases may happen, and while the thinkers writing about them are providing a commendable service in doing so, these cases will be *exceedingly* rare — the demandingness problem in contrast isn't unlikely at all: in fact it's *guaranteed* to face Daimler et al. on a *continuous* basis. In addition, the philosopher's dilemma cases for self-driving cars are not ones that involve massive amounts of utility and disutility hanging in the balance.[32]

But what is the solution to the tentacular demandingness problem that we have described? Some readers may think that nothing could be easier to surmount, because the humans are after all in charge, and we can simply engineer the AIs, whether of the vehicular variety or not, in such a way that they don't seek to leverage their percepts and their power to maximize utility and minimize disutility. Unfortunately, since, as we have pointed out, the theory of utilitarianism is the dominant guiding theory for engineering AI in the service of humanity, it would be strangely *ad hoc* to restrict the power of self-driving ICAP cars to make the world a better place.

Providing a solution here to the tentacular demandingness problem is beyond the scope of our present objectives. We rest content with the observation that the problem can't be solved without the assistance of the logicians, who will need to be called upon to apply their formal techniques and tools to a tricky philosophical problem, something they have been doing for many, many centuries.

8 Conclusion; Next Steps

We conclude that those organizations intent on building intelligent self-driving cars, in light of the fact that these cars, for reasons we have given, will be ICAP ones, must hire the computational logicians for (at least) the five determinate reasons we have given. We are well aware of the fact that the budgets, engineering approaches, and business models of companies busy striving to bring self-driving cars to market are in large measure threatened by what we have said. For example,

[31] Many readers will no doubt at this point hear echoes of The Singularity, the imagined point in the future when computing machines suddenly become more intelligent than humans, and then in turn build new machines that are even smarter, with the cycle continuing until by comparison with the machines we are the intellectual equivalent of ants. While The Singularity is chimerical (Bringsjord, Bringsjord & Bello 2013), the hive intelligence we portray herein is thoroughly realistic, and in line with what Bringsjord & Bringsjord (2016) have dubbed "The Mini-Maxularity," the future timepoint when today's technology simply matures in harsh but predictable ways.

[32] In contrast, consider the *real-life* nuclear cases chronicled in (Schlosser 2013).

in our experience, companies are often intent on using clunky, unverified legacy software. In the case of self-driving ICAP cars, what we have revealed will require an engineering cleanliness that can only be obtained if OS-level code is rewritten from scratch in accordance with the Curry-Howard Isomorphism to reach pristine rock-solidness, and then connected to logicist code at the module level for ethical control on an ongoing basis. Along the same disruptive lines, for reasons given above, self-driving ICAP cars cannot be wisely or even prudently engineered on the basis of machine-learning techniques — and yet such techniques are the dominant ones in play today in AI. This must change, fast.

Obviously much logicist research remains to be carried out in the self-driving-car space — and indeed, because the topics traversed above are not in the least restricted only to cars, we can say that much logicist work remains to be performed in the *self-moving-vehicle* space.[33] Since we have only discussed very rapidly the first three things (R1–R3) that, by $\mathcal{A}$, computational logicians must be hired to provide, the trio needs to be taken up in sustained fashion in future work. For example, if we are right that ICAP self-moving vehicles must have the ability to (in some cases) cogently self-justify why they did what they did (esp. if what they did was objectionable to relevant humans), why they are doing what they're doing, and why they propose performing some future actions, it will be necessary that NLP engineering of a logicist sort be carried out by the relevant companies and engineers. By definition, cogent justifications are based on interleaved language and logic (whether or not in that interleaving the underlying logic is formal or informal, and whether or not deductive or inductive logic is used); yet, to our knowledge, none of the relevant formal theory, let alone the technology upon which it would be based, has been developed by the corporations busy building self-moving vehicles. This state-of-affairs needs to change *post haste*.

There is of course a *specific* problem that should be targeted in near-term subsequent work: In light of how problematic is the use of utilitarianism as an undergirding moral theory for ethical control of self-moving vehicles, what should be done? How should the engineering of ethical control for vast numbers of interconnected self-moving vehicles proceed, in light of the hive-mind version of the demandingness objection revealed above? Bringsjord's view, notwithstanding the fact that much public policy is implicitly based on utilitarian calculation, is that the undergirding moral theory for ethical control of self-moving vehicles should not be utilitarianism, nor for that matter *any* form of consequentialism, but should be based on Leibnizian ethics and the hierarchy $\mathscr{EH}$. A defense of this view, and of a better foundation for ethical control, will need to wait for another day.

[33] Some take 'vehicle' to connote land-based transport, but we take the work to be fully general, and hence it includes aircraft (e.g., ICAP UAVs). Note that in point of fact, the analysis and argument we give in the present paper applies to all ICAP robots, period.

References

Almeida, J., Frade, M., Pinto, J. & de Sousa, S. (2011), *Rigorous Software Development: An Introduction to Program Verification*, Spinger, New York, NY.

Alpaydin, E. (2014), *Introduction to Machine Learning*, MIT Press, Cambridge, MA.

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

Arkoudas, K. & Bringsjord, S. (2007), 'Computers, Justification, and Mathematical Knowledge', *Minds and Machines* **17**(2), 185–202.
URL: *http://kryten.mm.rpi.edu/ka_sb_proofs_offprint.pdf*

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*

Arkoudas, K., Zee, K., Kuncak, V. & Rinard, M. (2004), Verifying a File System Implementation, *in* 'Sixth International Conference on Formal Engineering Methods (ICFEM'04)', Vol. 3308 of *Lecture Notes in Computer Science (LNCS)*, Springer-Verlag, Seattle, USA, pp. 373–390.

Bekey, P. L. G. & Abney, K., eds (2011), *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT Press, Cambridge, MA.

Bringsjord, E. & Bringsjord, S. (2014), Education and ... Big Data Versus Big-But-Buried Data, *in* J. Lane, ed., 'Building a Smarter University', SUNY Press, Albany, NY, pp. 57–89. This url goes to a preprint only.
URL: *http://kryten.mm.rpi.edu/SB_EB_BBBD_0201141900NY.pdf*

Bringsjord, S. (1998), 'Chess is Too Easy', *Technology Review* **101**(2), 23–28.
URL: *http://kryten.mm.rpi.edu/SELPAP/CHESSEASY/chessistooeasy.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_ab-toc_031607.pdf*

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

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

Bringsjord, S. (2015*b*), 'A Vindication of Program Verification', *History and Philosophy of Logic* **36**(3), 262–277. This url goes to a preprint.
URL: *http://kryten.mm.rpi.edu/SB_progver_selfref_driver_final2_060215.pdf*

Bringsjord, S. (2015*c*), Theorem: *General Intelligence Entails Creativity*, assuming ..., *in* T. Besold, M. Schorlemmer & A. Smaill, eds, 'Computational Creativity Research: Towards Creative Machines', Atlantis/Springer, Paris, France, pp. 51–64. This is Volume 7 in *Atlantis Thinking Machines*, edited by Kühnbergwer, Kai-Uwe of the University of Osnabrück, Germany.
URL: *http://kryten.mm.rpi.edu/SB_gi_implies_creativity_061014.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. & Bringsjord, A. (2016), The Singularity Business: Toward a Realistic, Fine-grained Economics for an AI-Infused World, *in* T. Powers, ed., 'TBD', Springer, New York, NY. This volume is part of Springer's Philosophical Studies series.
URL: *http://kryten.mm.rpi.edu/SBringsjord_ABringsjord_SingularityBiz_0915151500.pdf*

Bringsjord, S., Bringsjord, A. & Bello, P. (2013), Belief in the Singularity is Fideistic, *in* A. Eden, J. Moor, J. Søraker & E. Steinhart, eds, 'The Singularity Hypothesis', Springer, New York, NY, pp. 395–408.

Bringsjord, S. & Ferrucci, D. (2000), *Artificial Intelligence and Literary Creativity: Inside the Mind of Brutus, a Storytelling Machine*, Lawrence Erlbaum, Mahwah, NJ.

Bringsjord, S., Ferrucci, D. & Bello, P. (2001), 'Creativity, the Turing Test, and the (Better) Lovelace Test', *Minds and Machines* **11**, 3–27.
URL: *http://kryten.mm.rpi.edu/lovelace.pdf*

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. & Licato, J. (2012), Psychometric Artificial General Intelligence: The Piaget-MacGuyver Room, *in* P. Wang & B. Goertzel, eds, 'Foundations of Artificial General Intelligence', Atlantis Press, Amsterdam, The Netherlands, pp. 25–47. This url is to a preprint only.
URL: *http://kryten.mm.rpi.edu/Bringsjord_Licato_PAGI_071512.pdf*

Charniak, E. & McDermott, D. (1985), *Introduction to Artificial Intelligence*, Addison-Wesley, Reading, MA.

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.

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

Ebbinghaus, H. D., Flum, J. & Thomas, W. (1994), *Mathematical Logic (second edition)*, Springer-Verlag, New York, NY.

Ellis, S., Haig, A., Govindarajulu, N., Bringsjord, S., Valerio, J., Braasch, J. & Oliveros, P. (2015), Handle: Engineering Artificial Musical Creativity at the 'Trickery' Level, *in* T. Besold, M. Schorlemmer & A. Smaill, eds, 'Computational Creativity Research: Towards Creative Machines', Atlantis/Springer, Paris, France, pp. 285–308. This is Volume 7 in *Atlantis Thinking Machines*, edited by Kühnbergwer, Kai-Uwe of the University of Osnabrück, Germany.
URL: *http://kryten.mm.rpi.edu/SB_gi_implies_creativity_061014.pdf*

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

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, W., Nyberg, E., Prager, J., Schlaefer, N. & Welty, C. (2010), 'Building Watson: An Overview of the DeepQA Project', *AI Magazine* pp. 59–79.
URL: *http://www.stanford.edu/class/cs124/AIMagzine-DeepQA.pdf*

Futrelle, J. (2003), *The Thinking Machine: The Enigmatic Problems of Professor Augustus S. F. X. Van Dusen, Ph.D., LL.D., F.R.S., M.D., M.D.S.*, The Modern Library, New York, NY. The book is edited by Harlan Ellison, who also provides an introduction. The year given here is the year of compilation and release from the publisher. E.g., Futrelle published "The Problem of Cell 13" in 1905.

Glucksberg, S. (1968), ''Turning On' New Ideas', *Princeton Alumni Weekly* **69**, 12–13. Specifically, November 19.

Goodall, N. (2014), 'Ethical Decision Making During Automated Vehicle Crashes', *Transportation Research Record: Journal of the Transportation Research Board* **2424**, 58–65.

Govindarajulu, N., Licato, J. & Bringsjord, S. (2014), Toward a Formalization of QA Problem Classes, *in* B. Goertzel, L. Orseau & J. Snaider, eds, 'Artificial General Intelligence; LNAI 8598', Springer, Switzerland, pp. 228–233.
URL: *http://kryten.mm.rpi.edu/NSG_SB_JL_QA_formalization_060214.pdf*

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

Inhelder, B. & Piaget, J. (1958), *The Growth of Logical Thinking from Childhood to Adolescence*, Basic Books, New York, NY.

Johnson-Laird, P. N. (1983), *Mental Models*, Harvard University Press, Cambridge, MA.

Kierkegaard, S. (1992), *Either/Or: A Fragment of Life*, Penguin, New York, NY. *Either/Or* was originally published in 1843.

Klein, G. (2009), 'Operating System Verification—An Overview', *Sādhanā* **34, Part 1**, 27–69.

Knobe, J., Buckwalter, W., Nichols, S., Robbins, P., Sarkissian, H. & Sommers, T. (2012), 'Experimental Philosophy', *Annual Review of Psychology* **63**, 81–99.

Li, M. & Vitányi, P. (2008), *An Introduction to Kolmogorov Complexity and Its Applications*, Springer, New York, NY. Third edition.

Lin, P. (2015), Why Ethics Matters for Autonomous Cars, *in* M. Maurer, C. Gerdes, B. Lenz & H. Winner, eds, 'Autonomes Fahren: Technische, Rechtiche und Gesellschaftiche Aspekte', Springer, Berlin, Germany, pp. 69–85.

Malle, B., Scheutz, M., Arnold, M., Voiklis, J. & Cusimano (2015), Sacrifice One For the Good of Many? People Apply Different Moral Norms to Human and Robot Agents, *in* 'Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI 2015)', ACM, New York, NY, pp. 117–124.
URL: *http://dx.doi.org/10.1145/2696454.2696458*

Mikhail, J. (2011), *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*, Cambridge University Press, Cambridge, UK. Kindle edition.

Mueller, E. (2014), *Commonsense Reasoning: An Event Calculus Based Approach*, Morgan Kaufmann, San Francisco, CA.

Pereira, L. & Saptawijaya, A. (2016), *Programming Machine Ethics*, Springer, Basel, Switzerland. This book is in Springer's SAPERE series, Vol. 26.

Ranta, A. (2011), *Grammatical Framework: Programming with Multilingual Grammars*, CSLI, Stanford, CA. ISBN-10: 1-57586-626-9 (Paper), 1-57586-627-7 (Cloth).

Reiter, E. & Dale, R. (2000), *Building Natural Language Generation Systems*, Cambridge University Press, Cambridge, UK. This book in in CUP's *Studies in Natural Language Processing* series.

Russell, S. & Norvig, P. (2009), *Artificial Intelligence: A Modern Approach*, Prentice Hall, Upper Saddle River, NJ. Third edition.

Scheffler, S. (1982), *The Rejection of Consequentialism*, Clarendon Press, Oxford, UK. This is a revised edition of the 1982 version, also from Clarendon.

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

Schlosser, E. (2013), *Command and Control: Nuclear Weapons, the Damascus Accident, and the Illusion of Safety*, Penguin, New York, NY.

Si, M., Marsella, S. & Pynadath, D. (2010), 'Modeling Appraisal in Theory of Mind Reasoning', *Journal of Agents and Multi-Agent Systems* **20**, 14–31.

Trappl, R. (2015), *A Construction Manual for Robots' Ethical Systems: Requirements, Methods, Implementations*, Springer, Basel, Switzerland.

Turing, A. (1950), 'Computing Machinery and Intelligence', *Mind* **LIX (59)**(236), 433–460.

Vinyals, O., Toshev, A., Bengio, S. & Erhan, D. (2015), 'Show and Tell: A Neural Image Caption Generator'. **URL:** *http://arxiv.org/pdf/1411.4555.pdf*