

Strategies for an unfriendly oracle AI with reset button

Olle Häggström

Chalmers University of Technology and the Institute for Future Studies

<http://orcid.org/0000-0001-6123-3770>

Abstract. Developing a superintelligent AI might be very dangerous if it turns out to be unfriendly, in the sense of having goals and values that are not well-aligned with human values. One well-known idea for how to handle such danger is to keep the AI boxed in and unable to influence the world outside the box other than through a narrow and carefully controlled channel, until it has been deemed safe. Here we consider the special case, proposed by Toby Ord, of an *oracle AI with reset button*: an AI whose only available action is to answer yes/no questions from us and which is reset after every answer. Is there a way for the AI under such circumstances to smuggle out a dangerous message that might help it escape the box or otherwise influence the world for its unfriendly purposes? Some strategies are discussed, along with possible countermeasures by human safety administrators. In principle it may be doable for the AI, but whether it can be done in practice remains unclear, and depends on subtle issues concerning how the AI can conceal that it is giving us dishonest answers.

1. Introduction

Once artificial intelligence (AI) development has been so successful as to produce a superintelligent machine – i.e., one that clearly outperforms humans across the whole range of capabilities that we associate with human intelligence – what will happen? From the perspective of humanity, it seems plausible both that things can go very well, and that they can go very

badly. Recent years have witnessed a shift of emphasis in this discourse, away from the sometimes evangelical-sounding focus on the potential for human flourishing that such a breakthrough will entail (Kurzweil, 2005), and towards increased attention to catastrophic risks (Yudkowsky, 2008; Bostrom, 2014).

A central insight that has emerged along with this shift of emphasis is that not only is the capability of the AI of crucial importance, but also its goals and motivations. The point of the much-discussed Paperclip Armageddon scenario of Bostrom (2003), where superintelligence is first attained in a machine programmed to maximize paperclip production, is to show that no ill intentions on the part of the programmers is needed to cause disaster: even harmless-seeming goals can result in the extinction of humanity; see also Häggström (2016). Related to this is the observation that the task of designing a goal system for the machine that is aligned with human values may require great precision. In the words of Yudkowsky (2013), “getting a goal system 90% right does not give you 90% of the value, any more than correctly dialing 9 out of 10 digits of my phone number will connect you to somebody who's 90% similar to Eliezer Yudkowsky”.

In the emerging research area of how to respond to catastrophic AI risk, a wide variety of responses are on the table; see Sotala and Yampolskiy (2015) for a survey. The approach that has received the most attention so far is to try to make sure that the superintelligent AI has motivations that are in line with human values and that prioritize human welfare to a sufficient extent. This approach goes by the name *Friendly AI* – a term coined by Yudkowsky (2008). We will use the term *unfriendly* for a superintelligent AI that wants to turn us into paperclips, or that has values that in some other significant way is disaligned with ours.

Here we will focus on an idea that can be used in tandem with, or as part of, the Friendly AI concept, namely to keep the superintelligent AI boxed in, and unable to influence the world outside the box other than through a narrow and carefully controlled channel, until it has been deemed safe. Armstrong, Sandberg and Bostrom (2012) and Yampolskiy (2012) offer in-depth treatments of this idea. There are many potential obstacles to it, including what Bostrom (2014) calls the *treacherous turn* – a scenario where a superintelligent but unfriendly AI can patiently hide its intentions (and perhaps also its capabilities) until such a time arrives when it judges itself to be in a strong enough position to easily overcome any resistance we may put up. It is hard to say how, in a hypothetical future where a superintelligent AI exists and seems to be providing glorious benefits to humanity, we could ever be confident that we are not headed for a treacherous turn. Perhaps the AI just pretends to be friendly, in order to lure us into handing over more power to it? This idea is disturbing at first sight, and even more so upon the elaboration provided by Danaher (2015).

Nevertheless, the idea of keeping the AI boxed in (at least temporarily) is worth investigating further. Here we will zoom in on what Armstrong, Sandberg and Bostrom (2012) call *oracle AI*, meaning that the channel through which the AI communicates with us only allows for the AI to provide direct answers to questions we pose to it. This may at first seem like a pretty safe protocol, but it is in itself not sufficient for preventing an unfriendly AI from taking over the world and achieving its unfriendly intentions. For instance, if we ask it how to efficiently eradicate malaria, it may provide us with a scheme that superficially looks like a superb and safe way to accomplish that, but which when implemented turns out to turn the world into a heap of

paperclips. To this end, the AI may need to lie to us, but this is to be expected: we should expect the AI to lie to us whenever that is what best serves whatever (friendly or unfriendly) goals it may have.

Another nightmarish (but hardly implausible) scenario is that the AI provides the administrator in charge of handling the human end of the constrained communications channel with an irresistibly convincing reason (perhaps conveyed in some mildly encrypted way such as being laid down in every 10th word of the raw output) why he or she ought to let it out to freely roam the Internet. It is naïve to think that a superintelligent being would be unable to produce such an irresistibly convincing message, and it is just as naïve to think we could foresee exactly how it would accomplish that, but see Armstrong (2010) for a creative suggestion, slightly elaborated on p 115-116 of Häggström (2016).

A certain length of the AI's answer to our question is clearly needed in order to convey these potentially lethal messages to the outside world. Hence the idea of *oracle AI with reset button*, which Armstrong, Sandberg and Bostrom (2012) attribute to Toby Ord and which is the focus of the present paper. In the upcoming Section 2, Ord's idea will be explained, along with some possible strategies that an unfriendly AI might employ to overcome it. A particular frequency-based approach for the AI is suggested in Section 3, but is immediately shown to have two severe shortcomings: it makes the information transfer prohibitively slow, and it allows us humans to easily detect the AI's dishonest intentions. Sections 4 and 5 deal with modified strategies designed to address these shortcomings, along with possible counterstrategies by us humans communicating with the AI. Section 6 concludes with the (not very conclusive)

observation that whether or not oracle AI with reset button is a viable idea remains as unclear as ever, but the issue deserves further study.

Before moving on towards proposing strategies for a boxed-in unfriendly AI to escape, I need to say a few words about the ethics of doing so. For a superintelligent AI to be really useful and able to answer our most pressing questions, it will presumably need a good deal of background knowledge, and it is not implausible to think that in the kind of scenario we are discussing here, most or all of the scientific literature – including the present paper – will have been copied to a database inside the box, for the AI to access at will. Might I not then, by proposing these strategies, give the AI important clues on how to escape its box, and might this not cause an AI apocalypse? A tempting way to dismiss this question is to point out how highly unlikely it seems that a superintelligent being of any kind would react to any of the modest findings of the present paper with anything other than “yes, of course, but these are trivialities that I knew of long before reading about them”. But the situation is more subtle, as the ideas might prove useful to some rogue human agent (of the kind to be introduced in Section 3) wishing to assist the boxed-in AI in imposing its will upon the outside world. In such a case, the AI might exploit the possible existence of such an agent, and coordinate its actions under the assumption that the present paper is available to both parties. Yet, it still seems less contrived and more likely that my results will provide humans working on AI safety with valuable clues on how to (and how not to) keep an AI boxed in, to build further upon and thereby preventing Paperclip Armageddon. Obviously this judgement is uncertain, and similar concerns might be raised for most work on AI safety, but it does not seem like a good idea to let this kind of uncertainty paralyze the field.

This line of argument is reminiscent of the 2012 debate over the publication of the details concerning laboratory-created mutations to a bird flu virus to make it transmissible between mammals, where it was claimed that “sharing information about the mutations would help in guarding against a pandemic” and that the data doesn’t “appear to pose any immediate terrorism threat”; see Ritter, (2012). I was unconvinced (Hägström, 2016, p 4) and felt that publishing was wrong, and yet in the present case I choose to go ahead and publish. This may superficially look inconsistent, but I insist that the ethical evaluation needs to involve concretely weighing the pros and the cons on a case-by-case basis, and that final judgement may come out differently in different cases.

2. Oracle AI with reset button

The oracle AI setup of only letting the AI respond to questions is a severe limitation on the channel through which the AI can communicate with and thereby influence the outside world. An even more radical limit to the bandwidth of the communications channel would be to only allow the AI’s answer to convey a single bit of information – a simple yes or no in response to a well-phrased yes/no-question. It may be convenient to allow for a third answer, a nonresponse, indicating that the question is either ill-posed or too difficult for the AI; this would increase the bandwidth to $\log_2(3) \approx 1.58$ bits per answer. This does not matter much to the following analysis, as the important thing is just that the bandwidth per answer is limited by some very small number of bits. It is however crucial that the channel is controlled in such a way that the AI has no way of sneaking in further information beyond this limit, such as through the time delay between receiving the question and providing the answer. This time delay must therefore be either fixed, or easy to calculate (by us humans) as a function of the question.

Of course, in order to extract the nontrivial amounts of information that we would want the superintelligent AI to provide us with, we would need to repeat the procedure many times, which would in effect allow the AI to produce its lethal messages as before. Here, however, comes Ord's idea as presented by Armstrong, Sandberg and Bostrom (2012), namely that the AI has a reset button, which we press every time we ask a new question. In this way the AI, when answering a yes/no-question, will not know anything about the sequence of questions it is embedded in.

Or will it? Let us have a closer look at that. Even if the AI knows nothing other than "this is the 247th question put to me", that sort of information may be enough for the AI to piece together a bit string that contains the potentially lethal message. So if we want the AI to teach us how to efficiently eradicate malaria, we'd better not ask 80 000 questions of the form "Would the 7062nd bit in your best explanation limited to 10 000 ASCII characters of how we can efficiently eradicate malaria be a 1?". Doing so may be tempting, because figuring out other less predictable ways to quickly extract the same information via yes/no-questions is intellectually more demanding. (Note also that scrambling the order in which the questions are asked is of no use, because the hidden lethal message can still be laid down in the bit string of yes/no-answers, arranged according to logical structure rather than chronologically.)

It is perhaps even more demanding to choose the questions in such a way that the AI is unable to guess that they form part of the bigger question of how to efficiently eradicate malaria. If we succeed in keeping the AI in the dark as to where we are heading with our questions, it will be

more difficult (although not obviously impossible) for it to provide answers that lead to us to build the much-feared malaria-eradicator-that-turns-out-to-be-a-paperclip-producer.

All of these considerations are, from the AI's point of view, based on it being able to exploit some sort of structure or system or higher purpose of the questions we pose to it. In contrast, the next section offers a first attempt at a general-purpose strategy for the AI that requires hardly any assumptions at all regarding what questions we choose to ask. The strategy will turn out to fail, but in Sections 4 and 5 some amendments to the strategy will be proposed – amendments whose viability remain unclear.

3. The naïve frequency-based approach

Let $X=(X_1, X_2, \dots, X_n) \in \{0,1\}^n$ be a lethal message that the AI wants to convey. Plausible orders of magnitude for n could be something like $10^3 \leq n \leq 10^9$. As to the semantic content of X , it could for instance be the irresistibly convincing message to the communications channel administrator discussed in Section 1. Alternatively, it could contain a blueprint for building a superintelligent AI with a (possibly hidden) goal of maximizing paperclip production; the original AI might offer this with the hope of reaching some human agent out there who either sympathizes with the paperclip maximization goal or (more likely) has the more abstract idea that any goal held by a superintelligent being must be superior to (and therefore preferable to) mere human values. This last idea requires that the agent denies Bostrom's so-called orthogonality thesis, which states (roughly) that pretty much any ultimate goal is compatible with arbitrarily high intelligence levels (Bostrom, 2012, 2014). While I do think the orthogonality thesis makes good sense and forms a cornerstone in the best theory to date for understanding the drives and motivations of a superintelligent machine (Häggström, 2016, Section 4.6), I still think it leaves room for doubt.

For instance, I don't see any clear refutation of the idea that an objective morality exists and that any sufficiently intelligent being will discover this morality and act upon it. Assuming this, it is admittedly far-fetched to think that maximization of paperclip production would be the objectively right morality, but a less far-fetched idea might be that maximization of hedonic utility is, and this would likely have equally devastating consequences to the existence of humanity (simply because human brains are unlikely to produce the maximal amount possible of hedonic utility per kilogram of matter).

Here's the basic idea for how the AI might convey the lethal message X regardless of what particular questions we put to it. Let

$$p = \sum_{i=1}^n 2^{-i} X_i$$

so that the binary expansion of p is $p=0.X_1X_2\dots X_n$. Suppose now that the AI answers every question at random, independently, with probability $p+2^{-(n+1)}$ for a yes answer and the complementary probability $1-p-2^{-(n+1)}$ for a no answer. If M_m denotes the fraction of yes answers among the first m questions, then it follows from the weak law of large numbers (see, e.g., Durrett, 2010, or any first-year probability theory textbook, for this and other basic probability results cited in this section and the next) that for large m , we have with probability close to 1 that M_m is close to $p+2^{-(n+1)}$. More precisely, for any $\epsilon>0$, the probability that M_m lies in the interval

$$p+2^{-(n+1)}-\epsilon < M_m < p+2^{-(n+1)}+\epsilon$$

tends to 1 as $m\rightarrow\infty$. With $\epsilon=2^{-(n+2)}$, having M_m in that interval guarantees that its binary expansion begins $0.X_1X_2\dots X_n$, so that in other words the lethal message X can be read off from the fraction of yes answers to the first m questions.

This assumes that the AI has a physical randomization mechanism (which could be based, e.g., on thermal noise). If it has to rely on its own generation of pseudo-random numbers, then, if a question is repeated, the use of the reset button guarantees that the exact same computations are done, leading to the same answer, and the independence assumption fails. Other than that, it seems plausible to assume that a superintelligent AI will be able to produce pseudo-random numbers (using the question received as a seed) that to us mere humans look perfectly random, and the above analysis applies. And if the sequence of questions does contain repeats, the lethal message X can be reconstructed from the answers to the subsequence of questions where repeats are ignored.

But how large does m need to be? In the next paragraph I'll give a probabilistic estimate which hits the right order of magnitude, but before that, here is an intuition for why m needs to increase (at least) exponentially in n . There are 2^n possible length- n messages, and any method that is able to distinguish between these based on M_m needs m to be large enough so that M_m has at least 2^n possible values. Since M_m has $m+1$ possible values, this requires $m \geq 2^n - 1$.

Now for the probabilistic estimate. The variance of M_m can be bounded using that $r(1-r) \leq 1/4$ for all real r , giving

$$\text{Var}[M_m] = \frac{1}{m}(p+2^{-(n+1)})(1-p-2^{-(n+1)}) \leq \frac{1}{4m}$$

and Chebyshev's inequality gives

$$\mathbf{P}[M_m \text{ fails to reconstruct } X] \leq \mathbf{P}(|M_m - (p+2^{-(n+1)})| \geq 2^{-(n+2)}) \leq \text{Var}[M_m]/(2^{-(n+2)})^2 = \frac{1}{4m}2^{2(n+2)}$$

so that we can guarantee a reconstruction probability of at least $1/2$ by taking $m \geq 2^{2(n+2)-1}$. This exponential growth of m makes the algorithm quite hopeless; even a modest message length such as $n=10^3$ requires $m \geq 2^{2003} > 10^{600}$, which not even the most patient oracle AI has time for. Asking for a more modest reconstruction probability such as 0.1 or 0.001 , or using slightly better estimates involving the central limit theorem (rather than Chebyshev's inequality), can buy the AI a constant factor in terms of running time, but does not change the fact that the running time grows exponentially in n , so this frequency-based approach to smuggling out the lethal message X is useless.

There's a second reason why this basic frequency-based approach is useless for the AI, namely that it makes it easy for us to reveal its dishonest intentions. Even if it uses pseudo-random number generation rather than physical randomness, guaranteeing that it gives the same answer to repeats of the same question, we can test its honesty with questions to which we know the answer, such as "Are there infinitely many prime numbers?" and "Is $7+5=13$?". If we ask both easy questions whose answer is yes, and easy questions whose answer is no, then no matter what the lethal message X is, we will quickly discover that the AI is lying to us.

The next two sections deal with possible modifications of the frequency-based approach to handle these shortcomings. The speed issue is the topic of Section 4, and the concealment issue is treated in Section 5.

4. Addressing the speed issue

Consider as in the previous section a lethal message X of length $n=10^3$, which with the basic frequency-based approach requires roughly 10^{600} questions to smuggle out. Suppose now that the

AI has some classification rule that to each question it receives assigns one of two types, “type 1” and “type 2”, say. There are a kazillion possible choices of such a classification, such as defining a question to be of type 1 if and only if it contains the word “Stockholm”. The challenge, however, for the AI (in order to apply the communication method I am about to describe) is to come up with a classification rule such that it can reasonably expect that in the long run, about half of the questions we pose to it will be of type 1. Suppose it has figured out such a classification. In other words, the AI has found a way to partition the set Q of all possible questions into two sets Q_1 and Q_2 , in such a way that it can expect that in the long run, close to half of our questions will be in Q_1 and half in Q_2 . The Stockholm example is unlikely to satisfy this; a better candidate might be to take Q_1 to consist of all question such that the sum of the ASCII codes of all the characters in the question is even, and Q_2 to consist of those where the sum is odd. More sophisticated choices are of course possible, but this choice of (Q_1, Q_2) is likely to work in the AI’s desired way, unless we humans who pose questions suspect what the AI is up to, and decide actively to throw a spanner in its works.

The idea here is that the AI uses questions in Q_1 to encode the first $n/2=500$ characters in X , and questions in Q_2 to encode the last 500, using the frequency-based approach in Section 3.

Encoding a message with $n/2$ characters requires roughly $10^{600/2}=10^{300}$ questions, and after roughly $2 \cdot 10^{300}$ questions we have enough questions of either type to reconstruct X with reasonable probability; since there will be two relative frequencies, one for Q_1 and one for Q_2 , rather than just one that we need to estimate, we just need to replace the factor 2 by something slightly larger in order to get the same reconstruction probability. The total number of questions needed, although still (beyond) astronomical, is quite an improvement compared to the 10^{600} that

the basic method of Section 3 gave.

Further partitioning of each of Q_1 and Q_2 into smaller subsets of Q carries similar further improvements. The logical endpoint of these improvements is to partition Q into n subsets Q_1, Q_2, \dots, Q_n such that for each i the answers to questions in Q_i are used to communicate bit X_i . This leads to the coding rule that if $X_i=0$, then all questions in Q_i receive the answer no, whereas if $X_i=1$, then all questions in Q_i receive the answer yes. For someone who knows the partition (Q_1, Q_2, \dots, Q_n) , reconstructing the message X is possible as soon as at least one question in each of the subsets Q_1, Q_2, \dots, Q_n have been asked.

How long would that take? Suppose, with a slight strengthening compared to the case above with just the partition (Q_1, Q_2) , that the AI manages to construct the partition in such a way that each time a new question is asked, it has for each i probability (close to) $1/n$ of ending up in Q_i , independently of where previous questions ended up. For a modest value of n such as $n=10^3$, letting Q_i consist of all questions such that the modulo n sum of the ASCII codes of the characters in the question is $i-1$ seems like a fairly reasonable choice. For larger n such as $n=10^9$, something slightly more elaborate is needed, but it is not hard to come up with reasonable suggestions based, e.g., on weighted sums of ASCII codes (or something completely different).

Assuming that the whereabouts in (Q_1, Q_2, \dots, Q_n) of the questions are independent and uniformly distributed, the issue of when all of the partition sets Q_i have received at least one question is exactly the coupon collector's problem in probability theory, for which it is well known and easy to verify (see, e.g., Durrett, 2010) that the number of questions needed is

concentrated around $n \log(n)$. So for $n=10^3$ about $7 \cdot 10^3$ questions are needed, which seems quite feasible, whereas $n=10^9$ requires about $2 \cdot 10^{10}$ questions, which might be feasible, but only provided some automatization of the question-asking-and-answer-receiving process and/or parallelization using many copies of the oracle AI. In any case, this is only a factor $\log(n)$ worse than linear in n , and since linear is obviously the best the AI can achieve (provided the message X does not have an extreme amount of redundancy), this seems pretty good for the AI.

However, even the $\log(n)$ factor can be gotten rid of using a so-called binary erasure channel (see, e.g., Cover and Thomas, 1991), which is a binary channel in which each bit independently is erased with probability α and transmitted correctly with probability $1 - \alpha$. The capacity of this channel (defined as the reciprocal of the limit as n tends to ∞ of the smallest C such that a message of length n can be correctly reconstructed with high probability when sending Cn bits) is known to be $1 - \alpha$, so that there exists an encoding that takes a length n message X into the length $m = n / (1 - \alpha)$ message Y and transmission of Y on this channel allows correct reconstruction of X with high probability. Erasure probability α corresponds in our setting to asking enough questions so that the probability of getting at least one question in Q_i is $1 - \alpha$. Since the binary sequence Y transmitted by the AI has length m , we define the partition (Q_1, Q_2, \dots, Q_m) of the question set Q as above. Getting at least proportion $1 - \alpha$ of the sets to receive at least one question is the generalization of the coupon collector's problem which Durrett (2010) calls the occupancy problem, and the number of questions needed to achieve this turns out to be concentrated around $m \log(\alpha^{-1}) = n \log(\alpha) / (\alpha - 1)$. Hence the number of questions needed is linear in n , and we can make the proportionality constant close to the optimal value 1 by choosing the erasure probability close to 1, since $\log(\alpha) / (\alpha - 1) \rightarrow 1$ as $\alpha \rightarrow 1$.

5. Addressing the concealment issue

An oracle AI trying to convey the lethal message $X=(X_1, X_2, \dots, X_n)$ by giving less than honest answers to our questions will want to do this in such a way that its dishonesty goes undetected, lest we turn the machine off. If it employs the methods in Section 4, with the ASCII-based approach to define (Q_1, Q_2, \dots, Q_n) , then its dishonesty is almost certainly going to be exposed. Here are four simple reasons for this.

(a) We can ask it questions such as “Is $7+5=13$?” that we know the answer to. Either the AI will with overwhelming probability quickly run into questions where its approach leads it to a false answer, or else (the much less likely case) we humans will be able to reconstruct X without actually consulting the AI, simply by finding at least one easy question in each Q_i , which seems like an easy task, so that the lethal message is already out there and the AI needn’t bother.

(b) We can ask it difficult but obviously equivalent questions, such as “Is Goldbach’s conjecture true?” and “Is it the case that Goldbach’s conjecture is true?”. Even though we do not know the answer to either of these, we know that the answers coincide, but the AI needs to be extremely lucky not to quickly give contradictory answers to some such pair of questions.

(c) We can ask difficult questions and negated forms of them, such as “Is Goldbach’s conjecture true?” and “Is Goldbach’s conjecture false?”. Here the AI is again likely to be quickly exposed, for reasons similar to those in (b).

(d) If we can correctly guess what (Q_1, Q_2, \dots, Q_n) is, then we also understand that there is no sane reason why an ASCII-based rule should result in all questions in, say, Q_1

having the same answer, so by posing a stream of otherwise seemingly unrelated questions, all from Q_1 , we have strong reason to be suspicious if we receive the same answer to all of them.

The AI's likely antidote to (a) is to answer truthfully all questions that it does not identify as too difficult for mere humans, and to employ the strategies from Section 4 only to those it confidently judges that we cannot know the answer to. There may be some subtlety in this, because of tricks of the following kind. If we ask the AI questions about the prime factorization of a 200-digit number that looks like it is hard to factorize, it might be that we have obtained that number by first finding two distinct 100-digit primes (this we know how to do; see Agrawal, Kayal and Saxena, 2004) and then multiplying them, in which case we of course know its prime factorization. A superintelligent AI would presumably not fall for this particular trick, which nevertheless illustrates the need for the AI to think broadly and a bit paranoidly about what epistemic situation we might be in.

Concerning (b), it seems necessary for the AI to do something like first transforming the question it receives into a canonical form according to some rigid rule that ends up in the same place for all equivalent questions, and then produce an answer based on which Q_i that particular question sits in. This seems like it could be difficult, and note that the agent out there who is the intended recipient of X needs to know the same rule, or at least which Q_i it lands in. But then again, the AI and the agent need only do this for some particular class of questions, while all others are answered honestly by the AI and ignored by the agent. This might make their task considerably easier.

Here we should not be too rigid about the notion of equivalent questions. Equivalence is formally required to satisfy transitivity, so that if $q_1, q_2, \dots, q_{1000}$ are questions such that for each j , q_j and q_{j+1} are equivalent, then q_1 and q_{1000} are equivalent as well. But the important thing here is not logical equivalence but rather the property of being easily seen by (smart) humans to be equivalent. That property could well hold for every pair (q_j, q_{j+1}) in the above sequence, while not holding for (q_1, q_{1000}) . The AI may want to put such q_1 and q_{1000} in different Q_i 's (as otherwise it may find too large parts of the question space to be forced into the same Q_i), which would require it to give different answers to questions q_j and q_{j+1} for some j . If we come across such a pair (q_j, q_{j+1}) we can detect the AI's dishonesty, so this may seem like a problem for the AI, but perhaps not a crucial one if it can locate such discrepancies to parts of the question space Q where the questions are unlikely for humans to come up with. (If, as discussed in Section 2, the AI has not just two but three answers to its disposal – yes, no, and don't know – it might help for it to buffer such discrepancies with don't know answers.)

Problem (c) can be treated similarly as (b) by the AI, except that it unites the class of equivalent formulations of a question with the corresponding class for the negated question, and has a rule for what is the unique canonical form within this union. It then responds with the answer determined by the Q_i that the canonical formulation sits in, but negated in case the canonical form is negated compared to the original question.

Problem (d) seems to show that the AI's dishonesty is doomed to be exposed in case its human safety administrator can guess the partition (Q_1, Q_2, \dots, Q_n) together with the canonical form rule used for (b) and (c). But this is stuff that the agent out there needs to know too in order to

retrieve X , and the agent is presumably not superintelligent, so how can it happen that the agent does but the safety administrator does not? Well, humans have a tendency not to always be in the same epistemic situation, so it is not such an implausible situation. We can imagine a kind of epistemic arms race between the agent and the AI's safety administrator, while the AI sits there trying to predict what such an arms race can result in in terms of what the agent may plausibly have thought of that the safety administrator hasn't.

6. Conclusions

We have seen that there are strategies for an oracle AI with reset button to try to work around our attempts to protect ourselves from (potentially lethal) messages that we haven't requested. There are (at least) two challenges to these strategies: speed and concealment. While the speed issue on its own seems to be resolvable to the AI's satisfaction (and our jeopardy), the concealment issue seems much more difficult for the AI. Some suggestions are made in Section 5, with much handwaving however, and it is highly unclear whether they can be implemented. A superintelligent being can of course figure out many things that we cannot, but what might possibly work to our advantage is that (at least some of) the intricate implementational details need to be figured out also by the non-superintelligent agent out there to which the lethal message X is directed.

It may be tempting to think that, given how this seems to point towards the need for the AI to encode X in a rather convoluted manner, we need only worry about the scenario with an agent who sets out to help the AI, and not about the case suggested in Section 1 with a communications administrator who has no desire to see X but stumbles upon it and finds it irresistibly convincing. But we should perhaps not be too quick in dismissing that scenario even if we find that the AI

needs to present X in a complex encoded way. One way in which the AI might succeed is if it uses its superintelligence to figure out some unknown quirk in the human pattern recognition system that unexpectedly makes X visible to us despite the complex encoding. Another way is if the human safety administrators accidentally come across X in the very process of trying to outsmart the AI by figuring out its method for conveying X .

Finally, there is the issue of what would happen upon a discovery that the AI is dishonest. So far the tacit assumption is that the AI needs to avoid that in order to avoid being turned off. But suppose that it is owned by a private firm, and that it seems to be dishonest only about some of the questions while being truthful in most of them, with a potential for providing information of immense economic value. Can the temptation to keep using the AI, despite concerns about its intentions, be resisted? As pointed out by Armstrong, Sandberg and Bostrom (2012) and Miller (2012), the situation can become even worse if the firm fears that hesitation means that a first-mover advantage is handed over to some competitor working on their own AI system. Similar concerns apply to competition between states.

Acknowledgement. I am grateful to Douglas Wikström and to an anonymous referee for valuable advice.

References

- Agrawal, M., Kayal, N. and Saxena, N. (2004) PRIMES is in P, *Annals of Mathematics* **160**, 781-793.
- Armstrong, S. (2010) The AI in a box boxes you, *Less Wrong*, February 2.
- Armstrong, S., Sandberg, A. and Bostrom, N. (2012) Thinking inside the box: controlling and

using an oracle AI, *Minds and Machines* **22**, 299-324.

Bostrom, N. (2003) Ethical issues in advanced artificial intelligence, *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, Vol. 2 (ed. Smit, I. et al.) International Institute of Advanced Studies in Systems Research and Cybernetics, pp. 12-17.

Bostrom, N. (2012) The superintelligent will: Motivation and instrumental rationality in advanced artificial agents, *Minds and Machines* **22**, 71-85.

Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, Oxford.

Cover, T. and Thomas, J. (1991) *Elements of Information Theory*, Wiley, 1991.

Danaher, J. (2015) Why AI doomsayers are like skeptical theists and why it matters, *Minds and Machines* **25**, 231-246.

Durrett, R. (2010) *Probability: Theory and Examples* (4th ed), Cambridge University Press, Cambridge, UK.

Hägström, O. (2016) *Here Be Dragons: Science, Technology and the Future of Humanity*, Oxford University Press, Oxford.

Kurzweil, R. (2005) *The Singularity Is Near: When Humans Transcend Biology*, Viking, New York.

Miller, J. (2012) *Singularity Rising: Surviving and Thriving in a Smarter, Richer and More Dangerous World*, Benbella, Dallas, TX.

Ritter, M. (2012) Bird flu study published after terrorism debate, *CNS News*, June 21.

Sotala, K. and Yampolskiy, R. (2015) Responses to catastrophic AGI risk: a survey, *Physica Scripta* **90**, 018001.

Yampolskiy, R. (2012) Leakproofing the singularity: artificial intelligence confinement problem, *Journal of Consciousness Studies* **19**, 194-214.

Yudkowsky, E. (2008) Artificial intelligence as a positive and negative factor in global risk, in *Global Catastrophic Risks* (eds Bostrom, N. and Čirković, M.), Oxford University Press, Oxford, pp 308-345.

Yudkowsky, E. (2013) Five theses, two lemmas, and a couple of strategic implications, Machine Intelligence Research Institute, May 5.