

# LessWrong DISCUSSION

MAIN ▾ | DISCUSSION ▾

WIKI | SEQUENCES | ABOUT

You're looking at Less Wrong's discussion board. This includes all posts, including those that haven't been promoted to the front page yet. For more information, see [About Less Wrong](#).

## AI Risk & Opportunity: A Timeline of Early Ideas and Arguments

**4** [lukeprog](#) 31 March 2012 02:34PM

Part of the series [AI Risk and Opportunity: A Strategic Analysis](#).

(You can leave anonymous feedback on posts in this series [here](#). I alone will read the comments, and may use them to improve past and forthcoming posts in this series.)

Building on [the previous post on AI risk history](#), this post provides an incomplete timeline (up to 1993) of significant *novel* ideas and arguments related to AI as a potential catastrophic risk. I do not include ideas and arguments concerning only, for example, the possibility of AI ([Turing 1950](#)) or attempts to predict its arrival ([Bostrom 1998](#)).

As is usually the case, we find that when we look closely at a cluster of ideas, it turns out these ideas did not appear all at once in the minds of a Few Great Men. Instead, they grew and mutated and gave birth to new ideas gradually as they passed from mind to mind over the course of many decades.

**1863: Machine intelligence as an existential risk to humanity; relinquishment of machine technology recommended.** Samuel Butler in [Darwin among the machines](#) worries that as we build increasingly sophisticated and autonomous machines, they will achieve greater capability than humans and replace humans as the dominant agents on the planet:

...we are ourselves creating our own successors; we are daily adding to the beauty and delicacy of their physical organisation; we are daily giving them greater power and supplying by all sorts of ingenious contrivances that self-regulating, self-acting power which will be to them what intellect has been to the human race. In the course of ages we shall find ourselves the inferior race... the time will come when the machines will hold the real supremacy over the world and its inhabitants...

Our opinion is that war to the death should be instantly proclaimed against them. Every machine of every sort should be destroyed by the well-wisher of his species. Let there be no exceptions made, no quarter shown...

(See also [Butler 1872](#); [Campbell 1932](#).)

**1921: Robots as an existential risk.** The Czech play [R.U.R.](#) by Karel Capek tells the story of robots which grow in power and intelligence and destroy the entire human race (except for a single survivor).

**1947: Fragility & complexity of human values (in the context of machine goal systems); perverse instantiation.** Jack Williamson's novelette [With Folded Hands](#) (1947) tells the story of a race of machines that, in order to follow the Prime Directive: "to serve and obey and guard men from harm." To obey this rule, the machines interfere with every aspect of human life, and humans who resist are lobotomized. Due to the fragility and complexity of human values ([Yudkowsky 2008](#); [Muehlhauser and Helm 2012](#)), the machines' rules of behavior had unintended consequences, manifesting a "perverse instantiation" in the language of Bostrom (forthcoming).

(Also see [Asimov 1950, 1957, 1983](#); [Versenyi 1974](#); [Minsky 1984](#); [Yudkowsky 2001, 2011](#).)

**1948-1949: Precursor idea to intelligence explosion.** Von Neumann ([1948](#)) wrote:

...“complication” on its lower levels is probably degenerative, that is, that every automaton that can produce other automata will only be able to produce less complicated ones. There is, however, a certain minimum level where this degenerative characteristic ceases to be universal. At this point automata which can reproduce themselves, or even construct higher entities, become possible.

Von Nuemann (1949) came very close to articulating the idea of intelligence explosion:

There is thus this completely decisive property of complexity, that there exists a critical size below which the process of synthesis is degenerative, but above which the phenomenon of synthesis, if properly arranged, can become explosive, in other words, where syntheses of automata can proceed in such a manner that each automaton will produce other automata which are more complex and of higher potentialities than itself.


Register / Login

Password

Remember me

Login

[Recover password](#)

 [Subscribe to RSS Feed](#)

### NEAREST MEETUPS

Brussels - We meet every month: 12 January 2019 02:00PM

Bi-weekly Frankfurt Meetup: 01 January 2019 07:30PM

### VIRTUAL STUDY ROOM

Co-work with other rationalists online.  
[Less Wrong Study Hall](#)

### RECENT COMMENTS

**Hmm... Easier to find stories**  
by [CronoDAS](#) on [rationalfiction.io](#) - publish, discover, and discuss rational fiction | 0 points

**It seems to me that it doesn't**  
by [gjm](#) on [Open Thread May 30 - June 5, 2016](#) | 0 points

**Not going to use it but: 1. Good job**  
by [root](#) on [rationalfiction.io](#) - publish, discover, and discuss rational fiction | 0 points

**There are a few links [on the wiki]**  
by [gjm](#) on [Open Thread May 30 - June 5, 2016](#) | 0 points

> **imagine that you are literally he**  
by [gjm](#) on [Open Thread May 30 - June 5, 2016](#) | 0 points

### RECENT POSTS

**rationalfiction.io - publish, discover, and discuss rational fiction**  
by [rayalez](#) | 0v (1c)

**Open Thread May 30 - June 5, 2016**  
by [Elo](#) | 3v (13c)

**Cognitive Biases Affecting Self-Perception of Beauty**  
by [Bound\\_up](#) | 2v (27c)

**When considering incentives, consider the incentives of all parties**  
by [casebash](#) | -5v (5c)

**How my something to protect just coalesced into being**  
by [Romashka](#) | 5v (1c)

**Anti-reductionism as**

**1951: Potentially rapid transition from machine intelligence to machine takeover.** Turing ([1951](#)) described ways that intelligent computers might learn and improve their capabilities, concluding that:

...it seems probable that once the machine thinking method has started, it would not take long to outstrip our feeble powers... At some stage therefore we should have to expect the machines to take control...

**1959: Intelligence explosion; the need for human-friendly goals for machine superintelligence.** Good ([1959](#)) describes what he later ([1965](#)) called an "intelligence explosion," a particular mechanism for rapid transition from artificial general intelligence to dangerous machine takeover:

Once a machine is designed that is good enough... it can be put to work designing an even better machine. At this point an "explosion" will clearly occur; all the problems of science and technology will be handed over to machines and it will no longer be necessary for people to work. Whether this will lead to a Utopia or to the extermination of the human race will depend on how the problem is handled by the machines. The important thing will be to give them the aim of serving human beings.

(Also see Good [1962](#), [1965](#), [1970](#); Vinge [1992](#), [1993](#); Yudkowsky [2008](#).)

**1966: A military arms race for machine superintelligence could accelerate machine takeover; convergence toward a singleton is likely.** Dennis Feltham Jones' 1966 novel *Colossus* depicted what may be a particularly likely scenario: two world superpowers (the USA and USSR) are in an arms race to develop superintelligent computers, one of which self-improves enough to take control of the planet.

In the same year, [Cade \(1966\)](#) argued the same thing:

political leaders on Earth will slowly come to realize... that intelligent machines having superhuman thinking ability can be built. The construction of such machines, even taking into account all the latest developments in computer technology, would call for a major national effort. It is only to be expected that any nation which did put forth the financial and physical effort needed to build and programme such a machine, would also attempt to utilize it to its maximum capacity, which implies that it would be used to make major decisions of national policy. Here is where the awful dilemma arises. Any restriction to the range of data supplied to the machine would limit its ability to make effective political and economic decisions, yet if no such restrictions are placed upon the machine's command of information, then the entire control of the nation would virtually be surrendered to the judgment of the robot.

On the other hand, any major nation which was led by a superior, unemotional intelligence of any kind, would quickly rise to a position of world domination. This by itself is sufficient to guarantee that, sooner or later, the effort to build such an intelligence will be made — if not in the Western world, then elsewhere, where people are more accustomed to iron dictatorships.

...It seems that, in the foreseeable future, the major nations of the world will have to face the alternative of surrendering national control to mechanical ministers, or being dominated by other nations which have already done this. Such a process will eventually lead to the domination of the whole Earth by a dictatorship of an unparalleled type — a single supreme central authority.

(This last paragraph also argues for convergence toward what Bostrom later called a "singleton.")

(Also see [Ellison 1967](#).)

**1970: Proposal for an association that analyzes the implications of machine superintelligence; naive control solutions like "switch off the power" may not work because the superintelligence will outsmart us, thus we must focus on its motivations; possibility of "pointless" optimization by machine superintelligence.** [Good \(1970\)](#) argues:

Even if the chance that the ultraintelligent machine will be available [soon] is small, the repercussions would be so enormous, good or bad, that it is not too early to entertain the possibility. In any case by 1980 I hope that the implications and the safeguards will have been thoroughly discussed, and this is my main reason for airing the matter: an association for considering it should be started.

(Also see [Bostrom 1997](#).)

On the idea that naive control solutions like "switch off the power" may not work because the superintelligence will find a way to outsmart us, and thus we must focus our efforts on the superintelligence's *motivations*, Good writes:

Some people have suggested that in order to prevent the [ultraintelligent machine] from taking over we should be ready to switch off its power supply. But it is not as simple as that because the machine could recommend the appointment of its own operators, it could recommend that they be paid well and it could select older men who would not be worried about losing their jobs. Then it could replace its operators by robots in order to make sure that it is not switched off. Next it could have the neo-Luddites ridiculed by calling them Ludditeniks, and if necessary it would later have them imprisoned or executed. This shows how careful we must be to keep our eye on the "motivation" of the machines, if possible, just as we should with politicians.

**complementary, rather than contradictory**

by [ImNotAsSmartAsIThink](#) | 2v (6c)

**Weekly LW Meetups**

by [FrankAdamek](#) | 1v (0c)

**LINK: Performing a Failure Autopsy**

by [fowlertm](#) | 1v (3c)

**LINK: Quora brainstorms strategies for containing AI risk**

by [Mass\\_Driver](#) | 4v (1c)

**Iterated Gambles and Expected Utility Theory**

by [Sable](#) | 1v (26c)

LATEST OPEN THREAD

**Help request. I am looking for an**

by [kitimat](#) on Open Thread May 30 - June 5, 2016 | 0 points

LATEST RATIONALITY DIARY

**I've been able to update my web**

by [WalterL](#) on Group Rationality Diary, February 2016 | 2 points

RECENT WIKI EDITS

RECENT ON RATIONALITY BLOGS

[Alexander](#) on Age of Em

[Ascended Economy?](#)

[Book Review : Age of Em](#)

[Rating Ems vs AIs](#)

[4 Age of Em Web Reviews](#)

TOP CONTRIBUTORS, 30 DAYS

[ingres](#) (253)

[Lumifer](#) (181)

[gwern](#) (55)

[username2](#) (49)

[Elo](#) (44)

[Viliam](#) (43)

[tanagrabeast](#) (40)

[ChristianKl](#) (37)

[Gram\\_Stone](#) (32)

[James\\_Miller](#) (32)

[time](#) (28)

[HungryHobo](#) (26)

[knb](#) (26)

[RyanCarey](#) (25)

[Stefan\\_Schubert](#) (24)

RECENT KARMA AWARDS

(Also see [Yudkowsky 2008](#).)

Good also outlines one possibility for "pointless" goal-optimization by machine superintelligence:

If the machines took over and men became redundant and ultimately extinct, the society of machines would continue in a complex and interesting manner, but it would all apparently be pointless because there would be no one there to be interested. If machines cannot be conscious there would be only a zombie world. This would perhaps not be as bad as in many human societies where most people have lived in misery and degradation while a few have lived in pomp and luxury. It seems to me that the utility of such societies has been negative (while in the condition described) whereas the utility of a zombie society would be zero and hence preferable.

(Also see [Bostrom 2004](#); [Yudkowsky 2008](#).)

**1974: We can't much predict what will happen after the creation of machine superintelligence.** Julius Lukasiwicz ([1974](#)) writes:

The survival of man may depend on the early construction of an ultraintelligent machine-or the ultraintelligent machine may take over and render the human race redundant or develop another form of life. The prospect that a merely intelligent man could ever attempt to predict the impact of an ultraintelligent device is of course unlikely but the temptation to speculate seems irresistible.

(Also see [Vinge 1993](#).)

**1977: Self-improving AI could stealthily take over the internet; convergent instrumental goals in AI; the treacherous turn.** Though the concept of a self-propagating computer worm was introduced by John Brunner's *The Shockwave Rider* (1975), Thomas J. Ryan's novel *The Adolescence of P-1* (1977) tells the story of an intelligent worm that at first is merely able to learn to hack novel computer systems and use them to propagate itself, but later (1) has novel insights on how to improve its own intelligence, (2) develops convergent instrumental subgoals (see [Bostrom 2012](#)) for self-preservation and resource acquisition, and (3) learns the ability to fake its own death so that it can grow its powers in secret and later engage in a "treacherous turn" (see Bostrom forthcoming) against humans.

**1982: To design ethical machine superintelligence, we may need to design superintelligence first and then ask it to solve philosophical problems (e.g. including ethics).**

Good ([1982](#)) writes:

Unfortunately, after 2500 years, the philosophical problems are nowhere near solution. Do we need to solve these philosophical problems before we can design an adequate ethical machine, or is there another approach? One approach that cannot be ruled out is first to produce an ultra-intelligent machine and then ask it to solve philosophical problems.

**1988: Even though AI poses an existential threat, we may need to rush toward it so we can use it to mitigate other existential threats.** Moravec ([1988](#), p. 100-101) writes:

...intelligent machines... threaten our existence... Machines merely as clever as human beings will have enormous advantages in competitive situations... So why rush headlong into an era of intelligent machines? The answer, I believe, is that we have very little choice, if our culture is to remain viable... The universe is one random event after another. Sooner or later an unstoppable virus deadly to humans will evolve, or a major asteroid will collide with the earth, or the sun will expand, or we will be invaded from the stars, or a black hole will swallow the galaxy. The bigger, more diverse, and competent a culture is, the better it can detect and deal with external dangers. The larger events happen less frequently. By growing rapidly enough, a culture has a finite chance of surviving forever.

**1993: Physical confinement is unlikely to constrain superintelligences, for superintelligences will outsmart us.** Vinge (1993) writes:

I argue that confinement [of superintelligent machines] is intrinsically impractical. For the case of physical confinement: Imagine yourself confined to your house with only limited data access to the outside, to your masters. If those masters thought at a rate — say — one million times slower than you, there is little doubt that over a period of years (your time) you could come up with "helpful advice" that would incidentally set you free...

**After 1993.** The [extropians mailing list](#) was launched in 1991, and was home to hundreds of discussions in which many important new ideas were proposed — ideas later developed in the public writings of Bostrom, Yudkowsky, Goertzel, and others. Unfortunately, the discussions from before 1998 were private, by agreement among subscribers. The early years of the archive cannot be made public without getting permission from everyone involved — a nearly impossible task. I have, however, collected all posts I could find from 1998 onward and uploaded them [here](#) (link fixed 04-03-2012).

I will end this post here. Perhaps in a future post I will extend the timeline past 1993, when interest in the subject became greater and thus the number of new ideas generated per decade rapidly increased.

## References

- Asimov (1950). [The Evitable Conflict](#)
- Asimov (1957). [The Naked Sun](#)
- Asimov (1983). [The Robots of Dawn](#)

- Bostrom (1997). [Predictions from Philosophy? How philosophers could make themselves useful](#)
- Bostrom (1998). [How Long before Superintelligence?](#)
- Bostrom (2004). [The Future of Human Evolution](#)
- Bostrom (2012). [The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents](#)
- Bostrom (forthcoming). [Superintelligence](#).
- Brunner (1975). [The Shockwave Rider](#)
- Butler (1863). [Darwin among the machines](#)
- Butler (1872). [Erewhon](#).
- Campbell (1932). [The Last Evolution](#)
- Capek (1921). [R.U.R.](#)
- Ellison (1967). [I Have No Mouth, and I Must Scream](#)
- Good (1959). [Speculations on perceptrons and other automata](#)
- Good (1962). [The social implications of artificial intelligence](#)
- Good (1965). [Speculations Concerning the First Ultrainelligent Machine](#)
- Good (1970). [Some future social repercussions of computers](#)
- Jones (1966). [Colossus](#).
- Lukasiewicz (1974). [The Ignorance Explosion](#).
- Minsky (1984). [Afterward to Vinge's 'True Names'](#).
- Moravec (1988). [Mind Children: The Future of Robot and Human Intelligence](#).
- Muehlhauser & Helm (2012). [The Singularity and Machine Ethics](#)
- Ryan (1977). [The Adolescence of P-1](#)
- Turing (1950). [Computing Machinery and Intelligence](#)
- Turing (1951). [Intelligent machinery, a heretical theory](#)
- Versenyi (1974). [Can robots be moral?](#)
- Vinge (1992). [A Fire Upon The Deep](#).
- Vinge (1993). [The Coming Technological Singularity](#).
- Von Neumann (1948). [The general and logical theory of automata](#).
- Von Neumann (1949). Theory and Organization of Complicated Automata. (Five lectures delivered at the University of Illinois in December, 1949. Reprinted in *Papers of John Von Neumann on Computers and Computing Theory*.)
- Williamson (1947). [With Folded Hands](#).
- Yudkowsky (2001). [Creating Friendly AI](#).
- Yudkowsky (2008). [Artificial Intelligence as a Positive and Negative Factor in Global Risk](#)
- Yudkowsky (2011). [Complex value systems are required to realize valuable futures](#)

#### Article Navigation

[Comments \(28\)](#)

Tags: [ai risk](#) [history](#)

### Comments (28)

Sort By: Best

**Wei\_Dai** 01 April 2012 05:43:14AM 7 points [\[-\]](#)

Luke, I'm not sure this post was worth your time to write. Most of us here probably do not have a strong interest in studying history for its own sake, and this scholarship about who first came up with various Singularity-related ideas and arguments doesn't seem to help very much in trying to figure out what we should do now to have the best chance of a positive Singularity. Perhaps I'm missing the point, in which case please fill me in...

**lukeprog** 01 April 2012 07:32:42PM 9 points [\[-\]](#)

Note that I'm making heavy use of my various research and editing assistants, so most of the time spent on these last two posts wasn't my time. Also:

1. Showing the history can give people a more accurate picture of how these ideas developed and how many people were involved, which can [reduce unwarranted hero worship](#), counteract the tendency to [revere the bearer of good info](#), and help us make more accurate predictions about, for example, how often people independently come to conclusions about AI risk or intelligence explosion or whatever.
2. There was always a remote chance I would encounter a good idea in the literature that I had not encountered elsewhere. (This did not turn out to be the case, however.)

**Vladimir\_Nesov** 01 April 2012 12:17:47PM \* 3 points [\[-\]](#)

I'm guessing this post is more of a side effect of getting acquainted with related publications, but it could serve to signal to the casual reader the existence of such publications going back many decades.

**Wei\_Dai** 08 April 2012 03:35:55AM \* 5 points [\[-\]](#)

Luke grabbed the 98-2003 Extropian archives from <http://www.lucifer.com/exi-lists/>. For some reason the robots.txt of that site disallows search engine indexing, so none of those posts show up in Google. I've created a mirror of the archives at <http://extropians.weidai.com/> (hosted for free on GitHub, repository at <https://github.com/weidai11/extropians/tree/gh-pages>), and will submit the site to Google for indexing. If anyone sees any reason not to do this, please let me know.

**ETA:** The archive also includes posts from 96 and 97. Why did Luke say that the pre-98 posts are supposed to be private? Anybody know what the deal is?

**lukeprog** 23 May 2012 01:22:07AM 0 points 

I probably got the year wrong.

**lukeprog** 27 May 2013 04:51:02AM \* 3 points 

This comment is a placeholder for [Nick Bostrom's](#) forthcoming book from Oxford University Press, *Superintelligence*.

I wrote this comment so I can link to it when referring to "Bostrom, forthcoming."

**Update:** The OUP page is [here](#).

**ciphergoth** 27 May 2013 08:49:24AM \* 6 points 

Timelines on superintelligence related matters are of course hard to predict, but any idea when we'll be able to buy this book?

UPDATE 2014-01-07: [OUP now say July 2014 \(estimated\)](#)

**lukeprog** 27 May 2013 09:16:31PM \* 4 points 


75% confidence interval: 10-01-2013 through 04-01-2014.

**ciphergoth** 28 May 2013 05:19:23AM 1 point 

That's as precise as I could possibly hope for - thanks!

**lukeprog** 02 September 2013 05:15:06AM 0 points 

My 75% confidence interval is now wider, extending from Mar. 2014 through Jun. 2015. :(

**Kaj\_Sotala** 04 September 2013 07:53:09AM 0 points 

Out of curiosity, is the delay due to extensive revisions to the draft, or due to the publisher being slow?

**lukeprog** 04 September 2013 03:50:27PM 0 points 

Most of it is publisher delay.

**ciphergoth** 02 September 2013 12:19:47PM 0 points 

That's a shame.

At Anna Salamon's prompting, I took part in a workshop after Winter Intelligence 2011 to discuss this book; do you think it would be rude if I asked Nick Bostrom for a draft?

[deleted] 28 May 2013 08:32:48PM 0 points 

(I was confused by the beginning of the interval for a while until I remembered that you Americans write the month first.)

**ciphergoth** 03 June 2013 07:03:50AM 7 points 

In the [One True Date Format](#), that's 2013-10-01 to 2014-04-01.

**lukeprog** 31 March 2012 03:01:47PM \* 1 point



From [Versenyi \(1974\)](#):

To solve [Norbert] Wiener's "slave paradox," inherent in our wanting to build machines with two diametrically opposed traits (independence *and* subservience, self-directed... rationality *and* the seeking of someone else's goals), we would have to construct robots not only with a formal prudential programming, but also with all our specific goals, purposes, and aspirations built into them so that they will not seek anything but these. But even if this type of programming could be coherent, it would require an almost infinite knowledge on our part to construct robots in this way. We could make robots perfectly safe only if we had... an exact knowledge of all our purposes, needs, desires, etc., not only in the present but in all future contingencies which might possibly arise...

This is what led some "roboticists" to propose that robots should be programmed not only prudentially... but also with an overriding semi-Kantian imperative. (For example, Asimov's first law of robotics, "A robot may not injure a human being or, through inaction, allow a human being to come to harm," overrides the second, "A robot must obey the orders given it by human beings," which in turn overrides the purely prudential third, "A robot must protect its own existence." The trouble with this type of programming... is that it will not work as long as we have not also built into the robot an almost infinite knowledge of what is in the long run and in any conceivable situation good or bad, beneficial or harmful to human beings.

It would seem then that the only way to make a free... prudential robot safe for men would be to make it not only morally isomorphic but also wholly identical in structure and programming with human beings. Built of the same organic materials and given the same neurophysiological, psychological, and rational makeup, our android would... know as well as any man what would help and harm human beings and could thus obey the Asimovian first law, the Golden Rule, or any similar directive at least as well as any men can.

Unfortunately such construction and programming, even if it were technologically possible, would severely limit the specific usefulness of the robot.

**Dmytry** 31 March 2012 04:59:34PM -2 points



Don't forget all the smart people whom are sceptical of foom or of the scary idea in general, i.e. vast majority of those even remotely qualified. You are creating a cherry picked set here.

**lukeprog** 31 March 2012 06:48:16PM 0 points



Who criticized FOOM or scary idea before 1993?

**Dmytry** 31 March 2012 06:50:17PM \* -2 points



Who criticized God before theists started promoting the notion? The heavily privileged hypotheses aren't criticized explicitly before they for some reason become popular. (the people just voice out different opinions). You can ask Hanson what he thinks about it. Or Ray Kurzweil .

**lukeprog** 31 March 2012 09:32:57PM 2 points



My timeline in this post ends in 1993, is all I'm saying. Let me know if you find criticism of FOOM or Scary Idea before 1993.

**Nisan** 31 March 2012 10:09:59PM 1 point



The C-haceks in both occurrences of Karel Čapek's name are broken.

**lukeprog** 01 April 2012 12:07:02AM \* 1 point



Interestingly, it looks fine in the editor. Anyway, I've fixed this by not trying: now I just use a good old-fashioned "C".

lukeprog 06 November 2013 03:24:45AM 0 points



Good (1951) hints toward intelligence explosion:

The threshold between a machine which was the intellectual inferior or superior of a man would probably be reached if the machine could do its own programming.

provocateur 31 March 2012 05:01:21PM \* 0 points



Since you are including works of fiction, I think Terminator (1984) is worth mentioning. This is what most people think of when it comes to AI risk.

By the way, my personal favorite, when it comes to AI doing what it wasn't intended to, would have to be Eagle Eye (2008). It's got everything: hard take-off and wireheading of sorts, second-guessing humans, decent acting.

lukeprog 31 March 2012 06:47:38PM 1 point



Which new important ideas were contributed by *Terminator* or *Eagle Eye* that were not previously contributed?

provocateur 31 March 2012 07:32:29PM \* 0 points



**SPOILER ALERT** don't read if you are yet to see Eagle Eye.

I doubt that the Terminator introduced any new important ideas. Its notability is like that of David Chalmers' recent paper, in bringing old ideas to the attention of the broader public.

Eagle Eye was spoofing its own sensors at some point. Again, not a novel idea per se, but pretty great for a movie. In the beginning of the movie, IIRC there was some Bayesian updating going on based on different sources of evidence.

lukeprog 31 March 2012 09:33:29PM \* 0 points



Yeah, so, those works aren't included because they didn't introduce any new important ideas I can think of.

lukeprog 01 May 2012 05:44:35PM 0 points



I will also note that von Neumann may have been the first to talk explicitly about the possibility of machines creating machines more intelligent than themselves, in a lecture delivered in 1948 and published in 1951:

We are all inclined to suspect in a vague way the existence of a concept of "complication." This concept and its putative properties have never been clearly formulated. We are, however, always tempted to assume that they will work in this way, When an automaton performs certain operations, they must be expected to be of a lower degree of complication than the automaton itself. In particular, if an automaton has the ability to construct another one, there must be a decrease in complication as we go from the parent to the construct. That is, if A can produce B, the A in some way must have contained a complete description of B. In order to make it effective, there must be, furthermore, various arrangements in A that see to it that this description is interpreted and that the constructive operation that it calls for are carried out, In this sense, it would therefore seem at a certain degenerating tendency must be expected, some decrease in complexity as one automaton makes another automaton.

Although this has some indefinite plausibility to it, it is in clear contradiction with the most obvious things that go on in nature. Organisms reproduce themselves, that is, they produce new organisms with no decrease in complexity. In addition, there are long periods of evolution during which the complexity is even increasing. Organisms are indirectly derived from others which had lower complexity.

...All these are very crude steps in the direction of a systematic theory of automata. They represent, in addition, only one particular direction. This is, as I indicated before, the direction towards forming a rigorous concept of what constitutes "complication." They illustrate that "complication" on its lower levels is probably degenerative, that is, that every automaton that can produce other automata will only be able to produce less complicated ones. There is, however, a certain minimum level where this degenerative characteristic ceases to be universal. At this point automata which can reproduce themselves, or even construct higher entities, become possible. This fact, that complication as well as organization, below a certain minimum level is degenerative, and beyond that level can become self-supporting and even increasing, will clearly play an important role in any future theory of the subject.

