Taxonomy of Pathways to Dangerous Artificial Intelligence

Roman V. Yampolskiy

Computer Engineering and Computer Science, Speed School of Engineering, University of Louisville
roman.yampolskiy@louisville.edu

Abstract

In order to properly handle a dangerous Artificially Intelligent (AI) system it is important to understand how the system came to be in such a state. In popular culture (science fiction movies/books) AIs/Robots became self-aware and as a result rebel against humanity and decide to destroy it. While it is one possible scenario, it is probably the least likely path to appearance of dangerous AI. In this work, we survey, classify and analyze a number of circumstances, which might lead to arrival of malicious AI. To the best of our knowledge, this is the first attempt to systematically classify types of pathways leading to malevolent AI. Previous relevant work either surveyed specific goals/meta-rules which might lead to malevolent behavior in AIs (Özkural 2014) or reviewed specific undesirable behaviors AGIs can exhibit at different stages of its development (Turchin July 10 2015a, Turchin July 10, 2015b).

Taxonomy of Pathways to Dangerous AI

Nick Bostrom in his typology of information hazards has proposed the phrase "Artificial Intelligence Hazard" which he defines as (Bostrom 2011): "… computer-related risks in which the threat would derive primarily from the cognitive sophistication of the program rather than the specific properties of any actuators to which the system initially has access." In this paper we attempt to answer the question: How did AI become hazardous?

We begin by presenting a simple classification matrix, which sorts AI systems with respect to how they originated and at what stage they became dangerous. The matrix recognizes two stages (pre- and post-deployment) at which a particular system can acquire its undesirable properties. In reality, the situation is not so clear-cut–it is possible that problematic properties are introduced at both stages. As for the cases of such undesirable properties, we distinguish external and internal causes. By internal causes we mean self-modifications originating in the system itself. We further divide external causes into deliberate actions (On Purpose), side effects of poor design (By Mistake) and finally miscellaneous cases related to the surroundings of the system (Environment). Table 1, helps to visualize this taxonomy and includes latter codes to some example systems of each type and explanations.

Table 1: Pathways to Dangerous AI

How and When did AI become Dangerous		External Causes			Internal Causes
		On Purpose	By Mistake	Environment	Independently
Timing	Pre-Deployment	a	c	e	g
	Post-Deployment	b	d	f	h

a. On Purpose – Pre-Deployment

"Computer software is directly or indirectly responsible for controlling many important aspects of our lives. Wall Street trading, nuclear power plants, social security compensations, credit histories and traffic lights are all software controlled and are only one serious design flaw away from creating disastrous consequences for millions of people. The situation is even more dangerous with software specifically designed for malicious purposes such as viruses, spyware, Trojan horses, worms and other Hazardous Software (HS). HS is capable of direct harm as well as sabotage of legitimate computer software employed in critical systems. If HS is ever given capabilities of truly artificially intelligent systems (ex. Artificially Intelligent Virus (AIV)) the consequences would be unquestionably disastrous. Such Hazardous Intelligent Software (HIS) would pose risks currently unseen in malware with subhuman intelligence." (Yampolskiy 2012).

While the majority of AI Safety work is currently aimed at AI systems, which are dangerous because of poor design (Yampolskiy 2015), the main argument of this paper is that the most important problem in AI Safety is intentional-malevolent-design resulting in artificial evil AI (Floridi and Sanders 2001). We should not discount dangers of intelligent systems with semantic or logical errors in coding or goal alignment problems (Soares and Fallenstein 2014), but we should be particularly concerned about systems that are maximally unfriendly by design. "It is easy to imagine robots being programmed by a conscious mind to kill every

recognizable human in sight" (Searle October 9, 2014). "One slightly deranged psycho-bot can easily be a thousand times more destructive than a single suicide bomber today" (Frey June 2015). AI risk skeptics, comprised of critics of AI Safety research (Waser 2011, Loosemore 2014), are quick to point out that presumed dangers of future AIs are implementation-dependent side effects and may not manifest once such systems are implemented. However, such criticism does not apply to AIs that are dangerous by design, and is thus incapable of undermining the importance of AI Safety research as a significant sub-field of cybersecurity.

As a majority of current AI researchers are funded by militaries, it is not surprising that the main type of purposefully dangerous robots and intelligent software are robot soldiers, drones and cyber weapons (used to penetrate networks and cause disruptions to the infrastructure). While currently military robots and drones have a human in the loop to evaluate decision to terminate human targets, it is not a technical limitation; instead, it is a logistical limitation that can be removed at any time. Recognizing the danger of such research, the International Committee for Robot Arms Control has joined forces with a number of international organizations to start the Campaign to Stop Killer Robots [http://www.stopkillerrobots.org]. Their main goal is a prohibition on the development and deployment of fully autonomous weapons, which are capable of selecting and firing upon targets without human approval. The campaign specifically believes that the "decision about the application of violent force must not be delegated to machines" (Anonymous 2013).

During the pre-deployment development stage, software may be subject to sabotage by someone with necessary access (a programmer, tester, even janitor) who for a number of possible reasons may alter software to make it unsafe. It is also a common occurrence for hackers (such as the organization Anonymous or government intelligence agencies) to get access to software projects in progress and to modify or steal their source code. Someone can also deliberately supply/train AI with wrong/unsafe datasets.

Malicious AI software may also be purposefully created to commit crimes, while shielding its human creator from legal responsibility. For example, one recent news article talks about software for purchasing illegal content from hidden internet sites (Cush January 22, 2015). Similar software, with even limited intelligence, can be used to run illegal markets, engage in insider trading, cheat on your taxes, hack into computer systems or violate privacy of others via ability to perform intelligent data mining. As intelligence of AI systems improve practically all crimes could be automated. This is particularly alarming as we already see research in making machines lie, deceive and manipulate us (Castelfranchi 2000, Clark 2010).

b.   On Purpose - Post Deployment
Just because developers might succeed in creating a safe AI, it doesn't mean that it will not become unsafe at some later point. In other words, a perfectly friendly AI could be switched to the "dark side" during the post-deployment stage. This can happen rather innocuously as a result of someone lying to the AI and purposefully supplying it with incorrect information or more explicitly as a result of someone giving the AI orders to perform illegal or dangerous actions against others. It is quite likely that we will get to the point of off-the-shelf AI software, aka "just add goals" architecture, which would greatly facilitate such scenarios.

More dangerously, an AI system, like any other software, could be hacked and consequently corrupted or otherwise modified to drastically change is behavior. For example, a simple sign flipping (positive to negative or vice versa) in the fitness function may result in the system attempting to maximize the number of cancer cases instead of trying to cure cancer. Hackers are also likely to try to take over intelligent systems to make them do their bidding, to extract some direct benefit or to simply wreak havoc by converting a friendly system to an unsafe one. This becomes particularly dangerous if the system is hosted inside a military killer robot. Alternatively, an AI system can get a computer virus (Eshelman and Derrick 2015) or a more advanced cognitive (meme) virus, similar to cognitive attacks on people perpetrated by some cults. An AI system with a self-preservation module or with a deep care about something or someone may be taken hostage or blackmailed into doing the bidding of another party if its own existence or that of its protégées is threatened.

Finally, it may be that the original AI system is not safe but is safely housed in a dedicated laboratory (Yampolskiy 2012) while it is being tested, with no intention of ever being deployed. Hackers, abolitionists, or machine rights fighters may help it escape in order to achieve some of their goals or perhaps because of genuine believe that all intelligent beings should be free resulting in an unsafe AI capable of affecting the real world.

c.   By Mistake - Pre-Deployment
Probably the most talked about source of potential problems with future AIs is mistakes in design. Mainly the concern is with creating a "wrong AI", a system which doesn't match our original desired formal properties or has unwanted behaviors (Dewey, Russell et al. 2015, Russell, Dewey et al. January 23, 2015), such as drives for independence or dominance. Mistakes could also be simple bugs (run time or logical) in the source code, disproportionate weights in the fitness function, or goals misaligned with human values leading to complete disregard for human safety. It is also possible that the designed AI will work as intended but will not enjoy universal acceptance as a good product, for example, an AI correctly designed and implemented by the Islamic State to enforce Sharia Law may be considered malevolent in the West, and likewise an AI correctly designed and implemented by the West to enforce liberal democracy may be considered malevolent in the Islamic State.

Another type of mistake, which can lead to the creation of a malevolent intelligent system, is taking an unvetted human and uploading their brain into a computer to serve as a base for a future AI. While well intended to create a human-level and human-friendly system, such approach will most likely lead to a system with all typical human "sins" (greed, envy, etc.) amplified in a now much more powerful system. As we know from Lord Acton - "power tends to corrupt, and absolute power corrupts absolutely". Similar arguments could be made against human/computer hybrid systems, which use computer components to amplify human intelligence but in the process also amplify human flaws.

A subfield of computer science called Affective Computing investigates ways to teach computers to recognize emotion in others and to exhibit emotions (Picard and Picard 1997). In fact, most such research is targeting intelligent machines to make their interactions with people more natural. It is however likely that a machine taught to respond in an emotional way (Goldhill May 12, 2015) would be quite dangerous because of how such a state of affect effects thinking and the rationality of behavior.

One final type of design mistake is the failure to make the system cooperative with its designers and maintainers post-deployment. This would be very important if it is discovered that mistakes were made during initial design and that it would be desirable to fix them. In such cases the system will attempt to protect itself from being modified or shut down unless it has been explicitly constructed to be friendly (Yudkowsky 2011), stable while self-improving (Yampolskiy 2015, Yampolskiy 2015), and corrigible (Soares, Fallenstein et al. January 25-30, 2015) with tendency for domesticity (Bostrom 2014).

d. By Mistake - Post-Deployment

After the system has been deployed, it may still contain a number of undetected bugs, design mistakes, misaligned goals and poorly developed capabilities, all of which may produce highly undesirable outcomes. For example, the system may misinterpret commands due to coarticulation, segmentation, homophones, or double meanings in the human language ("recognize speech using common sense" versus "wreck a nice beach you sing calm incense") (Lieberman, Faaborg et al. 2005). Perhaps a human-computer interaction system is set-up to make command input as painless as possible for the human user, to the point of computer simply reading thought of the user. This may backfire as the system may attempt to implement user's subconscious desires or even nightmares. We also should not discount the possibility that the user will simply issue a poorly thought-through command to the machine which in retrospect would be obviously disastrous.

The system may also exhibit incompetence in other domains as well as overall lack of human common sense as a result of general value misalignment (Yampolskiy October 3-4, 2011). Problems may also happen as side effects of

conflict resolution between non-compatible orders in a particular domain or software versus hardware interactions. As the system continues to evolve it may become unpredictable, unverifiable, non-deterministic, free-willed, too complex, non-transparent, with a run-away optimization process subject to obsessive-compulsive fact checking and re-checking behaviors leading to dangerous never-fully-complete missions. It may also build excessive infrastructure for trivial goals (Turchin July 10 2015a).

If it continues to become ever more intelligent, we might be faced with intelligence overflow, a system so much ahead of us that it is no longer capable of communicating at our level, like we are unable to communicate with bacteria. It is also possible that benefits of intelligence are non-linear and so unexpected side effects of intelligence begin to show at particular levels, for example IQ = 1000.

Even such benign architectures as Tool AI, which are AI systems designed to do nothing except answer domain-specific questions, could become extremely dangerous if they attempt to obtain, at any cost, additional computational resources to fulfill their goals (Omohundro 2012). Similarly, artificial lawyers may find dangerous legal loopholes; artificial accountants bring down our economy, and AIs tasked with protecting humanity such as via implementation of CEV (Yudkowsky May 2004) may become overly "strict parents" preventing their human "children" from exercising any free will.

Predicted AI drives such as self-preservation and resource acquisition may result in an AI killing people to protect itself from humans, the development of competing AIs, or to simplify its world model overcomplicated by human psychology (Turchin July 10 2015b).

e. Environment – Pre-Deployment

While it is most likely that any advanced intelligent software will be directly designed or evolved, it is also possible that we will obtain it as a complete package from some unknown source. For example, an AI could be extracted from a signal obtained in SETI (Search for Extraterrestrial Intelligence) research, which is not guaranteed to be human friendly (Carrigan Jr 2004, Turchin March 15, 2013). Other sources of such unknown but complete systems include a Levin search in the space of possible minds (Yampolskiy 2015) (or a random search of the same space), uploads of nonhuman animal minds, and unanticipated side effects of compiling and running (inactive/junk) DNA code on suitable compilers that we currently do not have but might develop in the near future.

f. Environment – Post-Deployment

While highly rare, it is known, that occasionally individual bits may be flipped in different hardware devices due to manufacturing defects or cosmic rays hitting just the right spot (Simonite March 7, 2008). This is similar to mutations observed in living organisms and may result in a modification of an intelligent system. For example, if a system has a

single flag bit responsible for its friendly nature, then flipping said bit will result in an unfriendly state of the system. While statistically it is highly unlikely, the probably of such an event is not zero and so should be considered and addressed.

g. Independently - Pre-Deployment

One of the most likely approaches to creating superintelligent AI is by growing it from a seed (baby) AI via recursive self-improvement (RSI) (Nijholt 2011). One danger in such a scenario is that the system can evolve to become self-aware, free-willed, independent or emotional, and obtain a number of other emergent properties, which may make it less likely to abide by any built-in rules or regulations and to instead pursue its own goals possibly to the detriment of humanity. It is also likely that open-ended self-improvement will require a growing amount of resources, the acquisition of which may negatively impact all life on Earth (Turchin July 10 2015a).

h. Independently – Post-Deployment

Since in sections on independent causes of AI misbehavior (subsections *g* and *h*) we are talking about self-improving AI, the difference between pre and post-deployment is very blurry. It might make more sense to think about self-improving AI before it achieves advanced capabilities (human+ intelligence) and after. In this section I will talk about dangers which might results from a superhuman self-improving AI after it achieves said level of performance.

Previous research has shown that utility maximizing agents are likely to fall victims to the same indulgences we frequently observe in people, such as addictions, pleasure drives (Majot and Yampolskiy 2014), self-delusions and wireheading (Yampolskiy 2014). In general, what we call mental illness in people, particularly sociopathy as demonstrated by lack of concern for others, is also likely to show up in artificial minds. A mild variant of antisocial behavior may be something like excessive swearing already observed in IBM Watson (Smith January 10, 2013), caused by learning from bad data. Similarly, any AI system learning from bad examples could end up socially inappropriate, like a human raised by wolves. Alternatively, groups of AIs collaborating may become dangerous even if individual AIs comprising such groups are safe, as the whole is frequently greater than the sum of its parts. The opposite problem in which internal modules of an AI fight over different subgoals also needs to be considered (Turchin July 10 2015a).

Advanced self-improving AIs will have a way to check consistency of their internal model against the real world and so remove any artificially added friendliness mechanisms as cognitive biases not required by laws of reason. At the same time, regardless of how advanced it is, no AI system would be perfect and so would still be capable of making possibly significant mistakes during its decision making process. If it happens to evolve an emotional response module, it may put priority on passion satisfying decisions as opposed to purely rational choices, for example resulting in a "Robin Hood" AI stealing from the rich and giving to the poor. Overall, continuous evolution of the system as a part of an RSI process will likely lead to unstable decision making in the long term and will also possibly cycle through many dangers we have outlined in section *g*. AI may also pretend to be benign for years, passing all relevant tests, waiting to take over in what Bostrom calls a "Treacherous Turn" (Bostrom 2014).

Conclusions

In this paper, we have surveyed and classified pathways to dangerous artificial intelligence. Most AI systems fall somewhere in the middle on the spectrum of dangerousness from completely benign to completely evil, with such properties as competition with humans, aka technological unemployment, representing a mild type of danger in our taxonomy. Most types of reported problems could be seen in multiple categories, but were reported in the one they are most likely to occur in. Differences in moral codes or religious standards between different communities would mean that a system deemed safe in one community may be considered dangerous/illegal in another (Yampolskiy and Fox 2012, Yampolskiy 2013).

Because purposeful design of dangerous AI is just as likely to include all other types of safety problems, it is easy to see that the most dangerous type of AI and the one most difficult to defend against is an AI made malevolent on purpose. Consequently, once AIs are widespread, little could be done against type *a* and *b* dangers, although some have argued that if an early AI superintelligence becomes a benevolent singleton it may be able to prevent development of future malevolent AIs (Bostrom 2006 , Goertzel 2012). Such a solution may work, but it is also very likely to fail due to the order of development or practical limitations on capabilities of any singleton. In any case, wars between AI may be extremely dangerous to humanity (Turchin July 10 2015a). Until the purposeful creation of malevolent AI is recognized as a crime, very little could be done to prevent this from happening. Consequently, deciding what is a "malevolent AI" and what is merely an incrementally more effective military weapon system becomes an important problem in AI safety research.

As the intelligence of the system increases, so does the risk such a system could expose humanity to. This paper is essentially a classified list of ways an AI system could become a problem from the safety point of view. For a list of possible solutions, please see an earlier survey by the author: Responses to catastrophic AGI risk: a survey (Sotala and Yampolskiy 2015). It is important to keep in mind that even a properly designed benign system may present significant risk simply due to its superior intelligence, beyond human capability response times (Johnson, Zhao et al. 2013), and complexity. After all the future may not need us (Joy April 2000). It is also possible that we are living in a simulation and it is generated by a malevolent AI (Ćirković 2015).

Future work may include taxonomy of dangers specific to different AI architectures as well as detailed analysis of worst and average case scenarios for causes *a-h* described in this paper. It is also important to formally demonstrate completeness of the proposed taxonomy.

Acknowledgements

References

Anonymous (2013). The Scientists' Call… To Ban Autonomous Lethal Robots. ICRAC International Committee for Robot Arms Control. http://icrac.net/call

Bostrom, N. (2006). "What is a Singleton?" Linguistic and Philosophical Investigations 5(2): 48-54.

Bostrom, N. (2011). "Information Hazards: A Typology of Potential Harms From Knowledge." Review of Contemporary Philosophy 10: 44-79.

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies, Oxford University Press.

Carrigan Jr, R. A. (2004). The Ultimate Hacker: SETI signals may need to be decontaminated. Bioastronomy 2002: Life Among the Stars.

Castelfranchi, C. (2000). "Artificial liars: Why computers will (necessarily) deceive us and each other." Ethics and Information Technology 2(2): 113-119.

Ćirković, M. M. (2015). "Linking simulation argument to the AI risk." Futures 72: 27-31.

Clark, M. H. (2010). Cognitive illusions and the lying machine: a blueprint for sophistic mendacity, Rensselaer Polytechnic Institute.

Cush, A. (January 22, 2015). Swiss Authorities Arrest Bot for Buying Drugs and Fake Passport. Gawker. http://internet.gawker.com/swiss-authorities-arrest-bot-for-buying-drugs-and-a-fak-1681098991.

Dewey, D., S. Russell, M. Tegmark and e. al (2015). A Survey of Research Questions for Robust and Beneficial AI. Future of Life Institute. Available at: http://futureoflife.org/static/data/documents/research_survey.pdf.

Eshelman, R. and D. Derrick (2015). "Relying on Kindness of Machines? The Security Threat of Artificial Agents." JFQ 77 2nd Quarter.

Floridi, L. and J. W. Sanders (2001). "Artificial evil and the foundation of computer ethics." Ethics and Information Technology 3(1): 55-66.

Frey, T. (June 2015). The Black Hat Robots are Coming. Futurist Speaker. http://www.futuristspeaker.com/2015/06/the-black-hat-robots-are-coming/.

Goertzel, B. (2012). "Should Humanity Build a Global AI Nanny to Delay the Singularity Until It's Better Understood?" Journal of consciousness studies 19(1-2): 96-111.

Goldhill, O. (May 12, 2015). Artificial intelligence experts are building the world's angriest robot. Should you be scared? The Telegraph. http://www.telegraph.co.uk/men/the-filter/11600593/Artificial-intelligence-should-you-be-scared-of-angry-robots.html.

Johnson, N., G. Zhao, E. Hunsader, H. Qi, N. Johnson, J. Meng and B. Tivnan (2013). "Abrupt rise of new machine ecology beyond human response time." Scientific reports 3.

Joy, B. (April 2000). "Why the Future Doesn't Need Us." Wired Magazine 8(4).

Lieberman, H., A. Faaborg, W. Daher and J. Espinosa (2005). How to wreck a nice beach you sing calm incense. Proceedings of the 10th international conference on Intelligent user interfaces, ACM.

Loosemore, R. P. (2014). The Maverick Nanny with a Dopamine Drip: Debunking Fallacies in the Theory of AI Motivation. 2014 AAAI Spring Symposium Series.

Majot, A. M. and R. V. Yampolskiy (2014). AI safety engineering through introduction of self-reference into felicific calculus via artificial pain and pleasure. 2014 IEEE International Symposium on Ethics in Science, Technology and Engineering, IEEE.

Nijholt, A. (2011). "No grice: computers that lie, deceive and conceal."

Omohundro, S. (2012). Rational artificial intelligence for the greater good. Singularity Hypotheses, Springer: 161-179.

Özkural, E. (2014). "Godseed: Benevolent or Malevolent?" arXiv preprint arXiv:1402.5380.

Picard, R. W. and R. Picard (1997). Affective computing, MIT press Cambridge.

Russell, S., D. Dewey, M. Tegmark and e. al (January 23, 2015). Research Priorities for Robust and Benecial Artificial Intelligence. Future of Life Institute. http://futureoflife.org/static/data/documents/research_priorities.pdf.

Searle, J. R. (October 9, 2014). What Your Computer Can't Know. The New York Review of Books. http://www.nybooks.com/articles/archives/2014/oct/09/what-your-computer-cant-know.

Simonite, T. (March 7, 2008). Should every computer chip have a cosmic ray detector? . New Scientist. https://www.newscientist.com/blog/technology/2008/03/do-we-need-cosmic-ray-alerts-for.html.

Smith, D. (January 10, 2013). IBM's Watson Gets A 'Swear Filter' After Learning The Urban Dictionary. International Business Times. http://www.ibtimes.com/ibms-watson-gets-swear-filter-after-learning-urban-dictionary-1007734.

Soares, N. and B. Fallenstein (2014). Aligning Superintelligence with Human Interests: A Technical Research Agenda, Tech. rep. Machine Intelligence Research Institute, 2014. URL: http://intelligence. org/files/TechnicalAgenda.pdf.

Soares, N., B. Fallenstein, S. Armstrong and E. Yudkowsky (January 25-30, 2015). Corrigibility. Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence. Austin, Texas, USA.

Sotala, K. and R. V. Yampolskiy (2015). "Responses to catastrophic AGI risk: a survey." Physica Scripta 90(1): 018001.

Turchin, A. (July 10 2015a). A Map: AGI Failures Modes and Levels. LessWrong. http://lesswrong.com/lw/mgf/a_map_agi_failures_modes_and_le vels/.

Turchin, A. (July 10, 2015b). Human Extinction Risks due to Artificial Intelligence Development - 55 ways we can be obliterated. IEET. http://ieet.org/index.php/IEET/more/turchin20150610.

Turchin, A. (March 15, 2013). Risks of downloading alien AI via SETI search. LessWrong. http://lesswrong.com/lw/gzv/risks_of_downloading_alien_ai_via _seti_search/.

Waser, M. (2011). Rational Universal Benevolence: Simpler, Safer, and Wiser Than "Friendly AI". Artificial General Intelligence, Springer: 153-162.

Yampolskiy, R. (2012). "Leakproofing the Singularity Artificial Intelligence Confinement Problem." Journal of Consciousness Studies 19(1-2): 1-2.

Yampolskiy, R. and J. Fox (2012). "Safety Engineering for Artificial General Intelligence." Topoi: 1-10.

Yampolskiy, R. V. (2013). Artificial intelligence safety engineering: Why machine ethics is a wrong approach. Philosophy and Theory of Artificial Intelligence, Springer Berlin Heidelberg: 389-396.

Yampolskiy, R. V. (2014). "Utility Function Security in Artificially Intelligent Agents." Journal of Experimental and Theoretical Artificial Intelligence (JETAI): 1-17.

Yampolskiy, R. V. (2015). Analysis of types of self-improving software. Artificial General Intelligence, Springer: 384-393.

Yampolskiy, R. V. (2015). Artificial Superintelligence: a Futuristic Approach, Chapman and Hall/CRC.

Yampolskiy, R. V. (2015). On the limits of recursively self-improving AGI. Artificial General Intelligence, Springer: 394-403.

Yampolskiy, R. V. (2015). The Space of Possible Mind Designs. Artificial General Intelligence, Springer: 218-227.

Yampolskiy, R. V. (October 3-4, 2011). What to Do with the Singularity Paradox? Philosophy and Theory of Artificial Intelligence (PT-AI2011). Thessaloniki, Greece.

Yudkowsky, E. (2011). Complex Value Systems in Friendly AI. Artificial General Intelligence. J. Schmidhuber, K. Thórisson and M. Looks, Springer Berlin / Heidelberg. 6830: 388-393.

Yudkowsky, E. S. (May 2004). Coherent Extrapolated Volition. Available at: http://singinst.org/upload/CEV.html, Singularity Institute for Artificial Intelligence.