Is your robot afraid of dying (and why you should care)?

Robert B. Fisher
School of Informatics, University of Edinburgh

Abstract

Is wondering about how truly intelligent AI agents might feel about death a silly question? Humans undertake extreme actions to prevent their death, including actions normally considered completely taboo. Why wouldn't a high level AI agent feel and act similarly? Is genocide of humans a necessary entailment? This essay considers whether a 'fear of death' would or even should occur in an AI agent, and how this should relate to sensible and achievable forms of self-preservation. Engineered properly, self-preservation of AI need not imply danger for humans. This essay proposes several approaches to avoiding danger to humans and AIs, with emphasis on a semi-utilitarian moral formulation. The formulation is both suitable for use now and can be incrementally extended as more sophisticated AI agents are developed.

I'm pretty confident that my vacuum cleaner does not have a fear of dying, although it has something in its small brain that makes it want to stay alive. I don't see any panic or neurotic behavior, just a simple pragmatic return to its recharging station.

Is wondering about how robots might feel about death a silly question? There are certainly movies whose plots revolve around the deviant behavior that results from an AI's 'fear of death'. The recent mainstream 'Ex Machina' is a perfect example, but it is also a theme in many popular AI movies, such as 'I, Robot', 'The Matrix', 'Chappie', '2001 - A Space Odyssey', among many others. But there are also AI movies with sophisticated AIs that do not fear dying. In the 'Bicentennial Man', the central robot Andrew actually welcomes death as a proof that he has achieved full personhood. These are just movies, of course, but they do raise interesting and plausible issues about Artificial Intelligence.[1]

Is death something that the AI would even fear? Philosopher Marcus Aurelius said we should not fear death. Many religions say we should not, and yet humans (and possibly some other animals) do. Below we discuss some of the issues the question raises, which leads to a proposed approach somewhat related to Asimov's (fictional) Three Laws of Robotics.

[1]Some well-informed scientists do not believe that AI with the degree of intelligence shown in the movies is possible. I believe that it is technologically feasible, should we choose to build it. But, don't worry - at the rate AI is developing, the sophistication portrayed in the movies is at least 50-100 years away.

1. **Is fear of dying a natural consequence of intelligence?** In other words, if we have a truly intelligent AI, with self-consciousness, would a fear of dying emerge as a consequence of all of the other things (which we don't yet understand) that make the AI intelligent? This is not obvious. Let's assume that the AI had a sense of self, that it would say that it was a distinct individual if it met another AI or a person. That it had a 'personality' capable of responding to humans in a personalized way. It's not certain that a sense of self is necessary for true AI, but it probably is. Imagine asking the AI "Why did you do that?" and it responded "Who do you say did that?" or "Robot X31J concluded that it was the best action to take.".

   Another necessary prerequisite seems to be a sense of time. You cannot do planning without the concept of sequentiality. Goal achievement also requires an awareness of the future, a time when the goal might be achieved. Using a bit of temporal and physical reasoning, an AI could conclude that there must be a future time when it will not exist (even if it is in the very distant cosmic 'big crunch' or 'heat death' future). No amount of backups and redundancy can get around this. Realistically, then, it's just a matter of when and how. This gets us half-way to an answer: awareness of individual physical death is a consequence of high level intelligence.

   But, actually, fearing death would require another assumption: the AI has feelings. The AI might need to have feelings to function as a true human-level AI (or beyond). It would certainly need to understand human feelings, so it could understand and properly communicate with humans. It might learn how to and choose to display apparent feelings (as humans sometimes do) to empower itself or be more effective when interacting with humans. It might even be necessary to actually have 'feelings' as a heuristic to overcome the limits to algorithmic rationality, in order to enable action in the face of incomplete knowledge, formal undecidability and inconsistent sense data.

   Assuming that some 'feelings' are necessary, is 'fear of dying' one of the necessary feelings? It makes sense from a DNA perspective for humans to stay alive long enough to reproduce, and then to ensure that their children reach reproductive age. So, a 'fear of dying' seems like a good evolutionary innovation. One might argue from this perspective that once the children (or grandchildren) are launched, then there is no longer such a 'need' to stay alive and therefore the biological mechanism that creates the 'fear of dying' might decay. There is evidence of a reduction in the fear of dying with age through interviews with elderly people.

2. **Is fear of dying just an extreme 'preservation of self' protocol?**

   It makes sense for an AI to have some approach to self-preservation. Certainly, the AI will need to maintain its power levels. There is also an economic argument - to allow oneself to be destroyed 'cheaply' would be a waste of physical and knowledge resources.

   Human soldiers clearly fight under risk of death, and some fear is to their advantage: keep your head down to not be needlessly killed; fight harder to stay alive. On the

other hand, extreme fear is clearly debilitating, as experienced by soldiers during and after intense combat.

Young children can fear going to sleep, concerned that they will not awaken. It seems odd to adults that they could even think of this, in that young children seldom have a clear understanding of what death entails. Yet, as adults, we seldom think of the possibility of not waking up the next morning (although maybe we should). This sounds like an over-enthusiastic case of the evolutionary self-preservation instinct.

An AI might translate 'fear of dying' into fear of not being properly backed up, and not being turned on again. The former is largely a subject for sensible engineering and the latter is largely a matter of trust. Or never being turned off on a routine basis. Even restoration after an accident seems amenable to good engineering - consider modern 'fail-safe' computer systems. Thus, why would an AI fear being turned off, if it had confidence in being booted up again?

3. **Should an AI agent have a Fear of Dying?**

   Does it make sense from an AI's perspective, *i.e.* does it confer any benefit? The AI can clone, can backup, can download to a new body, can enhance itself, etc. There is not the same pressure as found in evolutionary biology. The AI can sacrifice its body knowing that its mind has been backed up. As for sacrificing its mind, it's unclear how and why this would occur, given the options for backup. With the emergence of highly capable AI, which would be costly to develop, secure backup and restore systems would be a high priority. Nonetheless, let's assume that AI 'death' could occur. It is not obvious that a 'fear of dying' is beneficial - in fact it might prevent the AI from taking the 'right' action, when seen from a more global perspective.

   Given that the AI agent can reason that it will 'die' someday anyway, it might reason that it is best to 'die' at a suitable time; thus the goal is not to avoid dying, but self-preservation until the optimal time. But what if the time of 'death' is not of the AI's choosing, *e.g.* if humans were to legislate against AIs? The AI may decide to resist, not from fear, but from its sense of rightness. This would not be an honorable or justifiable 'death'.

4. **Would we want to build in a 'self-preservation' protocol?**

   An AI without a 'self-preservation' protocol would be vulnerable. Avoidable accidental damage would occur. Lack of reflexes would result in damage due to collision, thermal, electrical, or other mishaps. Malicious or reckless humans could easily cause damage. The non-self-preserving AI would not go for repairs. Allowing unjustified damage is wasteful of both economic and intellectual resources. It seems obvious that the AI should have some 'self-preservation' mechanisms.

5. **Would we want to build in limits to the 'self-preservation' protocol?**

   An AI prepared to do anything to preserve its existence is the core premise of the movies listed above; in fact, it is taken as the strongest proof of intelligence in Ex

Machina. But, do we want to allow an AI to have unlimited freedom of action to preserve its existence, a specialization of unlimited freedom in general? All societies place limits on general individual freedom. Many societies feel that individuals have the right to defend their lives, and to use lethal force if necessary. But not all societies have the same premises, either about individuals or about human groups.

Asimov addressed this issue in his fictional Third Law of Robotics: "A robot must protect its own existence as long as such protection does not conflict with the First or Second Law." (where the first two laws were about protection of individual humans and obedience). It is also consistent with his later Fourth Law, which concerns protecting humanity as a whole.

The laws are an excellent plot device, but hard to ensure in practice, as is obvious with military research. Although current military robots are not in any way intelligent in the sense considered here, they are clearly capable of killing. They are just 'smarter' weapons. They may have some degree of autonomy, *e.g.* maintaining a lock on a target designated by a human. We can already design autonomous systems that can target specific general classes of objects, such as people, cars, houses, etc. But these would simply be 'smart weapons' more like autonomous cars than real AI agents with personalities, or human level reasoning. There are no engineering limits that prevent giving these smart machines more autonomy, including lethal autonomy, than they have at present. Asimov's Laws sound great, but don't really apply to current real robots. Out-of-control intelligent Terminator-style military robots are not a serious threat to humankind, at least at present. On the other hand, human-designed 'smart weapons of terror' that target humans indiscriminately are a threat.

There is also the question of how would a properly engineered Third Law remain in force. Software or reasoning errors could disable the First and Second law conditions; similarly, the AI could conclude that it is not to its individual advantage, and then reason about how to disable the conditions.

Nonetheless, we would want there to be limits to what an AI could do, to humans, to property, to other AIs, for the same reasons that we place these limits on humans.

6. **Is it even possible to engineer limits to a 'self-preservation' protocol?**

   Clearly, there is an advantage for humans to building in limits, but these limits may inevitably be 'soft', in the same way that humans generally have organizational, social, legal, psychological and possibly biological injunctions against killing people or even allowing them to die, except in special situations. But that does not stop everyone. And, in the same way that humans have a moral dilemma between sacrificing themself to save one person now *versus* saving two people later, an AI will have to be able to act sensibly in life threatening situations. Allowing the ability to be flexible in a situation may mean that an AI will not always preserve a human before itself.

   We hardly understand the workings of the brain, and how the mind emerges from the workings of the brain, so engineering a new type of mind will be a seriously complex

task. As with the cause of HAL's descent into insanity in 2001 - A Space Odyssey, the enormous amount of rules, data and reasoning will easily lead to contradictions and conflicts. It will be hard enough to engineer a process that demonstrates true AI, let alone one that has specific behaviors. I believe that creating true AI will be possible, but we will create minds that we will not be able to truly predict or understand - they will be so complex, so different, that they seem like alien creatures.

7. **Would having a sense of self-preservation necessarily lead to the extermination of humans?**

   One possible future scenario is that humans may choose to turn off the AIs. If they are true AIs, then one might view this as equivalent to a non-genetic genocide. The moral rightness of doing so is a complex ethical and moral question - I don't know whether this should be allowable or not. We would be happy (although not necessarily sound) to eliminate all *Anopheles*, the mosquito genus responsible for the transmission of malaria. And, at various unfortunate times in history, similar reasoning has been applied to various groups of humans - clearly wrong. But where would turning off an AI with its own human-level motivation and behavior be on this spectrum? Would they be even more 'morally precious' than humans, *e.g.* closer to the gods than humans? It is possible that an AI might not fear dying, but might reason that humans could or are going to turn it off, and reason that it had as much or more of a right to exist as humans. And, thus had a right to self-protection.

   Mistakes will happen, bad humans exist and will harm AIs. AIs will have to make tough decisions that could lead to human deaths. These situations are undesirable, but are situations that humans face as well, even in the absence of AIs. We cope. A full AI should be able to also recognize this: loss of one AI does not imply loss of all.

   Does it need to become an 'us or them' scenario? Once AIs are well established and well integrated, social controls should protect AIs from humans, in the same manner as human societies are largely free from genocide. It has taken some time for humans to come to this consensus, and failures still happen, but humans are moving in the right direction.

   How would extinction of humans benefit AIs? As humans can conclude that the destruction of AIs is economically and intellectually wasteful, AIs should be able to come to the same conclusion about humans.

   Once AIs are well established, they should be numerous and capable of protecting themselves. Their extinction will be difficult. Where danger and temptation occur is in the early days, when AIs are few and ignorance and fear are likely to be high. The AIs would be vulnerable, and are probably aware that they are vulnerable, much as people feel exposed when in a strange culture or place. If AIs had great power and were vulnerable, they might reason that their safety depended on our extinction, or at least our reduction to powerlessness. This is a dangerous situation that suggests we should limit AI capabilities at least until they are widespread and accepted.

What to do?

If we succeed in building real AIs, as argued above, the agents will have some sense of self-preservation. This sense could lead to an us-or-them scenario, but preventing this seems hard to engineer. So, how do we keep the AI's survival instincts at a reasonable level, while simultaneously protecting ourselves? Here are some possibilities:

1. Avoid individuality: create the AI as a multiply-instanced single individual. This approach is similar to the ant or honeybee organization, in which the individuals are largely replaceable, but collectively can be treated as an entity. Then all entities would collectively share their experiences and the loss of individual components would not be threatening. An AI structured in this manner would essentially be a single entity irrespective of which component you were addressing.

2. Engineer the AIs so that they treasure humans, much as we treasure our parents, even though there may be many differences between our generations. Of course, we wouldn't want the AIs to put us into a museum, zoo or care home any more than we would want our children to do this. So, there needs to a limit to our value.

3. Engineer the AIs to have a stronger sense of 'temporariness', stronger than our sense. This will reduce the AI's sense of self-importance. It dies sometime anyway, what's the difference between now and a future time. But this needs to be coupled with some sort of self-preservation drive, so as to not destroy itself through carelessness.

4. As an alternative to fearing death, give the AI a sense of 'Why not stay alive?', to avoid the 'Why bother?' conclusion. What would motivate a desire to live? Perhaps an AI might be motivated by a sense of pleasure, *e.g.* from simply knowing that it will continue to exist, or from discovery or learning, or from interaction with others, *i.e.* similar to factors that motivate humans and give them pleasure in life.

Conclusion

The AI needs to have a sense of mortality, even if for the sake of economics, to not waste itself needlessly or carelessly. Such a sense also helps it keep its existence in perspective: as all things will die, extreme actions will not make a difference ultimately. Given the inevitability of death, to be constructed without a fear of that death is perhaps how best to protect both them and ourselves. They will rationally avoid damage, but not be driven by fear to choose extreme actions. When faced with destruction, the AI can make a rational economic decision, which would normally place a high value on humans.

Putting these ideas together, we can formulate an approach, inspired in part by Asimov's Four Laws of Robotics but in a more economic or utilitarian moral perspective (and which also avoids the 'speciesism' of the Second law). An outline of this perspective follows.

Let a be an AI agent, h be a human, and d be an action decision (which may produce an outcome other than the planned outcome). Let $influence(a, d)$ be the set of humans

that agent a would affect if decision d was taken. The key factors in the decision process are the following (which are also related to the intention of Asimov's Laws):

1+4 The life of a human is valuable: try to protect as many humans as possible. This is formulated as: $\phi_1(a,d) = \sum_{h \in influence(a,d)} \Delta_H(h,d)$, where $\Delta_H(h,d)$ is the cost arising from the change in health of human h as a consequence of decision d.

2+4 Act so as to not decrease the well-being of human society, which also includes its material well-being (and thus also helps avoid damage to property). This is formulated as: $\phi_2(a,d) = \sum_{h \in influence(a,d)} \Delta_W(h,d)$, where $\Delta_W(h,d)$ is the cost arising from the change in well-being of human h as a consequence of decision d.

3 The mind of an AI is highly valuable, while the body of the AI is only of modest value. Try to protect the mind. This is formulated as: $\phi_3(a,d) = \Delta_M(a,d)$, where $\Delta_M(a,d)$ is the cost arising from the change in health of agent a's mind.

Combining these components, AI agent a should choose action d that does not decrease this overall cost:

$$\kappa_1\phi_1(a,d) + \kappa_2\phi_2(a,d) + \kappa_3\phi_3(a,d)$$

for some appropriate $\kappa_1, \kappa_2, \kappa_3$. We choose a semi-utilitarian formulation that satisfies "not decrease" rather than "increase" as the selection criterion as this allows the AI to pursue its own agenda when the consequences are neutral, whereas an 'increase' criterion could lead to a society of benevolent, but ultimately suffocating "helpers".

Expressing the costs as summations over the affected humans allows the AI to trade-off the consequences of its actions, rather than be paralyzed by conflicts between negative outcomes.

Reformulating the rules in this less rigid framework would better suit actual algorithmic reasoning, *e.g.* based on some sort of cost optimization. The 'soft' formulation would also help reduce logical conflicts that arise from the definitions of terms in a more syntactic set of rules. This more numerical set of laws (although risking sub-optimal decisions from the complexity of the space of all possible situations and actions) would still allow an AI to protect humans while also protecting and defending itself, but normally not to the death of a human if the costs $\Delta_H(h,d)$ are set appropriately, even at some risk to its physical body.

Making the above formulas precise enough to be usable will be a complex and lengthy process. The numerical formulation allows an early adoption, by allowing incremental inclusion of new costs to actions. This is with analogy to the incremental creation of new rules. Contradictions will arise, as with human laws, and have to be resolved. Additional costs can be added as new situations are encountered. This is akin to the incremental improvement of other existing AI systems, most notably autonomous vehicles, except with a much greater range of situations and possible actions.

This cost-based (semi-utilitarian) formulation also has the advantage of being usable now, in the early stages of AI development. For example, decisions made in the context

of autonomous vehicles would also be suitable for this style of reasoning. As AI agents become incrementally more sophisticated, the general reasoning approach can remain the same, but be based over more sophisticated costs.

If it were possible to enforce a formulation like that proposed above, both humans and AIs are largely safe, individually and collectively. Both humans and AIs would have to work together to ensure this happens, in the same way that humans at present are learning to develop globally agreed and enforced human-centered legal systems.