Can "Provably Beneficial AI" Save Us?

Selmer Bringsjord
Rensselaer AI & Reasoning Lab
RPI
Troy, USA
selmer.bringsjord@gmail.com

Naveen Sundar Govindarajulu
Rensselaer AI & Reasoning Lab
RPI
Troy, USA
naveen.sundar.g@gmail.com

John Licato
Advancing Machine & Human Reasoning Lab
University of South Florida
Tampa, USA
john.licato@gmail.com

***Abstract*—AI-polymath Stuart Russell, in the face of fear about superhuman AI arriving within 80 years and doing the human race in, commendably offers a recipe (based upon inductive reinforcement learning) for salvation quite different than our own (the sharing of which is beyond the current scope of the present paper). He does this in his recent book *Human Compatible*. Unfortunately, as we explain, Russell's recipe is afflicted by four fatal defects.**

***Index Terms*—machine ethics, robot ethics, inductive reinforcement learning**

I. Introduction: The Problem

AI-polymath[1] Stuart Russell, in the face of fear about superhuman AI arriving within 80 years and doing the human race in, offers a recipe for salvation quite different than our own (the sharing of which is beyond the current scope of the present short paper, but see e.g. [8]). He does this in his book *Human Compatible* [11]. Russell does not rely upon The Singularity (or any other such speculative thing) to justify his belief that superintelligent machines will arrive.[2] On the other hand, Russell is of the opinion that the arrival of superintelligent AI could very well be quite sudden. He writes:

> My timeline of, say, eighty years is considerably more conservative than that of the typical AI researcher. Recent surveys suggest that most active researchers expect human-level AI to arrive around the middle of this century. Our experience with nuclear physics suggests that it would be prudent to assume that progress could occur quite quickly and to prepare accordingly. If just one conceptual breakthrough were needed, analogous to Szilard's idea for a neutron-induced nuclear chain reaction, superintelligent AI in some form could arrive quite suddenly. The chances are that we would be unprepared: if we built superintelligent machines with any degree of autonomy, we would soon find ourselves unable to control them. I am, however, fairly confident that we have some breathing space because there are several major breakthroughs needed between here and superintelligence, not just one. [11, Chap. 3, § 7]

The remainder shall unfold straightforwardly as follows. In the next section we summarize what Russell offers as a solution to the threat to humanity from superintelligent AI. The section after that presents in sequence four problems that plague his proposal. Finally, the paper concludes with a brief discussion of the next steps to be taken in our assessment of Russell's approach, and in our consideration of competing approaches.

II. Russell's Proposed Solution

What is the solution Russell proposes? We cannot cover the ins and outs of his solution, as doing so would require a detailed explanation of *reinforcement learning* (RL), including *inverse* RL (IRL), upon which his proposal rests. While these forms of learning are mathematically simple frameworks in which agents gradually get better at reaching toward a goal, we nonetheless have not the time and space here to burn in exposition — and besides which RL and IRL are well-known to AI researchers. (Russell's [11] *Human Compatible* is in fact itself an excellent non-technical introduction to these forms of learning.) Fortunately, the core of Russell's proposed solution, what he calls "Provably Beneficial AI" (PBAI), can be quite efficiently conveyed here. The core of PBAI is that we take care to engineer robots driven solely by a "desire" to reach goals that accord with the goals of humanity. Of course, desire in the human case entais that the human doing the desiring has some states of "phenomenal" or "subjective" consciousness (what Block [1] calls 'p-consciousness'). This is so because, as we humans all know, when one desires something, one *feels* things, inevitably. For example, if one intensely desires to get some reward, and works ferociously toward it, but keeps failing to even get close to obtaining it, one is likely to e.g. feel frustrated, angry, despondent, and so on. Thus, we use scare quotes around 'desire' so as not to assume any such thing as that the robots Russell seeks will have p-consciousness.[3]

Encapsulated, what then in Russell's PBAI is the reward "desired" by the machines? He maintains that that reward will be none other than our own collective maximal well-being. Since we can safely assume that such goals in our case include that our species survives, and indeed overall thrives, if such a "desire" can be counted upon to really and truly drive our future robots, we should as a species be in good shape. In addition, we must be able to comfortingly *prove*

[1]Lead author of the encyclopedic, leading introduction and overview of AI, now out in its fourth edition: [12].

[2]The fact is, he does not really tell us in his book why he is so sure superintelligent AI will arrive — but he certainly is sure it will. Our educated guess is that Russell is content with his observing in his book the failure of numerous arguments against the proposition that superintelligent AI will arrive.

[3]According to the first author, they will have no such thing, and in fact no one at present has the slightest clue as to how to proceed with engineering that can be rationally regarded to move a nanometer closer to p-conscious AIs, as explained in [2].

 ISBN: 978-1-68558-014-8

that the robots are beneficial to humanity. Here is how Russell expresses overall his rather rosy take on things:

> [M]y proposal for beneficial machines: machines whose actions can be expected to achieve *our* objectives. Because these objectives are in us, and not in them, the machines will need [via IRL] to learn more about *what we really want* from observations of the choices we make and how we make them. Machines designed in this way will defer to humans: they will ask permission; they will act cautiously when guidance is unclear; and they will allow themselves to be switched off. [11, ¶ 2, § "Beneficial Machines" in Chap. 10 "Problem Solved?'; emphasis ours]

Unfortunately, while we have deep respect for the formality of Russell's approach (unsurprising since any real formality is rooted in formal logic and proofs therein: there is no other way to achieve a proof by to employ formal logic) there are four each-fatal-in-their-own-right problems plaguing Russell's proposal, as we now explain. Here now are these problems.

III. Four Problems Afflicting Russell's PBAI

As promised, we now proceed to explain, in turn, four defects (among others) that afflict PBAI.

A. Problem 1: Sola Utilitarianism?

The first problem is simple to grasp, and simply devastating; it is that Russell's proposal to save our race is based upon *only* the family of consequentialist ethical theories. This family includes the familiar ethical theory known as *act utilitarianism*, according to which what is obligatory are actions that maximize overall happiness; a precise account can be found in the classic [7]. But surely this particular family is only an *option* from among many families of ethical theories; and, these families are pairwise inconsistent. That is, pick any two families, and the definitions they include for the central operators of any ethical theory, for instance for *obligatory*, and one will arrives at contradictions, by elementary deductive reasoning over these definitions in garden-variety contexts. To see this, let us pick for consideration another ethical-theory family. Specifically, let us pick for expository purposes the family of *divine-command* ethical theories. Divine-command ethical theories are based upon the core notion that what is obligatory, permissible, forbidden, and so on is wholly determined by God's commands. A seminal presentation of a divine-command ethical theory is given by [10]. Exploration of divine-command ethical theories in a manner that conforms to what is needed in attempts to engineer morally correct machines is carried out in [4]. Note that when one considers the entire population of planet Earth, and subscription among its members to a dominant family of ethical theories, it is probably the divine-command family that has the largest number of adherents, by far.[4]

Now, given the setup supplied in the previous paragraph, here is a pair of relevant biconditionals, one from each of the two families we have just cited.[5] The first is part of act utilitarianism; the second is from all divine-command ethical theories.

Ob_{U} An agent (a category that includes human persons) is obligated at time t, given (context) Φ, to do action a at later time t' if and only if a, from among all viable alternative actions available to this agent, brings about the most happiness for the most people.

Ob_{DC} A human person is obligated at time t, given (context) Φ, to do action a at later time t' if and only if the performance of a has been commanded by God (or is deductively entailed by what has been commanded by God).

We are quite sure the reader can see the problem. By 'context' here, represented by 'Φ,' is meant simply a collection of declarative formulae, or for our somewhat informal exposition here, declarative propositions, that sets the situation. We can consider a hypothetical to make this more concrete: Molycarp is a devout Christian living under a brutal dictatorship whose key tenets include those of rabid and unrelenting atheism, and Molycarp is imprisoned, tortured, and asked to explicitly utter blasphemous and profane denial of his orthodox conception of Jesus as sinless and divine.[6] *Ex hypothesi*, Molycarp's agreeing to do this will save his life, ensure the well-being of his family (for which he is the breadwinner), and bring about many, many other happiness-bearing states-of-affairs through an endless array of chains of weal catalyzed by his subsequent actions. However, if he accepts death, only two terrestrial people will ever know what happened to him (the dictator and the executioner), as he will be incinerated, and in fact soon after his death everybody else will thoroughly forget about him. By a suitable instantiation of Ob_{DC}, Molycarp is obligated to proclaim his belief in Jesus and his divinity, and die a martyr; but in stark contrast, by a suitable instantiation of Ob_{U}, he is obligated to go through the motions of quickly spouting out a few words that will secure his freedom, and a lot of happiness that cannot otherwise be secured. Assuming that no one can be obligated to perform two actions that are impossible to both perform,[7] we have a contradiction.

There is more general, history-centric way to sum up Problem 1 for Russell, and for those inclined to follow him; it is to simply report that the discipline of systematic, theoretical ethics has been in progress since at least Aristotle, three centuries before the birth Christ, and if we know anything at all about the history of the discipline from that ancient timepoint we know that the human race has on hand myriad families of ethical theories, each none other than, as we have noted above, pairwise incompatible. It is thus rather doubtful that the solution to the problem posed by future superintelligent

[4]There are currently e.g. about 2.2 billion Christians on Earth, and about 2 billion Muslims. For both groups, by definition, it is first and foremost what God commands that determines what is obligatory. Orthodox and conservative Jews would of course be in precisely the same category. (This is of course not at all to say that the three religions here each perfectly agree on every attribute ascribed to God. The main ones, though — e.g. omnipotence, omniscience, omnipresence, omnibenevolence, creator of all contingent things — are indeed ascribed to God in the case of each of the trio of religions we cite here.

[5]For easing exposition, let us not worry about which particular ethical theory is in play here from each of the two families we have called out.

[6]The sinlessness and divinity of Jesus is a credal doctrine of orthodox Christianity. See e.g. [13]. Many readers will see in our use of 'Molycarp' a thinly disguised reference to the real martyr Polycarp, executed in 155 AD.

[7]This, that "ought implies can," is known as *Kant's Law*, and is a staple in deontic logic, the branch of logic devoted to logicizing ethical theories.

 ISBN: 978-1-68558-014-8

machines is to be found in the Russellian engineering of robots whose *modus operandi* is the following the dictates of only *one* family, consequentialism.

B. Problem 2: Mental States Not Inferrable from Behavior

The second problem afflicting Russell's approach to the threat to humanity is that this approach at its heart relies upon the ability of present and future AIs to infer a human's interior mentality from that human's exterior, readily observable behavior. After all, what Russell (admirably and rationally) wants is for the machines in question to place our happiness first among the goals they seek — but what is happiness if not a mental state, and as such an *invisible* state? (This is why we emphasized the phrase 'what we really want' in the quote of Russell just above.) This particular sentence is being written (at least in its first version) by author Bringsjord, who is thus simply staring at a screen and typing as characters appear on said screen for this eyes to take in. Okay, so suppose you walk up now to Bringsjord, who is seated, and look at his face, standing above him; and suppose that he stops typing and looks at your face. Can you tell if Bringsjord is happy? You may of course be able to rationally *assert* that he is happy, because you may have empirical data regarding his recent past (e.g., that he had a gourmet lunch featuring arctic char at Manhattan's Aquavit restaurant, a particular favorite of his, before his the current work session you just interrupted), and you may even happen to have a live feed from Selmer's iPhone somehow, giving you his vitals and perhaps all sorts of information about this bodily state, including its over internal condition in many regards, but — again, we assume here that Selmer is staring at you, expressionless — you will only be guessing. And in fact you would be wrong. Reason? Selmer happens to be thinking about an event in his childhood, a rather sad one: the death of his dog King, caused by a car; and his current state is far from a happy one, mixed as it is with some rather dreadful mental movies of what happened that fateful day just outside New York City.

Now, just replace you with a robot (or with an AI using sensors in the relevant room) looking at Selmer, and you will see the problem facing Russell. AIs cannot toil on our behalf by using inductive reinforcement learning because they cannot learn the nature of what they need to reduce or increase: namely, our mental states.

C. Problem 3: Cognition Ranges Beyond The Turing Limit

The next problem is quite simple to state. The robots that will be toiling in our favor are explicitly asserted by Russell to be boxed in by what a Turing machine can do. This is easy to confirm, because when he offers a theorem-schema that, when proved, will provide the ultimate assurance he seeks in the face of impending doom from superintelligent machines, that theorem-schema employs 'machine,' and this term means *Turing-level* machine. (We look at Russell's theorem-sketch below, in the final section.) Put another way, the robots with which Russell is concerned are all constrained by the Turing Limit, the level of computational power beyond which Turing machines (and lesser machines, e.g. linear-bounded Turing machines). But that means that if our cognition, our intellectual power, extends *beyond* this limit, the robots will not be able to grasp and abide by our cognition. But according to Bringsjord, human cognition is indeed of this nature; see for example assertions and defenses of this claim in [3, 5, 6].

It is important to grasp that the problem here for Russell's PBAI paradigm is not weak, vague, or haphazard; in fact the problem is logico-mathematical in nature. Suppose one computing machine m_1 is not capable of computing functions beyond some *bona fide* level L_1, and that some other computing machine m_2 is capable of computing functions at some level L_2 above L_1.[8] It then is an easy theorem that m_1, by observing the operation of the more powerful m_2, cannot compute functions at L_2, or for that matter one iota above L_1. Yet, Russell pins his hopes on robots that will observe us, and figure out how to work to our benefit. But what if our benefit requires doing things that demand as much cognitive power as we have? In that case it is mathematically impossible for his salvific recipe to work.

D. Problem 4: Humans Do Not Agree on Weighty Propositions

Let us suppose for the sake of argument that the Russellian beneficial-to-us robots can indeed somehow be magically engineered, so that at every moment of their existence, and perpetually so, they toil for *the* benefit of humanity: their sought-after reward is that very benefit. Notice our emphasis on the word 'the' in the previous sentence. That tiny little word, a so-called "determiner," creates a fatal problem for Russell. The problem is that there is no *the* thing that is humankind's benefit. What would this thing be, after all? Masochists seek their own harm and pain; sadists the harm and pain of others; criminals their own material benefit at the expense and pain of others; Christians perpetual bliss in an afterlife, this earthly life being no more than — quoting David — a vapor and — quoting Solomon — at its best filled with soul-making suffering; "brave" existentialists like Camus expend what they admit is pointless effort to stay alive even though this life is evanescent and absurd; and so on seemingly *ad infinitum* into never-ending heterogeneity. So, there is no *the* benefit, alas. The bottom line for Russell's PBAI explodes it; that bottom line is that each relevant group of humans, with enough wealth, is going to purchase a robot or robots in order to facilitate *their* priorities. If anything, this will just make the world as contentious and chaotic as it is now — maybe more so.

IV. Next Steps

The alert reader will recall that there is a 'P' for 'provably' in Russell's 'PBAI.' What is it that Russell says we need to prove in his approach? He gives the general shape of the theorems which, if proved, will constitute assurance. We read:

[8] We spare the reader technical bases beneath this imagined state-of-affairs, but mention here that this means that the levels must be ones in the Arithmetic Hierarchy or Analytic Hierarchy, and genuinely distinct ones therein. We cannot be referring to levels in the Polynomial Hierarchy, because all problems in that hierarchy are Turing-solvable.

 ISBN: 978-1-68558-014-8

> Let's look at the kind of theorem we would like eventually to prove about machines that are beneficial to humans. One type might go something like this:
>
> > Suppose a machine has components A, B, C, connected to each other like so and to the environment like so, with internal learning algorithms l_A, l_B, l_C that optimize internal feedback rewards r_A, r_B, r_C defined like so, and [a few more conditions] . . .. Then, with very high probability, the machine's behavior will be very close in value (for humans) to the best possible behavior realizable on any machine with the same computational and physical capabilities.

Russell's main point here is that such a theorem should hold regardless of how smart the components become — that is, "the vessel never springs a leak and the machine always remains beneficial to humans" ([11, Chap. 8, § "Mathematical Guarantees," ¶ 8]). The next step in our evaluation of PBAI is to investigate carefully how theorems of this general shape can *in fact* be proved. This will require formalizing the concepts that Russell leaves vague and undefined here. For example, what, logico-mathematically speaking, is a 'machine' in the theorem-sketch that Russell provides here?[9] Likewise, what precisely is 'the environment'? At the very least, we shall need to venture precise answers to these questions in order to understand what Russell is gesturing toward when he sketches the kind of theorem to target in PBAI. We will then need to see if in fact an actual theorem of this shape can be proved, and what the proof would need to be like. Following on this, another step will be to see if, in approaches very different than PBAI, theorems providing greater assurance can be obtained. After all, Russell here concedes, explicitly, that the best his approach can reach is only "very high probability" that the machines will operate in our interests. We believe that total assurance can in fact be secured on the strength of proving theorems of a different nature than what Russell describes, and will seek to demonstrate that our optimism is well-founded.

ACKNOWLEDGMENTS

We are indebted to Stuart Russell for bravely and perspicaciously dealing with an acute future danger that many may wish to ignore or at least severely downplay. We are deeply grateful to ONR for past, extended support of research in the area of robot ethics (that informs the present paper), in particular through a MURI grant on which both Bringsjord and Govindarajulu were central researchers (along with PI Matthias Scheutz and Co-PI Betram Malle).

REFERENCES

[1] N. Block. On a Confusion About a Function of Consciousness. *Behavioral and Brain Sciences*, 18:227–247, 1995.

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

[3] S. Bringsjord and K. Arkoudas. The Modal Argument for Hypercomputing Minds. *Theoretical Computer Science*, 317:167–190, 2004.

[4] S. Bringsjord and J. Taylor. The Divine-Command Approach to Robot Ethics. In P. Lin, G. Bekey, and K. Abney, editors, *Robot Ethics: The Ethical and Social Implications of Robotics*, pages 85–108. MIT Press, Cambridge, MA, 2012. URL http://kryten.mm.rpi.edu/Divine-Command_Roboeth

[5] S. Bringsjord and M. Zenzen. *Superminds: People Harness Hypercomputation, and More*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003.

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

[7] F. Feldman. *Introductory Ethics*. Prentice-Hall, Englewood Cliffs, NJ, 1978.

[8] N. Govindarajulu and S. Bringsjord. On Automating the Doctrine of Double Effect. In C. Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, pages 4722–4730. International Joint Conferences on Artificial Intelligence, 2017. ISBN 978-0-9992411-0-3. doi: 10.24963/ijcai.2017/658. URL https://doi.org/10.24963/ijcai.2017/658.

[9] J. D. Hamkins and A. Lewis. Infinite Time Turing Machines. *Journal of Symbolic Logic*, 65(2):567–604, 2000.

[10] P. Quinn. *Divine Commands and Moral Requirements*. Oxford University Press, Oxford, UK, 1978.

[11] S. Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Books, New York, NY, 2019. This is the ebook version, specifically an Apple Books ebook.

[12] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, New York, NY, 2020. Fourth edition.

[13] R. Swinburne. *Was Jesus God?* Oxford University Press, Oxford, UK, 2010.

[9] Apropos of the discussion above, what about computing machines that are provably capable of more than what can be done by a standard Turing machine? E.g., what about infinite-time Turing machines [9]?

 ISBN: 978-1-68558-014-8