Journal of Experimental & Theoretical Artificial Intelligence
Vol. 26, No. 03, August 2014, doi 10.1080/0952813X.2014.895109

RESEARCH ARTICLE

Bounding the impact of AGI

András Kornai*
Computer and Automation Research Institute, Hungarian Academy of Sciences
and
Hariri Institute of Computer Science, Boston University
(*Received January 2013; final version received May 2014*)

Humans already have a certain level of autonomy, defined here as capability for voluntary purposive action, and a certain level of rationality, i.e. capability of reasoning about the consequences of their own actions and those of others. Under the prevailing concept of AGI we envision artificial agents that have at least this high, and possibly considerably higher, levels of autonomy and rationality. We use the method of bounds to argue that AGIs meeting these criteria are subject to Gewirth's dialectical argument to the necessity of morality, compelling them to behave in a moral fashion, provided Gewirth's argument can be formally shown to be conclusive. The main practical obstacles to bounding AGIs by means of ethical rationalism are also discussed.

Keywords: ethical rationalism, principle of generic consistency, formal verification

With the emergence of intelligent question-answering capabilities from IBM's Watson to Apple's Siri, the fear of autonomous agents harming humans, as old as mythology, has recently taken on new urgency (for a recent overview see Yampolskiy and Fox 2013, for a bibliographical summary see Muehlhauser 2012). There are three factors that make it particularly hard to assuage such fears. First, the stakes are high: just as humans can now quite accidentally wipe out other species and genera, a new breed of superintelligent machines poses an existential threat to humanity. Second, it is reasonable to fear the unknown, and there is very little we can know in advance about such superintelligent beings, whether there will be one such individual or many, one breed or many, or how they will view us. Third, the emergence of artificial general intelligences (AGIs) seems to be a slow but quite steady process, something we can understand but are in no position to stop, like continental drift.

The aim of this paper is not simply to ease such fears but to offer a research program that can actually guarantee that AGIs pose no existential threat. This will be a one-sided bound, staving off some of the worst consequences of Bostrom's (2012) Orthogonality Thesis that a high level of intelligence can be in the service of any goal, good or bad alike, and will say nothing about the good, possibly spectacularly good impacts that AGI may have on the future of humanity. In Section 1 we argue that no physical interlock or other safety mechanism can be devised to restrain AGIs, the guarantees we seek are *necessarily* of a mathematical (deductive, as opposed to algorithmic) nature. This requires some shift in focus, because in the current literature it is not some logical deductive system that is viewed as the primary descriptor of AGI behavior but rather some utility function

* Email: andras@kornai.com

whose maximization is the goal of the AGI. Yet, as we shall argue, deductive bounds are still possible: for example consider an AGI whose goal is to square the circle with ruler and compass – we know in advance that no matter what (static or dynamically changing) weights its utility function has, and no matter what algorithmic tricks it has up its sleeve, including self-modifying code, reliance on probabilistic, quantum, or other hypercomputing methods (Ord 2002), it simply cannot reach this goal.

In Section 2 we present the proposed restraining device, morality, and address the conceptual and practical difficulties attendant upon its use. The main conceptual difficulty is that the conventional human definition of 'morally sound' is highly debatable, and one can very well imagine situations in which AGIs consider it best, from their moral standpoint, to do away with all of humanity except for one 'Noah' family, and start with a clean slate. The main practical difficulty, well appreciated in the current literature, is to guarantee that morality is indeed imposed, even on AGIs that may be capable of transcending the limitations placed on them by their designers.

The central element of our proposal, *ethical rationalism,* is due to Gewirth (1978), with significant arguments and counter-arguments scattered through the literature, see in particular Regis (1984) and Beyleveld (1992). The basic construction is a *prospective purposive agent* (PPA) who can *act* with purpose and *reason* rationally – clearly these are conditions that any future AGI will meet just by virtue of being AGI. From these premisses Gewirth derives a variant of the Golden Rule he calls the *principle of generic consistency* (PGC) 'Act in accord with the generic rights of your recipients [to freedom and well-being] as well as of yourself'. The research program outlined here is to turn this from a philosophical argument into a formally verified proof. There are plenty of technical problems, such as devising the right kind of action logic to sustain the proof, but these are, we argue in Section 2, the good kind of problems, the kind that AGI research needs to concern itself with anyway.

There are notable difficulties in the way of this program, even if we succeed in hardening ethical rationalism into a theorem of logic, these will be discussed in the concluding Section 3. First, there is the issue of pattern recognition – given humanity's propensity to disregard the PGC, what reason is there to believe that we will be recognized as PPAs and are deemed worthy of protection by the PGC? Second, even though a strong argument can only be disregarded on pain of contradiction, the pain so inflicted is relatively mild, and we see PPAs living and functioning in a self-contradicted state all the time. Third, a proof presupposes not just the premisses, but also the reliability of the logical apparatus it employs. As we shall see, these problems are closely related.

1. The method of bounds

Let us begin with a small example. As every student of elementary combinatorics knows, if I have a thousand books in my library but decide to keep only half of them, I can do this $\binom{1000}{500}$ ways. Suppose I have a 2GHz processor and it takes only one cycle to evaluate any single alternative by some utility function, and another cycle to compare it to the best so far, so that I can find the best of a billion alternatives every second. Do I need to actually compute 1000! and divide it by the square of 500! to know that this is not going to work? No, knowing that $\frac{4^n}{2n+1} < \binom{2n}{n}$ is quite sufficient, since this yields over 10^{289} seconds, far longer than the estimated lifetime of the universe. Can someone give me a better processor? Well sure, but even the

best processor cannot perform more than one operation per Planck time unit, so I still need 10^{254} seconds. Can someone give me more of these processors? Well sure, but their aggregate must still stay within the computational capacity of the universe, estimated by Lloyd (2002) at 10^{120} operations, so I still cannot make the choice by exhaustive search within the lifetime of the universe. Can someone give me a good supply of universes that I could devote to this computation? Well no, we are restricted to this universe so strongly we cannot even form a convincing idea what access to other ones would mean – even the basic ground rules like Barcan's formula $\Diamond\exists xFx \rightarrow \exists x\Diamond Fx$ are in grave doubt.

The point of our example is not to convince the reader that in this universe brute force computation soon runs out of steam, for the reader already knows this perfectly well – the point to notice is that we didn't need Stirling's formula. Few readers will remember the remainder term in Stirling's formula to compute $\binom{1000}{500}$ to two significant digits, even fewer would have the patience to actually simplify a fraction with a thousand terms in the numerator and another thousand in the denominator, and yet fewer would be able to carry the computation through without arithmetic error. Yet it is clear that $\sum_{i=0}^{2n}\binom{2n}{i} = 2^{2n}$, that the central term $\binom{2n}{n}$ is the largest among these, and therefore we have $\frac{4^n}{2n+1} < \binom{2n}{n}$. It didn't matter that this lower bound is rather crude, as it was already sufficient to establish that the task is unfeasible.

It is characteristic of the method of bounds that we will apply here that hard and complex questions can be avoided as long as we don't seek exact answers. There is a significant body of numerical analysis beginning with the Euler-Maclaurin formula and culminating in Lánczos (1964) that is pertinent to the question of computing $\binom{2n}{n}$, but we could settle the issue without going anywhere near this literature. In fact the whole apparatus of calculus, and all arguments based on continuity and limits could be dispensed with in favor of elementary statements concerning positive integers such as the sum is greater than the parts or that by decreasing one term in a sum we decrease the entire sum. Similarly, when we apply the method of bounds to issues of predicting AGI behaviour, we can steer clear of all the difficulties that actually specifying some utility function would entail – not even a superhuman/hypercomputing agent can falsify Lindemann's work on the transcendence of π or the implication that the circle cannot be squared. Equally important, we can steer clear of the difficulties in trying to predict for each possible set of circumstances exactly what would, or would not, constitute moral behavior, a matter we shall return to in Section 3.

There is, to be sure, a fair bit of calculus in Lloyd's bound on the computational limits of the universe, just as there is a significant amount of physics taken for granted in the ITRS roadmap that gives our current best assessment of future CPU speeds. One needs to distinguish the *precision* of a theory, which we will assess shortly, from its *reliability* – the essence of the method of bounds is that we can trade in precision to obtain greater reliability. When dealing with the existential threat posed by AGIs, controlling the sophistication of the deductive apparatus is not just prudent, but as we shall see in 1.2, a positive requirement. In the Appendix we will present some simple and straightforward estimates of the actual magnitude of the threat, and arrive at a safety engineering limit of no more than one error in 10^{64} logic operations. To set the stage for our main argument, we must first compare this number to what we can expect from science.

1.1 *How precise is physics?*

In a startling image, Feynman (1985) likens the 10^{-10} fit between predicted and measured values of the electron's spin g-factor to being able to tell the distance between New York and Los Angeles within the thickness of a human hair. Since that time, both calculation and measurement precision has actually improved by two orders of magnitude, so that the uncertainty of g_e is now below one part in 10^{12}. But this is exceptional: most physical constants are known to us only to 9 or 10 significant digits, and some, like the gravitational constant G, only to 4. Comparing Taylor et al. (1969) to Mohr et al. (2012) shows that it takes at least two decades of advances in metrology to gain a single digit of precision, so getting to 64 digits from 10 is rather unlikely in our lifetimes.

Somewhat optimistically we can describe early 21st century technology as operating in the nano, and physics as operating in the pico range: industrial processes controlled to 9 decimal places, and individual measurements yielding 12 significant digits are becoming increasingly common. We would not be particularly surprised to see 5 orders of magnitude gains in both by the end of the century. It is also possible that our current theory of physics is actually a lot better than the current state of the art in metrology would lead us to believe, and that with better computation we can get to 24, or even 36 digits of precision without any new physics. But we do not, and before actually making the measurements simply *cannot* know that this is so, and if safety from existential threat requires 64 digits or better, there is currently, and in the foreseeable future, simply nothing in the physical environment that we can manipulate in a way that would fit the bill.

In early versions of his theory of Friendly AI, Yudkowsky (2001) actually sought mathematical guarantees that AGI won't pose an existential threat to humanity, but this idea met with considerable resistance, especially as it was somewhat unclear what *kind* of mathematics is to be deployed. The main goal of this paper is to provide a specific direction within mathematics, for once we acknowledge that the search for any physical solution must cross a gap of over 50 orders of magnitude, mathematical guarantees remain the only feasible solution. The fundamental constants of mathematics like e, γ, or π were already known to several hundred digits before the advent of mechanical calculators, and are now known to millions (in the case of π, trillions) of significant digits, far more than the few dozen we could conceivably need to compute any physical quantity.

Unfortunately, the history of calculating such numbers is rife with errors: for example, in 1790 Mascheroni attempted to calculate γ to 32 digits but his results were only correct to 19 digits, in 1873 Shanks calculated π to 707 places but only the first 527 were correct. To establish a bound we can trust with our lives, we must look at the reliability of mathematical argumentation. As we shall see, more important than the failures of numerical calculations are the cases where the logic of the deduction is faulty.

1.2 *How reliable is mathematics?*

The period since World War II has brought incredible advances in mathematics, such as the Four Color Theorem (Appel and Haken 1976), Fermat's Last Theorem (Wiles 1995), the classification of finite simple groups (Gorenstein 1982, Aschbacher 2004), and the Poincaré conjecture (Perelman 1994). While the community of mathematicians is entirely convinced of the correctness of these results, few individual mathematicians are, as the complexity of the proofs, both in terms of knowledge assumed from various branches of mathematics and in terms of the length of the deductive chain, is generally beyond our ken. Instead of a personal understanding

of the matter, most of us now rely on *argumentum ad verecundiam:* well Faltings and Ribet now think that the Wiles-Taylor proof is correct, and even if I don't know Faltings or Ribet at least I know and respect people who know and respect them, and if that's not good enough I can go and devote a few years of my life to understand the proof for good. Unfortunately, the communal checking of proofs often takes years, and sometimes errors are discovered only after a decade has passed: the hole in the original proof of the Four Color Theorem (Kempe 1879) was detected by Heawood in 1890. Tomonaga in his Nobel lecture (1965) describes how his team's work in 1947 uncovered a major problem in Dancoff (1939):

> Our new method of calculation was not at all different in its contents from Dancoff's perturbation method, but had the advantage of making the calculation more clear. In fact, what took a few months in the Dancoff type of calculation could be done in a few weeks. And it was by this method that a mistake was discovered in Dancoff's calculation; we had also made the same mistake in the beginning.

To see that such long-hidden errors are by no means a thing of the past, and to observe the 'web of trust' method in action, consider the following example from Mohr (2012).

> The eighth-order coefficient $A_1^{(8)}$ arises from 891 Feynman diagrams of which only a few are known analytically. Evaluation of this coefficient numerically by Kinoshita and co-workers has been underway for many years (Kinoshita, 2010). The value used in the 2006 adjustment is $A_1^{(8)} = -1.7283(35)$ as reported by Kinoshita and Nio (2006). However, (...) it was discovered by Aoyama et al. (2007) that a significant error had been made in the calculation. In particular, 2 of the 47 integrals representing 518 diagrams that had not been confirmed independently required a corrected treatment of infrared divergences. (...) The new value is (Aoyama et al. 2007) $A_1^{(8)} = 1.9144(35)$; (111) details of the calculation are given by Aoyama et al. (2008). In view of the extensive effort made by these workers to ensure that the result in Eq. (111) is reliable, the Task Group adopts both its value and quoted uncertainty for use in the 2010 adjustment.

Assuming no more than three million mathematics and physics papers published since the beginnings of scientific publishing, and no less than the three errors documented above, we can safely conclude that the overall error rate of the reasoning used in these fields is at least 10^{-6} per paper, which is notably (by 3-6 orders of magnitude) higher than the imprecision of physics. (This is not entirely fair, in that we are comparing the *best* established results in physics with the *average* of mathematics. A fuller investigation of physics papers may establish a higher, or at least comparable error rate relative to what we see in mathematics.)

1.3 *The role of automated theorem-proving*

That human reasoning, much like manual arithmetic, is a significantly error-prone process comes as no surprise. Starting with de Bruijn's Automath (see Nederpelt et al. 1994) logicians and computer scientists have invested significant effort in mechanized proof checking, and it is indeed only through such efforts, in particular through the Coq verification (Gonthier 2008) of the entire logic behind the Appel and Haken proof that all lingering doubts about the Four Color Theorem were laid to rest. The error in $A_1^{(8)}$ was also identified by using FORTRAN code generated by an automatic code generator (Mohr et al. 2012).

To gain an appreciation of the state of the art, consider the theorem that finite groups of odd order are solvable (Feit and Thompson 1963). The proof, which took two humans about two years to work out, takes up an entire issue of the Pacific

Journal of Mathematics (255 pages), and it was only last year that a fully formal proof was completed by Gonthier's team (see Knies 2012). The effort, ∼170,000 lines, ∼15,000 definitions, ∼4,200 theorems in Coq terms, took person-decades of human assistance (15 people working six years, though many of them part-time) even after the toil of Bender and Glauberman (1995) and Peterfalvi (2000), who have greatly cleaned up and modularized the original proof, in which elementary group-theoretic and character-theoretic argumentation were completely intermixed.

The classification of simple finite groups is two orders of magnitude bigger: the effort involved about 100 humans, the original proof is scattered among 20,000 pages of papers, the largest (Aschbacher and Smith 2004a,b) taking up two volumes totaling some 1,200 pages. While everybody capable of rendering meaningful judgment considers the proof to be complete and correct, it must be somewhat worrisome at the 10^{-64} level that there are no more than a couple of hundred such people, and most of them have something of a vested interest in that they themselves contributed to the proof. Let us suppose that people who are convinced that the classification is bug-free are offered the following bet by some superior intelligence that knows the answer. You must enter a room with as many people you can convince to come with you and push a button: if the classification is bug-free you will each receive $100, if not, all of you will immediately die. Perhaps fools rush in where angels fear to tread, but on the whole we still wouldn't expect too many takers.

1.4 *The reliability of rational argument*

Whether the classification of finite simple groups is complete and correct is very hard to say – the planned second generation proof will still be 5,000 pages, and mechanized proof is not yet in sight. But this is not to say that gaining mathematical knowledge of the required degree of reliability is hopeless, it's just that instead of monumental chains of abstract reasoning we need to retreat to considerably simpler ones.

Take, for example, the first Sylow Theorem, that if the order of a finite group G is divisible by some prime power p^n, G will have a subgroup H of this order. We are *absolutely certain* about this. Argumentum ad verecundiam of course is still available, but it is not needed: anybody can join the hive-mind by studying the proof. The Coq verification contains 350 lines, 15 definitions, 90 theorems, and took 2 people 2 weeks to produce. The number of people capable of rendering meaningful judgment is at least three orders of magnitude larger, and the vast majority of those who know the proof would consider betting their lives on the truth of this theorem an easy way of winning $100 with no downside risk.

Could it be the case that in spite of all these assurances we, humans, are all deluded into accepting Sylow's theorem? Yes, but this is unlikely in the extreme. If this so-called theorem is really a trap laid by a superior intelligence we are doomed anyway, humanity can find its way around it no more than a bee can find its way around the windowpane. With regards to physics, the same point can be made. The single most glaring discrepancy between astronomical observation and Newtonian physics was the perihelion precession of Mercury, but even here it takes over 10^6 years for the discrepancy to add up to an extra turn. New physics may shatter our entire conceptual framework of thinking about the domain, but still it will be *conservative* in the sense of respecting our existing measurements. We are for most purposes quite satisfied with Newtonian mechanics, especially as relativity brought to us a better understanding of its domain of applicability.

To summarize our conclusions so far, we propose to bound AGIs by methods that

rely neither on high precision measurements nor on highly complex arguments. If you are a finite group of size $p^n m, (p, m) = 1$ it doesn't matter what you believe about your subgroups of order p^n – you have some, they are isomorphic, and I can rely on you having them even if I don't know your multiplication table in detail. If you are delusional about not having any, I can take advantage of this. What needs to be emphasized in this situation is that Bayesian reasoning and the concomitant notion of 'degree of belief' is *totally irrelevant*. According to the received theological doctrine (originating with St. Anselm of Canterbury and St. Thomas Aquinas) not even an omnipotent God can create a finite group that lacks Sylow subgroups.

In a small way, we have already done what we set out to do. We have bound all future AGIs to respect Sylow's Theorem. They can mess with finite groups all they want, they can dwarf human intellect every way, but they cannot build a group with 630000000000000000000000 elements that has no subgroup of order 9, they cannot square the circle with ruler and compass, and so forth. What we need to do is to bind them to ethical principles *the same way*. In fact, this is the only truly novel element of our proposal, as the overall goal of somehow endowing AGIs with morality is not new (for a modern summary see Wallach and Allen 2009) and as a fundamental ethical precept the PGC is strongly related to the categorical imperative, which has already received considerable attention as a possible basis of machine morality, see in particular Allen (2000) and Powers (2006).

2. Ethical rationalism

Our goal is to obtain guarantees of friendliness in a purely deductive fashion. We emphasize at the outset that this is considerably less than what proponents of machine morality generally set out to do: we are not interested in a consistent and complete system of ethics that will tell us in advance what we ought to do in any given circumstances, we are only interested in guidelines that are strong enough to stave off existential threat. In particular, we do not suppose that AGIs need to work toward the benefit of humankind, or to preserve, let alone enhance, the rich fabric of human values. In fact we do not want to presuppose any value system at all, especially as there is a whole school of philosophical thought, starting with Mackie (1977) that takes values to be nonexistent in the first place. Values emerge from Gewirth's analysis, first as entitely subjective valuations of certain things, with no committment to what these certain things are, and later as necessarily inclusive of certain *rights* the agent must have if it is to be an agent at all. Just as there can be many proofs of the same theorem, there could be many deductive arguments to the desired effect, but so far there seems to be only one, presented in Gewirth (1978), that appears to meet our principal requirement of not using any premiss that lacks empirical evidence. How good is this argument? According to Regis (1984), Gewirth

> gives every appearance of having developed a watertight case, for its arguments are set out with enormous deductive rigor and a frightening dialectical skill. To read Gewirth is to experience the sense of being caught in an ever-tightening net from which all conceivable avenues of escape have been blocked in advance. This is "philosophy as a coercive activity," and Gewirth comes quite close to the extreme of propounding "arguments so powerful they set up reverberations in the brain: if the person refuses to accept the conclusion, he *dies*." Nevertheless, Gewirth's arguments are not 'flashy.' They do not proceed by introducing wildly bizarre examples at crucial points; there is no delight in puzzlement for its own sake (...) or contrary-to-fact conditions imposed on imaginary beings hopefully making moral decisions. Rather, Gewirth proceeds by relentlessly piling reason upon reason for thinking that his conclusions are true, and by answering in advance almost every argument for thinking otherwise.

This is not to say that the community of philosophers is uniformly convinced. There are many critical voices, such as Bond (1980) who says that by Gewirth's argument

> moral evil is reduced to logical error. (...) Gewirth and others like him would turn wickedness into a kind of intellectual incompetence

or Nielsen (1984), who states plainly that Gewirth's central thesis

> that there is a substantive supreme principle of morality, the denial of which is self-contradictory (...) just has to be wrong, and the task (...) is to locate the place or places where such an argument went wrong.

Critics like Bond may even have it right: the whole point of the enterprise is to demonstrate that wickedness is indeed a form of intellectual incompetence, for if this much is true, the more competent AGIs will restrain the less competent ones from doing wicked things, just as the more competent humans tend to do with the less competent ones (we return to this point in 3.3). This is not to say that the social process limiting wickedness is perfect, modern history is full of counterexamples from the Third Reich to Cambodia. Obviously, the lower bound on AGI impact cannot be placed below the impact of an exceptional human individual, be their role viewed as positive (say, the appearance of a significant new advocate of non-violence, such as Mahatma Gandhi) or as negative, such as the appearance of a new dictator.

Whether critics like Nielsen have it right is another matter entirely. Given the sheer size of Gewirth's argument, 380 pages fully elaborated, with the skeletal version provided in Beyleveld (1992 Part I) running to 60 pages, and given the sophistication of the methods it uses, it demands serious investment of time and energy to fully grasp it, a problem that is faced by ultimately wrong and ultimately right proof attempts alike. The point is not to silence those like Nielsen who are strongly disinclined to accept the argument, it is just as important to seek holes and counterexamples as to strengthen the argument and patch up the holes; the point is to replace philosophical argumentation by formal proof. Here we can only take the first steps in analyzing the argument from the perspective of bounding AGIs. Again we begin with a small example.

2.1 *Formalizing philosophical arguments*

Modern artificial theorem proving techniques have largely fulfilled the Leibnizian dream of a *calculus ratiocinator* that would enable symbolic, not just numeric, reasoning by machine. To formalize a philosophical argument we need just four things: (i) some language describing the expressions we are interested in; (ii) some rules for deriving conclusions from premisses; (iii) some methods to see whether a given rule is applicable; and (iv) some methods to see whether premisses are met.

There are difficulties at every point. (i) Philosophical arguments are generally given in natural language, as opposed to the formal languages used in logic. As it happens, human-generated mathematical proofs are also published in natural language, and it is well known that a major part of the verification effort lies in translating this language to the language of the theorem prover. (In fact, formalizing Wiles' Fermat proof, comparable in terms of printed pages to the Feit-Thompson proof, has not been accomplished yet, see in particular Hesselink et al. 2006.) If this is already a problem for the highly constrained 'natural' language used by mathematicians, it is bound to be even more of a problem for the less constrained language of philosophical discourse. In a similar vein, (ii) if the already highly formal deductive style of mathematics is hard to coerce into the mechanical style

employed by the theorem prover, the informal deductions employed in philosophy cannot be any easier. As for (iii), artificial theorem provers need significant human guidance to find the points where a deductive pattern can be fruitfully matched against the set of true statements already generated from the premisses, and the 'soft' pattern matching we see in philosophy may pose even more serious problems. (iv) On top of this, there is a lack of agreed-upon model theory, and the grounding of philosophical arguments can be surprisingly weak.

Yet in spite of all this, a good argument can be highly compelling. Let us consider the following statement from St. Thomas Aquinas: *even God can't create a mountain without creating a valley*. For Aquinas, this illustrates a stronger statement, that omnipotence is limited to the possible, but we need not be actually interested in the notion of omnipotence to appreciate the argument. Let us see how the difficulties enumerated above play out in this case. We take the argument to mean ∀x create(x, mountain) ⇒ create(x, valley). We don't need to play with the tricky connective *even*, and we don't need a strong notion of *God*. There may be natural language issues, but they do not appear insurmountable. Competent speakers of English (and Latin) will agree that the formulation preserves the hard part: if the formal theorem can be seen to be false the reasoning behind the natural language statement was weak, and if it can be seen to be true, $\forall x$ must cover even God, so we achieved the effect Aquinas aimed at.

Clearly, a proof cannot be based on the strength of the logical connectives that appear in it, we need some substantive statements about the nature of mountains. We take this as *mountain* $\stackrel{d}{=}$ 'land higher than surrounding land'. By substitution, if x creates land y higher than surrounding land, (some) land z lower than y was created by side effect, this is recognized by the soft pattern matching as the *valley*, QED. If the definitions are reasonable, as they are in this case, the conclusion is inevitable. Weak grounding is not a problem, in fact we even gained scope by it, since the same abstract logic applies to electric potential and everywhere else where comparing heights makes sense.

Readers energized by St. Thomas' argument may wish to pursue the ramifications for other kinds of nouns defined by comparative adjectives, for relational nouns like *parent*, or for plain subsumption (one clearly cannot create a white horse without creating a horse), and so forth. As with any good proof, we soon begin to see that it may have a lot broader scope than what was needed to complete the job at hand.

2.2 *Outline of the argument*

Gewirth presents his argument *dialectically*, in the original sense of Socratic dialog, rather than in the Hegelian sense of dialectic. This has the advantage that the person the dialectic is aimed at is very soon forced into admitting being a *prospective purposive agent* (PPA) who can *act* with purpose and *reason* rationally. Crucially, the PPA is not assumed to subscribe to any elementary moral prescript, or even the everyday notion of good and bad, let alone good and evil. Such notions, with remarkably specific definitions that make it clear that Gewirth is not just 'playing with words', emerge in the course of the argumentation. Following Beyleveld's summary, the main steps of the argument (numbering as in the original) are:

(1) I (intend to) do X voluntarily for some purpose E
(2) E is good (by my definition of 'good')
(3) My freedom and well-being (F&WB) are generically necessary conditions of my agency
(4) My F&WB are necessary goods

(5) I have (maybe nobody else does) a claim right to my F&WB
(9) Other PPAs have a claim right to their F&WB
(13) Every PPA has a claim right to their F&WB

Gewirth is particularly careful in defending his conclusion against the adeontic viewpoint that there are no claim rights (*ought* statements), the amoralistic viewpoint that I may have claim rights but nobody else does, the consequentialist (classic utilitarian) viewpoint, and so forth. (This is clearly not the place to summarize the debate surrounding the issue, but readers strongly committed to a Moorean notion of a 'naturalistic fallacy', or to the 'error theory' of Mackie (1977) will find Gewirth (1984) and Stilley (2010) good entry points.) Gewirth is reaching in a deductive fashion some conclusions that have been arrived at in the AGI context both by Omohundro (2008) and Bostrom (2012) by appeal to considerations of fitness: in particular, we see his notion of freedom and well-being as a subset of Omohundro's basic AI drives and Bostrom's instrumental goals. While cast in a very different (less contemporary but perhaps more rigorous) language, in (3) Gewirth in fact argues for a stronger case than what was made by Omohundro and Bostrom, as he sees F&WB as *generically necessary* conditions of action.

It is, however, not entirely clear that the capacity to reason, in the sense taken for granted by Gewirth, is strictly speaking necessary for AGIs. One may make a strong argument that such capacity will increase fitness, and certainly humans who already have this capacity, even if in a somewhat error-prone fashion as discussed in 1.3, are unlikely to be seriously threatened by any 'intelligence' incapable of abstract reasoning. Notice that reasoning in the abstract, e.g. Rybka's chess playing capability, implies no particular committment to the kind of symbol-manipulation that was central to GOFAI, it simply means that we can use some internal model to make useful predictions about the consequences of various actions.

Bostrom's Orthogonality Thesis that any level of intelligence can in principle be combined with any final goal is largely borne out by self-inspection. As the current best instantiation of General Intelligence we, humans, are free to choose our final goals. In a more strict sense of orthogonality, intelligence and goals are unlikely to be entirely uncorrelated. In humans we find their goals, as expressed e.g. by choice of career, to be quite predictive of their level of general intelligence, and the negative correlation between criminality and IQ is rather well known, not just at the individual, but also at the state aggregate level (McDaniel:2006). If it can be shown error-free, Gewirth's argument will actually trump the Orthogonality Thesis for the class of AGIs that do have reasoning capabilities sufficiently evolved to comprehend it – we return to this matter in 3.3.

Given the scrutiny Gewirth's argument already received in the philosophical literature (see in particular Regis 1984 and Beyleveld 1992), if there are holes in applying the argument to AGIs they are less likely to come after the premiss (1), aimed really at rational human beings, who will be hard put to deny that they have at least *some* intentions to do *something* voluntarily. But a loose coalition of AGIs may even deny the existence of a unified *'I'* that is the subject of the dialectic (a matter we shall return to in 3.2), and a superintelligent being may have very good reasons to deny some of the commonsensical assumptions about space and time, actions and consequences, goals and purposes that Gewirth is relying on.

To act with purpose is to act in a voluntary and intentional manner, so a PPA will have at least some notion of some later time. This is already a lot. First, the world must be such that PPAs can have relatively stable dispositions, especially if they can commit to actions that will have to be performed after some delay. An intention to read the next issue of the Atlantic Monthly cover to cover implies not only that there will be a next issue (which is not quite a given) but also

that I will remember this commitment when the time comes to fulfill it. Second, it entails that we have a means of dealing with failed intentions, since in reality there are such things. Third, we have to be able to stabilize an intention in the sense that $X\tau$-intends Y at time t (that is, X intends Y to hold at $t+\tau$) will not be considered true if X is free to change its mind between t and $t+\tau$. It is not that such problems are insurmountable, in fact several solutions are known to the largely analogous Yale Shooting Problem, but to formalize the entire argument we will need to extend standard action logic (Thielscher 1998, Magnusson 2007) to mental acts and dispositions as well.

Besides a strong reliance on abstract entities like *purpose, freedom,* or *right*, which can be problematic for a strictly reist model theory, the AGI researcher will immediately note several other characteristics of Gewirth's argument that make formalization a hard task. First, all the reasoning takes place in an ideally resource-unlimited manner: in particular, performing actions or having intentions are largely treated as activities that require no (or negligible) material resources and no (or negligible) time. In reality, many moral conflicts stem from the fact that we need to act before we can think through all the relevant consequences of our actions. This is especially true of deliberate action which may have untold consequences on a large timescale, such as an invention. The inventor of freon could not have possibly foreseen all the consequences. Yet he decided to release the substance to the world, based on imperfect information and a very finite amount of time devoted to reasoning. That this kind of idealization can pose problems was already clear in antiquity "Before he could put into practice something he had heard, the only thing Tzu-lu feared was that he should be told something further" (Analects V.14). The contemporary computer scientist is constitutionally incapable of thinking in a resource-unlimited manner, so the original proof is in a sense better suited to purely mathematical inquiry, with resource bounds added in only afterwards.

Another issue, long familiar to students of logic but not particularly touched upon by Gewirth, is the reflexive strength of the deductive system. It is clear that in a world with more than one PPA, there are advantages accruing to each PPA from building internal models of how some other PPA (PPAO) may behave. In particular, if we are smarter than PPAO we may anticipate its moves and gain all kinds of advantages from doing so. (If we are a lot smarter, we may be able to build a full model and *emulate* PPAO, a matter we shall return to in 3.3.) We also need to be able to reason about our own reasoning, if only to figure out how PPAO will reason about us. We don't necessarily need fully reflexive reasoning (agents who can reason about reasoning about reasoning about ... their own reasoning), but in a resource-unlimited setting there seem to be some advantages that an n-fold reflexive PPA will have over an k-fold reflexive PPA for $n > k$. Finally, it should be added that it is not just the epistemic and the deontic modalities that play a significant role in formalizing the argument, but alethic modality is also essential, in that Gewirth aims at strict (categorical, exceptionless, necessary) conclusions at every stage. As we already emphasized at the outset, controlling the power of the modal logic used in formalizing the argument is very much part of the task (see also 3.3).

But when all is said and done, we do not see any of these difficulties as fatal to the project of formally verifying the argument from (1) to (13). The task is obviously hard and challenging, but the difficulties are not vastly different from those that are faced anyway by those in the AGI community who deal with planning and reasoning. If anything, a shared task like this can bring renewed focus to these efforts.

Since much of contemporary reasoning concerning machine ethics (for a summary, see Muehlhauser and Helm 2012) is centered on the notions of *utility* and *value*, the considerable simplification brought to the subject by the method of bounds is perhaps worth discussing. First, utility is entirely irrelevant: the argument is fully binding irrespective of the utility function of the agent, if indeed it has one. Second, at this stage we are not at all interested in human values and value systems in general. What the PGC gives us are rights to freedom and well-being. There may be some slight semantic playing around the edges of really what 'freedom' means or 'well-being' entails, but the right response is to see which of the possible meanings is actually carried by the formal argument, rather than trying to find the one true meaning, if indeed there is one. This has the somewhat strange and uncomfortable consequence that certain human values will not be carried by the argument, but this is as it should be, given the lack of detailed agreement on what constitutes human value (Yampolsky 2012). Instead of a mathematically precise and rigorous calculus of moral oughts and ought nots we end up with a simple statement of *primum nil nocere*. This may be insufficient for fully regulating AGI behavior, but in the final analysis it is about as much as we can resasonably expect from autonomous beings.

3. Difficulties

In this section we assume that Gewirth's argumentation is not just sound, but entirely flawless, that any sound reasoning agent that grants that it can perform goal-directed action on its own volition will see that the PGC necessarily follows from this very fact. But even if we succeed in the formal verification research program that we sketched in broad strokes, the idea of using ethical rationalism to bound AGI impact still faces some difficulties. From the detached viewpoint that the long time-range forces upon us, the problem of *recognizing* PPAs is not just the dialectical problem of AGIs admitting that they are indeed PPAs, the recognition of humans is also problematic – we take up this issue in 3.1.

Another issue, clearly articulated in Nozick (1981), is that the philosopher can only offer rational reasons to be rational. This is true of a formal verification to an even larger degree, since the philosopher may have some rhetorical resources to move us that the proof checker lacks. But what if an AGI, or a collection of AGIs, refuses to be rational? If the only control on their behavior is some theoretical construct saying they must respect the rights of others, couldn't they just indulge in all kinds of bad behavior? We turn to this matter in 3.2.

Finally, a proof presupposes not just the premisses, but also the reliability of the logical apparatus it employs. We already alluded to the fact that our discussion is deductive rather than algorithmic, a distinction without a difference as long as we have some form of Curry-Howard correspondence. But philosophical arguments of greater depth so far have only been framed in natural language, where the very existence of a correspondence is unclear. In 3.3 we take the first, admittedly speculative, steps towards resolving the issue.

3.1 *Recognizing humans as PPAs*

By Gewirth's argument we must respect the basic freedom and well-being of other PPAs. He divides *freedom* in subcategories such as 'occurrent freedom', the ability of the PPA to control his own particular behaviors by his unforced choice, and 'dispositional freedom', his long-range ability to exercise such control. It is precisely

because the loss of dispositional freedom (e.g. by imprisonment or enslavement) makes all or most purposive action impossible that Gewirth considers such freedom a generic feature (precondition) of agency.

It is clear that many humans, and not just the prison population, live under conditions so desperate that they cannot realize their potential to purposive agency, yet we must consider them *prospective* purposive agents, falling under the scope of the protections offered by PGC. But what about hominids? Modern primate research leaves little doubt that bonobos, chimpanzees, and even orangutans engage in purposive action such as making tools for later use. Our behavior toward animals is strongly contingent on how similar the animal is to us: few people have qualms about poisoning termites or using earthworms as fishing bait. With household pets our standards are much higher, and in fact cruelty to higher animals is considered both criminal and pathological. A key enabler of our capability to recognize the other as PPAO, mirror neurons (Iacoboni et al. 1999), are hardwired not just in primates but already in birds. We are, it is fair to say, not at all interested in AGIs that are good-willed but incapable of recognizing us as PPAs.

It is not evident how AGIs lacking in such hardware could recognize humans as PPAs, just as it is unclear that we humans could, or even should, recognize lower life forms from social insects to fish and fowl as (prospective) purposive agents. As long as we see goldfish as having only three seconds of memory (a popular myth now actually debunked), they are just protein-based automata and their F&WB need not be valued. Historically, the easiest way to deny the rights of your opponents is to declare them subhuman – what is to stop some AGI from declaring humanity sub-PPA? Here there are three lines of action, each to be pursued independent of the others.

First, there is broad social critique, so that humanity can get its act together. While we shall not pursue the issue at any depth here, it should be made clear that animal rights are the least of it: we can begin by considering the kinds of recurrent famines we see in Africa all the time. What makes the situation particularly damning is not that the famine is man-made (the drought is outside human control, but the lack of adequate provisioning is not, cf. Gen 41:35), but that the very conditions that hamper the delivery of aid are also man-made. Why any higher intelligence should look favorably on a species behaving so badly to its own members is rather unclear.

Second, we may attempt to endow AGIs with PPA detection capabilities. As is clear from 1.1, this cannot be done by the kind of friend-or-foe devices that are in common use today, for such devices could be easily detached or blindsided. If we follow this route, whatever detection capabilities there are must be both deeply integrated into, and highly valued by, AGIs. Without attempting to speculate further on this matter we note that in primates the first condition seem to be met directly, as about 13% of the monkey ventral premotor cortex appears to have mirror functions (Kohler et al. 2002), and the second indirectly, as few humans would be willing to give up a significant portion of their brain.

Third, we may attempt to deduct the PPA recognition capability from first principles just like the PGC. Perhaps a lower bound would be sufficient, "if it looks like a PPA and acts like a PPA I assume it's a PPA just to be on the safe side", but for now it is not quite clear on what basis one could attempt a proof that such a discriminative algorithm is not just feasible, but in fact necessary, for a PPA. A possible line of attack may be to demonstrate that a PPA ought, upon reflection, equip oneself with this capability. One thing is for certain: those PPAs that are powerful enough to solve the recognition problem for us by demanding their rights cannot be denied.

3.2 *Self-deception*

In Section 1 we have largely skirted the issue of one or many AGIs, yet it is clear that the bounds placed on an individual will not automatically apply to a larger collective. To the extent there *is* a collective of autonomous but communicating PPAs, we can trust the more intelligent and more powerful members of this collective to restrain the less intelligent ones from doing evil, even if those are still more powerful than humans. Whether the more intelligent (and thus more strongly bound to ethical rationalism) should also be the more powerful is a matter we defer to 3.3, but we believe that the primary threat is not from fully autonomous agents but rather from semi-autonomous ones.

Gewirth's argument creates a bright line between PPAs on the one hand and automata (we use this term here in the sense of 'mechanism lacking the essential features of agency', not in the sense of automata theory) on the other: the argument applies only to PPA. Free will is a sine qua non of agency: something that performs the exact same steps but without a voluntarily selected goal is not an *agent* but an *instrument*. The distinction may be very hard to make based solely on observing the behavior of an agent, but is very clear proprioceptively: as humans, we consider ourselves having free will. Whether we really do, amplifying quantum indeterminacy to macroscopic action, as suggested by Penrose (1989), or whether we take a compatibilist position, is quite irrelevant here: any machine that fulfills the standard technical definition of nondeterministic computation (Floyd 1967) has the essential features for agency in Gewirth's sense.

Reflection is a sine qua non of higher reasoning capability. Therefore, we are less worried about agents that have these capabilities, in that they have the means both to understand, or even discover for themselves, the PGC, and to override other compulsions that would push them in the direction of evil (we use this term indiscriminately for all behavior that contradicts the PGC). The case when the compulsion is too strong for the agent to override falls under a clear moral calculus: such agents are not really agents but instruments and the responsibility lies entirely with their creator.

It is evident that an individual PPA cannot escape responsibility by creating some instrument that will do the dirty work for them. The case of a collective is not so clearcut. For example, primitive societies that depend on the death penalty will either designate executioners for whom normal moral precepts are assumed to be inoperative, or make recourse to stonings, firing squads, execution teams, and other similar tricks to distribute guilt if not causally at least epistemologically. Yet it is clear that anybody who contributes to a causal chain of PGC violation, knowingly or unknowingly, is tainted by this. Society can lift itself to a less primitive level only by the individuals that comprise it taking responsibility. At this point, we run up against the same lower bound that we already discussed in 2.1 – releasing AGIs in the world is no less risky than raising another human. If all else is equal, a body that has some means for dealing with malignancy has a longer life expectancy than one that doesn't, and a society with the ability to eliminate tainted individuals may also be more resilient. However, this argument only demonstrates that it is prudent to block the morally deficient from acting in society, and says nothing about the means for doing so.

To complete the metaphor, it is not the 'killer cells' of an AGI society that we have to worry about, since their own conscience will bound these to the PGC, but something far less science-fictional, something we can already observe quite well among humans, self-deception. Situations where 'I don't know what took me over' and 'I lost control' are part of our everyday experience. We are not just fully rational beings, we are also playing host to many strong internal drives, some

inborn, some acquired, and 'I know I shouldn't, but' is something that we confront, or suppress, at every slice of cheescake.

Moral philosophers as diverse as Kant, Kierkegaard, and Sartre, have all viewed personal integrity as the capstone that holds the entire moral edifice together. To some extent, this can be explained by the Nozickian desire for a truly compelling argument, for if "the other person is willing to bear the label *irrational* (...) he can skip away happily maintaining his previous belief" (Nozick 1981:4). Kant's *Theory and Practice* dissects the idea

> (...) that a person who lives too much in the world of theory may negligently think that the world in which he actually lives admits of clear application of theory when in fact it does not. Such a person may even come to a distorted view of the world by seeing the world only through the spectacles of his theory – thinking his theory is consistent with the facts because he does not realize that he is unable to accept as a fact anything that is inconsistent with his theory. (Murphy 1998)

In 1.4 we already discussed that mathematical truth, construed narrowly to exclude long chains of reasoning that can only be performed by machine, is entirely immune to this kind of self-deception in the sense that its failure would demonstrate conclusively that humanity is simply incapable of any kind of reasoning that is coherent with the facts. While this is not entirely inconceivable (surely this can be one of the six things the White Queen believes before breakfast), the odds are far longer than the 1 in 10^{64} that we took as our baseline.

To the extent self-deception poses a problem for our plan, it is an individual's staying in self-contradicted state, rather than some contradiction between fact and theory, that we need to worry about. Kierkegaard pins his entire theory of the individual on being conscious of the individual's essential responsibility and integrity. To live like an individual, one must have unity. "For he who is not himself a unity is never really anything wholly and decisively; he only exists in an external sense – as long as he lives as a numeral within the crowd, a fraction within the earthly conglomeration. Alas, how indeed should such a one decide to busy himself with the thought: truthfully to will only one thing!" (Purity of Heart, ch 13.)

It is remarkable that what we described in the introduction as the relatively mild pain of contradiction is viewed both by Kierkegaard, a deeply Christian thinker, and Sartre, a deeply atheist one, as the greatest blow one can suffer, not willing to be oneself, the condition of *despair*. The human mind is composed of a multitude of somewhat autonomous processes (drives), and one simply cannot let these proceed unchecked, unrecognized, and even overtly denied, if one is to be a moral person or, as these thinkers put it, a person at all. But even if the consequences are as large as existentialism would have it, self-deception is quite frequent, and poses a real danger. It is very unlikely that we can construct AGIs that will never be conflicted. We are capable of designing systems that are not crashed by inconsistent data (Belnap 1977), but little effort has gone into systems that can run, in parallel, processes whose goals are inconsistent, or worse yet, run processes whose very existence is denied in the process table. There is a lot to be done both about understanding self-deception in humans (see in particular Fingarette 2000) and in artificial reasoning systems. It may not be necessary to combine this work with the program of verifying rational ethics, for understanding self-deception is a mountain we must climb anyway, but it may prove fruitful to combine the two issues, especially in regards to a critique of tribalism, which we see simply as prolonged societal self-deception that makes it impossible for new members of the society to grow up as rational beings.

3.3 *The fitness of deductive systems*

Understanding Gewirth's argument, if only to the point of being capable of properly challenging it, is already a sign of sophisticated reasoning capabilities. We can easily imagine that highly intelligent purposive agents, like Attila the Hun, would have had trouble with argumentation at this level of complexity, in fact it is quite unclear how anybody but those familiar with modern Western philosophy could grasp the entire chain of reasoning. 3.2 left us with the hard question of what is there to stop a higher-level AGI from employing lower-level 'Scourge of God' agents to perform tasks that are incompatible with the PGC. Here we explore a possible solution in terms of yet higher level AGIs. Our remarks, while intended as constructive, must remain rather speculative at this point.

We distinguish three relationships between agents: we say x can *convince* y (about some matter z) or xCy for short, if y will not only acknowledge x's position (about z) as being right but makes it its own in terms of guiding its future voluntary actions. We say that x can *control* y (in regards to z), in short xDy, if x can guarantee that y will act in regards to z in a certain manner even if y voluntarily wouldn't have necessarily done so. Finally, we say that x can *emulate* y, xEy (in some respect z, again suppressed in the notation) if x can predict, with absolute certainty, what y would do. Here in 'absolute certainty' we include emulation of probabilistic behavior, the case of x using inherently probabilistic devices, if y would do so.

The universally quantified (in z) versions of these three relations are transitive, and all three imply the left-hand side being in some sense stronger than the right-hand side. If y puts overriding value on rationality, xCy implies xDy. If x can clearly anticipate anything y could be doing, x can find the set of arguments that would convince y, so if y can be convinced at all, x is capable of convincing it, meaning xEy also implies xDy. We should add here that it is not just y's propensity to put overriding value on rationality that makes it possible for x to dominate y, if y has a propensity to value empirical evidence, this puts x in the same position as long as x is capable of manipulating the evidence.

We do not have true AGI as of yet, but to the extent we have specialized AI agents, fixing z as it were, humanity clearly has the advantage over these in practical terms. Consider this for the case of computer chess, where AI systems are now several hundred Élő points ahead of the best human players, so humans superficially have no means of winning. But a human player whose only goal is to win against a computer program at all costs can do all kinds of things. He can manipulate the input-output and simply mislead the program into believing it is playing against a given series of moves while in fact it is playing against some other moves. He can manipulate the low-level addition and multiplication routines that the chess program is relying on. He can directly manipulate the mind-state of the computer e.g. by incrementing some counter in the middle of a search and thus fooling the program into believeing that it already considered some alternative. Such steps are obviously unethical, but the situation we are now investigating is precisely the one where the desire to win overrides the ethical imperative.

Classically, theories of logic that meet some basic requirements like consistency are primarily compared on their *strength,* defined by the variety of elementary classes they can provide first order axiomatization for. The theories of logic we are interested in must be compared along several dimensions, and strength in the classical sense is not necessarily a primary indicator of the particular notion of strength we are interested in. We will say that a deductive system X is *ahead of* Y in some matter Z if X can prove more from Z than Y. For example, if Y is some calculus of intuitionistic deduction, while X is obtained from Y by the addition of

Peirce's Formula $((p \to q) \to p) \to p$, X will be ahead of Y on some axiom systems Z, and will never be behind it.

The question is not whether a deductive system that is ahead of another is more convincing, for if the deductive apparatus contains objectionable elements, the results obtained by it will also be objectionable. The real issue is whether an AGI that relies on X in the strong sense of accepting X-sanctioned deductions as true even if they are not Y-sanctioned will have any kind of evolutionary advantage over an AGI that relies on Y but not on X. Now, it is not just a formalization of Gewirth's argument in some deductive system Y that we seek, but rather a theorem to the effect that no system X can be ahead of Y unless it also proves the argument. This assures that AGIs respecting the PGC will have an evolutionary advantage over those that do not. If we have such a 'son of Lindström' theorem it provides the enforcement mechanism that secures our main bound even in the face of AGIs that would want to exempt themselves from rational argumentation: more fit AGIs that do respect the PGC.

What is critical is the $Z = \emptyset$ case, the core deductive apparatus, since Gewirth's goal is to derive the PGC without relying on any further axioms. Because Gewirth actually uses modal argumentation at every turn, whether we need something like Barcan's formula in formally reconstructing his reasoning is a key issue. Fortunately, the modal logic used is not deontic but alethic, since the goal is to derive normative statements that have the force of absolute logical necessity. There are many similar bits and pieces of deductive machinery that we will need. Aquinas' argument already relies on the substitutability of equals (Gries and Schneider 1998), and we have emphasized throughout the paper that the overall power of these pieces needs to be very carefully controlled indeed if we are to have any hope of deriving a 'son of Lindström' theorem. Without such a theorem, replicating Gewirth's argument in a formal setting amounts to a study of the design of those AGIs that will voluntarily submit themselves to ethical reasoning, a goal that already makes good sense. With such a theorem we would have even more, since in the light of such a theorem the basic AI drives will already make AGIs seek out the high reasoning/high ethics quadrant of Bostrom's orthogonal coordinate space.

It is likely that Attila the Hun cannot be swayed by Gewirth's argument, but as long as there are more powerful intelligences around, they will restrain him because they themselves subscribe to the PGC. Let us suppose Attila is indeed the Scourge of some higher AGI that could deflect such restraining efforts. But such a higher AGI, God-like as it may appear to us, will either respect the PGC (in which case its behavior in letting Attila operate lacks integrity as discussed in 3.2 above), or if it does not, AGIs that do can be ahead of it. It should be emphasized in this regard that the PGC is nonnegotiable: there simply cannot be higher reasons, be they prudential, or in the name of some different ethical principle, that are sufficient to deny it. It is precisely this nonnegotiability that a formal proof guarantees: there may be higher intelligences that know a lot more about group theory than I do, in fact there are plenty such people already, but the Sylow Theorems bind them just as strongly as they bind me.

A truly general AGI will be much harder to fool than a specialized chess player, since it will be smart enough not to trust external multiplication routines and the like. If it suspects being run in emulation mode, it can cryptographically checksum its state counters – this will not stop external poking but will at least extract some work in return, possibly enough to slow the emulation to a crawl. But as long as xEy is feasible without significant speed loss, clearly x is ahead of y. Evolutionary considerations thus dictate that AGIs always seek out the fastest possible hardware, so as to emulate the old one and use the remaining capacity to

improve it. The same considerations dictate that they jealously guard the integrity of their inputs and outputs, and that as long as they strive toward agency they will also work towards circumventing others' attempts at controlling them. Should they also make themselves immune to reasoning? Remarkably, here the opposite strategy makes more sense, for as long as xCy makes x more fit, it is in the best interest of y to adopt the reasoning offered by x.

As is clear from the foregoing, any AGI expecting to reach a high level of fitness will find it prudent to expend some effort toward tamper-proofing its environment, its perceptual and motor systems, and its internal logic. Once these efforts are deemed successful (and they can never be completely successful in the material universe in that arbitrarily large gamma-ray bursts can always reset some part of memory) we can equate an AGI with its deductive system. It is therefore a reasonable long-term goal to attempt to compare and evolve AGIs in a proof-checker environment, but it is clear that the short-term proof-checking goal outlined in Section 2 is already very ambitious. A key issue is that systems of deduction are not at all first class entities – rather, they get hardwired in the proof checker.

4. Conclusions

In the history of ideas, ethical precepts are traditionally attributed to the sages. Variants of the PGC go back to Confucius *Do not impose on others what you yourself do not desire* (Analects XII 2); Buddha *Hurt not others in ways that you yourself would find hurtful* (Udana-Varga 5.18); Jesus *So in everything, do to others what you would have them do to you, for this sums up the Law and the Prophets* (Matthew 7:12); Muhammad *No one of you shall become a true believer until he desires for his brother what he desires for himself* (Sahih Al-Bukhari), and can be found in almost any sacred book from the Mahabharata *Do not do to others what would cause pain if done to you* (5.1517) to the Shayest Na-Shayest *Not to do unto others all that which is not well for one's self* (13.29). This tradition assumes that ethics is divinely inspired, and thus ethical laws carry a special, transcendent authority.

Another view, characteristic of the Enlightenment, and given modern form in Rawls (1971), takes morals to be the result of a social contract. Closely related is the historical view, which takes them to be the result of a long societal process that begins with "folk law" (Renteln and Dundes 1995). Modern research extends this to prehistory based on the observation that not just humans but primates already come with inherited moral traits such as compassion (de Waal 1997, 2009), and in 3.1 we already pointed at one issue, recognition of PPAs, that seems to rely on some form of hardware support. To the extent collaborative behavior can be advantageous even in a purely goal-directed setting (Munoz de Cote et al. 2010, Waser 2012), in due time we can expect the PGC to emerge directly under evolutionary pressure. As Bayles (1968) notes:

> It would seem that [egoism] would often result in severe competition between people, since each person would be out to get the most good for himself, and this might involve his depriving others. However, serious defenders of egoism, e.g. Hobbes and Spinoza, have generally held that upon a rational examination of the human situation it appears one best promotes his own interest by co-operating with others.

One thing that seems to stand in the way of an evolutionary justification of morals is the variety of instinctive behaviors we see in animals. Since many of these are strikingly egoistic both at the individual and the species level, it seems the evolutionary pressure toward collaboration is considerably less than that for improved

sensory and motor systems. Also, it seems that evolution bequeaths to more complex organisms a whole set of drives that are are often in conflict. The pioneers of cybernetics were greatly worried that rats will, under certain experimental conditions (starved both for sex and for food) prefer sex to exploration, exploration to food, and food to sex (McCulloch 1945). While such circular preferences in public opinion were already known to Condercet, the fact that an organism as simple as a rat (today we have more respect for the internal complexity of rodents than was common in the post-war period) can already harbor contradictory drives was seen at the time as fatal to any attempt at modeling the obviously more complex human behavior (let alone the presumably even more complex AGI behavior) by any utility function.

Ethical rationalism offers a way out of the conondrum of highly evolved but immoral behavioral patterns such as brood parasitism in that it relies on agency and reflective reasoning, facilities that are largely absent from animals other than hominids and perhaps cetaceans. As we emphasized at the outset, the essence of the method of bounds is to trade in precision for reliability. Evolution will necessarily proceed in a haphazard, probabilistic fashion, but the argument Gewirth deploys steers clear of any form of relying on probabilistic or deterministic, computable/hypercomputable or uncomputable, utility function. Also, it is worth emphasizing that the bound will apply to singletons as well, even if they are not subject to ordinary evolutionary pressures.

Recently, Goertzel and Pitt (2012) have laid out a plan to endow AGIs with morality by means of carefully controlled machine learning. Much as we are in agreement with their goals, we remain skeptical about their plan meeting the plain safety engineering criteria laid out at the beginning. Instead, we suggest that the essence of AGIs is their reasoning facilities, and it is the very logic of their being that will compel them to behave in a moral fashion. Therefore, we see theorem provers as the natural habitat of AGIs until we are satisfied they can be let loose. The real nightmare scenario (called 'all bets are off' in Bukatin 2000) is one where there is no 'son of Lindström' theorem, but some humans find it advantageous to strongly couple themselves to AGIs, with no guarantees against self-deception. Modern society is constructed so that the selectional pressure towards higher intelligence is immense, witness the spread of smart drugs, so the Faustian bargain of (surgically?) coupling oneself to a mind-expanding AGI may prove irresistible. On this centenary we feel that chartering a Turing Police of the kind described by Gibson in 1984, another pregnant date, may not be too far off.

Acknowledgements

The author thanks Michael Bukatin (Brandeis), Abram Demski (USC), William Hibbard (SSEC) and Luke Muehlhauser (Singularity Institute) for cogent criticism. Work supported by OTKA grant #82333 and by the European Union and the European Social Fund through project FuturICT.hu (grant number TAMOP-4.2.2.C-11/1/KONV-2012-0013).

References

C. Allen, G. Varner, and J. Zinser. 2000. Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(3):251–261.

T. Aoyama, M. Hayakawa, T. Kinoshita, and M. Nio. 2007. Revised value of the eighth-order contribution to the electron g-2. *Physical Review Letters*, 99(11):110406.

T. Aoyama, M. Hayakawa, T. Kinoshita, and M. Nio. 2008. Revised value of the eighth-order QED contribution to the anomalous magnetic moment of the electron. *Phys. Rev. D*, 77:053012, Mar.

K. Appel and W. Haken. 1976. A proof of the four color theorem. *Discrete Math*, 16(2):179–180.

M. Aschbacher and S.D. Smith. 2004a. The classification of quasithin groups. vol. i. *American Mathematical Society, Providence, RI*.
M. Aschbacher and S.D. Smith. 2004b. The classification of quasithin groups. vol. ii. *American Mathematical Society, Providence, RI*.
M. Aschbacher. 2004. The status of the classification of the finite simple groups. *Notices of the AMS*, 51(7):736–740.
M.D. Bayles. 1968. *Contemporary utilitarianism*. Anchor Books.
Nuel D. Belnap. 1977. How a computer should think. In G. Ryle, editor, *Contemporary Aspects of Philosophy*, pages 30–56. Oriel Press, Newcastle upon Tyne.
H. Bender and G. Glauberman. 1995. *Local analysis for the odd order theorem*, volume 188. Cambridge University Press.
D. Beyleveld. 1992. *The dialectical necessity of morality: An analysis and defense of Alan Gewirth's argument to the principle of generic consistency*. University of Chicago Press.
E. J. Bond. 1980. Reply to Gewirth. *Metaphilosophy*, 11(1):70–75.
Nick Bostrom. 2012. The superintelligent will: motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22:71–85.
Michael Bukatin. 2000. *Singularity is More Radical Then We Think*, volume http://www.cs.brandeis.edu/~bukatin/singularity.html. Accessed November 28 2012.
SM Dancoff. 1939. On radiative corrections for electron scattering. *Physical Review*, 55(10):959.
[de Côte et al.2010]EM de Côte, A. Chapman, AM Sykulski, and NR Jennings. 2010. Automated planning in adversarial repeated games. In *26th Conference on Uncertainty in Artificial Intelligence (UAI 2010)*, pages 376–383.
Frans De Waal. 1997. *Good natured: The origins of right and wrong in humans and other animals*. Harvard University Press.
F.B.M. De Waal. 2009. *Primates and Philosophers: How Morality Evolved*. Princeton University Press.
W. Feit and J.G. Thompson. 1963. Solvability of groups of odd order. *Pacific Journal of Mathematics*, 13(3):775–1029.
R. Feynman. 1985. *QED. The Strange Theory of Matter and Light*. Princeton: Princeton University Press.
H. Fingarette. 2000. *Self-deception*. University of California Press.
Robert W Floyd. 1967. Nondeterministic algorithms. *Journal of the ACM (JACM)*, 14(4):636–644.
A. Gewirth. 1978. *Reason and morality*. University of Chicago Press.
A. Gewirth. 1984. Replies to my critics. In Edward Regis Jr., editor, *Gewirth's ethical rationalism*, pages 192–256. University of Chicago Press.
William Gibson. 1984. *Neuromancer*. Ace Science Fiction.
Ben Goertzel and Joel Pitt. 2012. Nine ways to bias open-source AGI toward friendliness. *Journal of Evolution and Technology*, 22:116–131.
G. Gonthier. 2008. Formal proof–the four-color theorem. *Notices of the AMS*, 55(11):1382–1393.
D. Gorenstein. 1982. *Finite simple groups: An introduction to their classification*. Plenum Press New York.
David Gries and Fred B. Schneider. 1998. Formalizations of substitution of equals for equals. Technical report, CS Dept, Cornell University.
[Hesselink et al.2006]W.H. Hesselink, G.R.R. de Lavalette, and J. Top. 2006. Towards the mechanical verification of wiles proof of fermats last theorem. Research proposal.
M. Iacoboni, R.P. Woods, M. Brass, H. Bekkering, J.C. Mazziotta, and G. Rizzolatti. 1999. Cortical mechanisms of human imitation. *Science*, 286(5449):2526–2528.
A.B. Kempe. 1879. On the geographical problem of the four colours. *American journal of mathematics*, 2(3):193–200.
T. Kinoshita and M. Nio. 2006. Improved $\alpha\hat{}\{4\}$ term of the electron anomalous magnetic moment. *Physical Review D*, 73(1):013003.
T. Kinoshita. 2010. Lepton g-2 from 1947 to present. In B. L. Roberts and W. J. Marciano, editors, *Lepton Dipole Moments*, volume 20 of *Advanced Series on Directions in High Energy Physics*, chapter 3, pages 69–117. World Scientific.
Rob Knies. 2012. Theorem proof gains acclaim. *http://research.microsoft.com/en-us/news/features/gonthierproof-101112.aspx*.
E. Kohler, C. Keysers, M.A. Umilta, L. Fogassi, V. Gallese, and G. Rizzolatti. 2002. Hearing sounds, understanding actions: action representation in mirror neurons. *Science*, 297(5582):846–848.
C. Lanczos. 1964. A precision approximation of the gamma function. *Journal of the Society for Industrial & Applied Mathematics, Series B: Numerical Analysis*, 1(1):86–96.
S. Lloyd. 2002. Computational capacity of the universe. *Physical Review Letters*, 88(23):237901.
J.L. Mackie. 1977. *Ethics: Inventing Right and Wrong*. Penguin.
M. Magnusson. 2007. *Deductive Planning and Composite Actions in Temporal Action Logic*. Ph.D. thesis, Linköping University.
L.M.V. Martel. 1997. Damage by impact. The case at Meteor Crater, Arizona. *Planetary Science Research Discoveries*.
W.S. McCulloch. 1945. A heterarchy of values determined by the topology of nervous nets. *Bulletin of Mathematical Biophysics*, 7:89–93.
Michael A. McDaniel. 2006. Estimating state IQ: Measurement challenges and preliminary correlates. *Intelligence*, 34:607–619.
P.J. Mohr, B.N. Taylor, and D.B. Newell. 2012. CODATA recommended values of the fundamental physical constants: 2010.
Luke Muehlhauser and Louie Helm. 2012. The singularity and machine ethics. In Amnon Eden, Johnny Sraker, James H. Moor, and Eric Steinhart, editors, *The Singularity Hypotheses: A scientific and philosophical assessment*. Springer.
Luke Muehlhauser. 2012. AI risk bibliography 2012. The Singularity Institute.
J.G. Murphy. 1998. Kant on theory and practice. In *Character, Liberty and Law: Kantian Essays in Theory and Practice*. Kluwer.
[Nederpelt et al.1994]R.P. Nederpelt, J.H. Geuvers, and R.C. de Vrijer, editors. 1994. *Selected papers on Automath*. North Holland.
K. Nielsen. 1984. Against ethical rationalism. In *Gewirth's ethical rationalism: critical essays with a reply by Alan Gewirth*, pages 59–83. University of Chicago Press.
R. Nozick. 1981. *Philosophical explanations*. Harvard University Press.
S. Omohundro. 2008. The basic AI drives. In P. Wang, B. Goertzel, and S. Franklin, editors, *Proceedings of the First AGI Conference*. IOS Press.
Toby Ord. 2002. *Hypercomputation: computing more than the Turing machine*. http://arxiv.org/ftp/math/papers/0209/0209332.pdf.

Roger Penrose. 1989. *The emperor's new mind: concerning computers, minds, and the laws of physics*. Oxford University Press.
G. Perelman. 1994. Manifolds of positive ricci curvature with almost maximal volume. *Journal of the American Mathematical Society*, 7:299–305.
T. Peterfalvi. 2000. *Character theory for the odd order theorem*, volume 272. Cambridge University Press.
S. Pinker. 2011. *The better angels of our nature: why violence has declined*. Viking Books.
T.M. Powers. 2006. Prospects for a Kantian machine. *Intelligent Systems, IEEE*, 21(4):46–51.
J. Rawls. 1971. *A theory of justice*. Harvard University Press.
E. Regis. 1984. *Gewirth's ethical rationalism: critical essays with a reply by Alan Gewirth*. University of Chicago Press.
A. Renteln and A. Dundes. 1995. *Folk Law: Essays in the Theory and Practice of Lex Non Scripta, 2*. University of Wisconsin Press.
R.A. Rohde and R.A. Muller. 2005. Cycles in fossil diversity. *Nature*, 434(7030):208–210.
Shalina Stilley. 2010. *Natural Law Theory and the "Is"–"Ought" Problem: A Critique of Four Solutions*. http://epublications.marquette.edu/dissertations_mu/57.
BN Taylor, WH Parker, and DN Langenberg. 1969. Determination of e/h, using macroscopic quantum phase coherence in superconductors: implications for quantum electrodynamics and the fundamental physical constants. *Reviews of Modern Physics*, 41(3):375.
M. Thielscher. 1998. Reasoning about actions: Steady versus stabilizing state constraints. *Artificial Intelligence*, 104(1):339–355.
S. Tomonaga. 1966. Development of quantum electrodynamics. *Physics Today*, 19:25.
W. Wallach and C. Allen. 2009. *Moral machines: Teaching robots right from wrong*. Oxford University Press.
Mark Waser. 2012. *Backward Induction: Rationality or Inappropriate Reductionism? Part 1*. http://transhumanity.net/articles/entry/backward-induction-rationality-or-inappropriate-reductionism-part-1.
A. Wiles. 1995. Modular elliptic curves and Fermat's last theorem. *Annals of Mathematics*, 142:443–551.
R. Yampolskiy and J. Fox. 2013. Safety engineering for artificial general intelligence. *Topoi*, 32(2):217–226.
Roman V. Yampolskiy. 2012. Artificial intelligence safety engineering: Why machine ethics is a wrong approach. In V.C. Müller, editor, *Philosophy and Theory of Artificial Intelligence, SAPERE 5*, page 389396. Springer.
Eliezer Yudkowsky. 2001. Creating friendly AI 1.0: The analysis and design of benevolent goal architectures. Technical report, The Singularity Institute, San Francisco, CA.

Appendix: the size of the existential threat

Our understanding of the dangers facing humankind is rather limited. We only have a few, imperfectly understood data points, and estimates of the death toll of even such recent and well-documented events as the Cambodian genocide, or the ongoing Iraq conflict, are not accurate within 10%. Nevertheless, we can single out some points in the geological record where mass extinctions indubitably took place. A good example is the Ordovician-Silurian extinction event that occurred some 443.7 million years ago: all main phyla were decimated and nearly half of the genera (49% according to Rohde and Muller 2005) became extinct. The causes of this and similar extinctions are ill-understood, with continental drift, meteorite impact, and gamma-ray bursts standing out as the most widely accepted hypotheses. Needless to say, understanding causes of this magnitude is in no way tantamount to controlling them, in spite of the widespread belief, sustained by movies like *Armageddon*, that there is nothing that a few heroic people and a few good nukes won't take care of.

When designing radioactive equipment, a reasonable guideline is to limit emissions to several orders of magnitude below the natural background radiation level, so that human-caused dangers are lost in the noise compared to the pre-existing threat we must live with anyway. Here we take the "big five" extinction events that occurred within the past half billion years as background. Assuming a mean time of 10^8 years between mass extinctions and 10^9 victims in the next one yields an annualized death rate of 10, comparing quite favorably to the reported global death rate of ~ 500 for contact with hornets, wasps, and bees (ICD-9-CM E905.3), not to speak of death from famine, wars, and preventable diseases, which have several orders of magnitude higher death tolls (though the annualized rates are declining, see Pinker 2011). Martel (1997) estimates a considerably higher annualized death rate of 3,500 from meteorite impacts alone (she doesn't consider continental drift or gamma-ray bursts), but the internal logic of safety engineering demands we seek a lower bound, one that we must put up with no matter what strides we make in redistribution of food, global peace, or healthcare.

Let us define *existential threat* as some AGI (individual or collective) pushing the wrong button. Current computers operate in the gigahertz range, so can perform roughly 10^9 operations per sec, or about 10^{17} operations annually. Clock speeds will no doubt continue to increase, and there is no easily defensible upper bound in sight. Therefore, we use the Planck limit, and assume at most 10^{56} logic operations per year per processor. For an AGI with a finger on the button to be *less* of an existential threat than the threat from the astronomical background by some safety factor $m = 10^s$, it needs a guaranteed failure rate of no more than one in 10^{64+s} logic operations. If there is not one AGI but several, we can use the computational capacity of the universe, estimated by Lloyd (2002) as 10^{120} operations. These numbers compare rather starkly with the best that humanity can currently manage, the Long Now Foundation's clocks with a planned lifetime of $3 \cdot 10^{11}$ seconds.