MIRI
MACHINE INTELLIGENCE
RESEARCH INSTITUTE

AI Risk Bibliography 2012

Luke Muehlhauser
Machine Intelligence Research Institute

Muehlhauser, Luke. 2012. *AI Risk Bibliography 2012*.
The Singularity Institute, San Francisco, CA. Last modified May 29, 2013.

The Machine Intelligence Research Institute was previously known as the Singularity Institute.

Introduction

For the purposes of this bibliography, AI risk is defined as the risk of AI-related events that could end human civilization.

This bibliography contains 89 entries. Generally, only sources with an extended analysis of AI risk are included, though there are some exceptions among the earliest sources. Listed sources discuss either the likelihood of AI risk or they discuss possible solutions. (This does not include most of the "machine ethics" literature, unless an article discusses machine ethics in the explicit context of artificial intelligence as an existential risk.)

Where possible, I have included links to open-access web versions of the entries. Unconventionally, I have listed the references chronologically by year of publication rather than alphabetically by first author. In the age of digital distribution, it is easy enough to find a particular author's publications by searching for their name with an application's 'Find' tool. Given this, I believe the chronological listing will give a clearer picture of the ongoing development of the field.

I presume I have overlooked some literature. Please send suggestions for new entries, or updated links for existing entries, to luke@intelligence.org.

I extend my thanks to Seth Baum for inspiring this bibliography with his global catastrophic risks bibliography (http://sethbaum.com/research/gcr/bibliography.pdf), to Jonathan Wang for helping me assemble the bibliography, and to Malo Bourgon for formatting the references.

References

Butler, Samuel [Cellarius, pseud.]. 1863. "Darwin Among the Machines." *Christchurch Press,* June 13. http://www.nzetc.org/tm/scholarly/tei-ButFir-t1-g1-t1-g1-t4-body.html.

Good, Irving John. 1959. *Speculations on Perceptrons and Other Automata.* Research Lecture, RC-115. IBM, Yorktown Heights, New York, June 2. http://domino.research.ibm.com/library/cyberdig.nsf/papers/58DC4EA36A143C218525785E00502E30/$File/rc115.pdf.

Good, Irving John. 1965. "Speculations Concerning the First Ultraintelligent Machine." In *Advances in Computers,* edited by Franz L. Alt and Morris Rubinoff, 31–88. Vol. 6. New York: Academic Press. doi:10.1016/S0065-2458(08)60418-0.

Good, Irving John. 1970. "Some Future Social Repercussions of Computers." *International Journal of Environmental Studies* 1 (1–4): 67–79. doi:10.1080/00207237008709398.

Versenyi, Laszlo. 1974. "Can Robots be Moral?" *Ethics* 84 (3): 248–259. http://www.jstor.org/stable/2379958.

Good, Irving John. 1982. "Ethical Machines." In *Intelligent Systems: Practice and Perspective,* edited by J. E. Hayes, Donald Michie, and Y.-H. Pao, 555–560. Machine Intelligence 10. Chichester: Ellis Horwood.

Minsky, Marvin. 1984. "Afterword to Vernor Vinge's novel, 'True Names.'" Unpublished manuscript, October 1. Accessed December 31, 2012. http://web.media.mit.edu/~minsky/papers/TrueNames.Afterword.html.

Moravec, Hans P. 1988. *Mind Children: The Future of Robot and Human Intelligence.* Cambridge, MA: Harvard University Press.

Crevier, Daniel. 1993. "The Silicon Challengers in Our Future." Chap. 12 in *AI: The Tumultuous History of the Search for Artificial Intelligence.* New York: Basic Books.

Vinge, Vernor. 1993. "The Coming Technological Singularity: How to Survive in the Post-Human Era." In *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace,* 11–22. NASA Conference Publication 10129. NASA Lewis Research Center. http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940022855_1994022855.pdf.

Hanson, Robin. 1994. "If Uploads Come First: The Crack of a Future Dawn." *Extropy* 6 (2). http://hanson.gmu.edu/uploads.html.

Whitby, Blay. 1996. *Reflections on Artificial Intelligence: The Legal, Moral, and Ethical Dimensions.* Exeter, UK: Intellect Books.

Bostrom, Nick. 1997. "Predictions from Philosophy? How Philosophers Could Make Themselves Useful." Unpublished manuscript. Last revised September 19, 1998. `http://www.nickbostrom.com/old/predict.html`.

Gubrud, Mark Avrum. 1997. "Nanotechnology and International Security." Paper presented at the Fifth Foresight Conference on Molecular Nanotechnology, Palo Alto, CA, November 5–8. `http://www.foresight.org/Conferences/MNT05/Papers/Gubrud/`.

Hanson, Robin. 1998. "Economic Growth Given Machine Intelligence." Unpublished manuscript. Accessed May 15, 2013. `http://hanson.gmu.edu/aigrow.pdf`.

Warwick, Kevin. 1998. *In the Mind of the Machine: Breakthrough in Artificial Intelligence.* London: Arrow.

Moravec, Hans P. 1999. *Robot: Mere Machine to Transcendent Mind.* New York: Oxford University Press.

Joy, Bill. 2000. "Why the Future Doesn't Need Us." *Wired,* April. `http://www.wired.com/wired/archive/8.04/joy.html`.

6, Perri [David Ashworth]. 2001. "Ethics, Regulation and the New Artificial Intelligence, Part I: Accountability and Power." *Information, Communication & Society* 4 (2): 199–229. doi:`10.1080/713768525`.

Hibbard, Bill. 2001. "Super-Intelligent Machines." *ACM SIGGRAPH Computer Graphics* 35 (1): 13–15. `http://www.siggraph.org/publications/newsletter/issues/v35/v35n1.pdf`.

Yudkowsky, Eliezer. 2001. *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures.* The Singularity Institute, San Francisco, CA, June 15. `http://intelligence.org/files/CFAI.pdf`.

Bostrom, Nick. 2002. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology* 9. `http://www.jetpress.org/volume9/risks.html`.

Goertzel, Ben. 2002. "Thoughts on AI Morality." *Dynamical Psychology.* `http://www.goertzel.org/dynapsyc/2002/AIMorality.htm`.

Hibbard, Bill. 2002. *Super-Intelligent Machines.* New York: Kluwer Academic / Plenum.

Bostrom, Nick. 2003. "Ethical Issues in Advanced Artificial Intelligence." In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence,* edited by Iva Smit and George E. Lasker, 12–17. Vol. 2. Windsor, ON: International Institute for Advanced Studies in Systems Research / Cybernetics.

Georges, Thomas M. 2003. *Digital Soul: Intelligent Machines and Human Values.* Boulder, CO: Westview.

Bostrom, Nick. 2004. "The Future of Human Evolution." In *Two Hundred Years After Kant, Fifty Years After Turing,* edited by Charles Tandy, 339–371. Vol. 2. Death and Anti-Death. Palo Alto, CA: Ria University Press.

Goertzel, Ben. 2004. "Encouraging a Positive Transcension: Issues in Transhumanist Ethical Philosophy." *Dynamical Psychology.* `http://www.goertzel.org/dynapsyc/2004/PositiveTranscension.htm`.

Goertzel, Ben. 2004. "The All-Seeing A(I): Universal Mind Simulation as a Possible Path to Stably Benevolent Superhuman AI." *Dynamical Psychology.* `http://www.goertzel.org/dynapsyc/2004/AllSeeingAI.htm`.

Posner, Richard A. 2004. "What are the Catastrophic Risks, and How Catastrophic Are They?" Chap. 1 in *Catastrophe: Risk and Response.* New York: Oxford University Press.

Yudkowsky, Eliezer. 2004. *Coherent Extrapolated Volition.* The Singularity Institute, San Francisco, CA, May. `http://intelligence.org/files/CEV.pdf`.

de Garis, Hugo. 2005. *The Artilect War: Cosmists vs. Terrans: A Bitter Controversy Concerning Whether Humanity Should Build Godlike Massively Intelligent Machines.* Palm Springs, CA: ETC Publications.

Hibbard, Bill. 2005. "The Ethics and Politics of Super-Intelligent Machines." Unpublished manuscript, July. Microsoft Word file, accessed December 31, 2012. `https://sites.google.com/site/whibbard/g/SI_ethics_politics.doc`.

Kurzweil, Ray. 2005. "The Deeply Intertwined Promise and Peril of GNR." Chap. 8 in *The Singularity Is Near: When Humans Transcend Biology.* New York: Viking.

Armstrong, Stuart. 2007. "Chaining God: A Qualitative Approach to AI, Trust and Moral Systems." Unpublished manuscript, October 20. Accessed December 31, 2012. `http://www.neweuropeancentury.org/GodAI.pdf`.

Bugaj, Stephan Vladimir, and Ben Goertzel. 2007. "Five Ethical Imperatives and Their Implications for Human-AGI Interaction." *Dynamical Psychology.* `http://goertzel.org/dynapsyc/2007/Five_Ethical_Imperatives_svbedit.htm`.

Dietrich, Eric. 2007. "After The Humans Are Gone." *Philosophy Now,* May/June. http://philosophynow.org/issues/61/After_The_Humans_Are_Gone.

Hall, John Storrs. 2007. *Beyond AI: Creating the Conscience of the Machine.* Amherst, NY: Prometheus Books.

Hall, John Storrs. 2007. "Ethics for Artificial Intellects." In *Nanoethics: The Ethical and Social Implications of Nanotechnology,* edited by Fritz Allhoff, Patrick Lin, James Moor, John Weckert, and Mihail C. Roco, 339–352. Hoboken, NJ: John Wiley & Sons.

Hall, John Storrs. 2007. "Self-Improving AI: An Analysis." *Minds and Machines* 17 (3): 249–259. doi:10.1007/s11023-007-9065-3.

Omohundro, Stephen M. 2007. "The Nature of Self-Improving Artificial Intelligence." Paper presented at Singularity Summit 2007, San Francisco, CA, September 8–9. http://selfawaresystems.com/2007/10/05/paper-on-the-nature-of-self-improving-artificial-intelligence/.

Blake, Thomas. 2008. "Robot Ethics: Why 'Friendly AI' Won't Work." In *Proceedings of the Tenth International Conference ETHICOMP 2008: Living, Working and Learning Beyond Technology,* edited by Terrel Ward Bynum, Maria Carla Calzarossa, Ivo De Lotto, and Simon Rogerson. ISBN: 9788890286995.

Hall, John Storrs. 2008. "Engineering Utopia." In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference,* edited by Pei Wang, Ben Goertzel, and Stan Franklin, 460–467. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS.

Hanson, Robin. 2008. "Economics of the Singularity." *IEEE Spectrum* 45 (6): 45–50. doi:10.1109/MSPEC.2008.4531461.

Omohundro, Stephen M. 2008. "The Basic AI Drives." In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference,* edited by Pei Wang, Ben Goertzel, and Stan Franklin, 483–492. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS.

Yudkowsky, Eliezer. 2008. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks,* edited by Nick Bostrom and Milan M. Ćirković, 308–345. New York: Oxford University Press.

Freeman, Tim. 2009. "Using Compassion and Respect to Motivate an Artificial Intelligence." Unpublished manuscript, March 8. Accessed December 31, 2012. http://fungible.com/respect/paper.html.

Shulman, Carl. 2009. "Arms Control and Intelligence Explosions." Paper presented at the 7th European Conference on Computing and Philosophy (ECAP), Bellaterra, Spain, July 2–4.

Shulman, Carl, Henrik Jonsson, and Nick Tarleton. 2009. "Machine Ethics and Superintelligence." In *AP-CAP 2009: The Fifth Asia-Pacific Computing and Philosophy Conference, October 1st-2nd, University of Tokyo, Japan, Proceedings,* edited by Carson Reynolds and Alvaro Cassinelli, 95–97. AP-CAP 2009. `http://kant.k2.t.u-tokyo.ac.jp/ap-cap09/proceedings.pdf`.

Sotala, Kaj. 2009. "Evolved Altruism, Ethical Complexity, Anthropomorphic Trust: Three Factors Misleading Estimates of the Safety of Artificial General Intelligence." Paper presented at the 7th European Conference on Computing and Philosophy (ECAP), Bellaterra, Spain, July 2–4.

Wallach, Wendell, and Colin Allen. 2009. *Moral Machines: Teaching Robots Right from Wrong.* New York: Oxford University Press. doi:`10.1093/acprof:oso/9780195374049.001.0001`.

Waser, Mark R. 2009. "A Safe Ethical System for Intelligent Machines." In *Biologically Inspired Cognitive Architectures: Papers from the AAAI Fall Symposium,* edited by Alexei V. Samsonovich, 194–199. Technical Report, FS-09-01. AAAI Press, Menlo Park, CA. `http://aaai.org/ocs/index.php/FSS/FSS09/paper/view/934`.

Chalmers, David John. 2010. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* 17 (9–10): 7–65. `http://www.ingentaconnect.com/content/imp/jcs/2010/00000017/f0020009/art00001`.

Fox, Joshua, and Carl Shulman. 2010. "Superintelligence Does Not Imply Benevolence." In *ECAP10: VIII European Conference on Computing and Philosophy,* edited by Klaus Mainzer. Munich: Dr. Hut.

Geraci, Robert M. 2010. *Apocalyptic AI: Visions of Heaven in Robotics, Artificial Intelligence, and Virtual Reality.* New York: Oxford University Press. doi:`10.1093/acprof:oso/9780195393026.001.0001`.

Goertzel, Ben. 2010. "Coherent Aggregated Volition: A Method for Deriving Goal System Content for Advanced, Beneficial AGIs." *The Multiverse According to Ben* (blog), March 12. `http://multiverseaccordingtoben.blogspot.ca/2010/03/coherent-aggregated-volition-toward.html`.

Goertzel, Ben. 2010. "GOLEM: Toward an AGI Meta-Architecture Enabling Both Goal Preservation and Radical Self-Improvement." Unpublished manuscript, May 2. Accessed December 31, 2012. `http://goertzel.org/GOLEM.pdf`.

Kaas, Steven, Steve Rayhawk, Anna Salamon, and Peter Salamon. 2010. *Economic Implications of Software Minds.* The Singularity Institute, San Francisco, CA, August 10. http://intelligence.org/files/EconomicImplications.pdf.

McGinnis, John O. 2010. "Accelerating AI." *Northwestern University Law Review* 104 (3): 1253–1270. http://www.law.northwestern.edu/lawreview/v104/n3/1253/LR104n3McGinnis.pdf.

Russell, Stuart J., and Peter Norvig. 2010. "Philosophical Foundations." Chap. 26 in *Artificial Intelligence: A Modern Approach,* 3rd ed. Upper Saddle River, NJ: Prentice-Hall.

Shulman, Carl. 2010. *Omohundro's "Basic AI Drives" and Catastrophic Risks.* The Singularity Institute, San Francisco, CA. http://intelligence.org/files/BasicAIDrives.pdf.

Shulman, Carl. 2010. *Whole Brain Emulation and the Evolution of Superorganisms.* The Singularity Institute, San Francisco, CA. http://intelligence.org/files/WBE-Superorgs.pdf.

Sotala, Kaj. 2010. "From Mostly Harmless to Civilization-Threatening: Pathways to Dangerous Artificial Intelligences." In *ECAP10: VIII European Conference on Computing and Philosophy,* edited by Klaus Mainzer. Munich: Dr. Hut.

Tarleton, Nick. 2010. *Coherent Extrapolated Volition: A Meta-Level Approach to Machine Ethics.* The Singularity Institute, San Francisco, CA. http://intelligence.org/files/CEV-MachineEthics.pdf.

Waser, Mark R. 2010. "Designing a Safe Motivational System for Intelligent Machines." In *Artificial General Intelligence: Proceedings of the Third Conference on Artificial General Intelligence, AGI 2010, Lugano, Switzerland, March 5–8, 2010,* edited by Eric B. Baum, Marcus Hutter, and Emanuel Kitzelmann, 170–175. Advances in Intelligent Systems Research 10. Amsterdam: Atlantis. doi:10.2991/agi.2010.21.

Dewey, Daniel. 2011. "Learning What to Value." In *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3–6, 2011. Proceedings,* edited by Jürgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks, 309–314. Lecture Notes in Computer Science 6830. Berlin: Springer. doi:10.1007/978-3-642-22887-2_35.

Hall, John Storrs. 2011. "Ethics for Self-Improving Machines." In *Machine Ethics,* edited by Michael Anderson and Susan Leigh Anderson, 512–523. New York: Cambridge University Press.

Muehlhauser, Luke. 2011. “So You Want to Save the World.” Last revised March 2, 2012. http://lukeprog.com/SaveTheWorld.html.

Waser, Mark R. 2011. “Rational Universal Benevolence: Simpler, Safer, and Wiser than ‘Friendly AI.’” In *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3–6, 2011. Proceedings,* edited by Jürgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks, 153–162. Lecture Notes in Computer Science 6830. Berlin: Springer. doi:10.1007/978-3-642-22887-2_16.

Yudkowsky, Eliezer. 2011. “Complex Value Systems in Friendly AI.” In *Artificial General Intelligence: 4th International Conference, AGI 2011, Mountain View, CA, USA, August 3–6, 2011. Proceedings,* edited by Jürgen Schmidhuber, Kristinn R. Thórisson, and Moshe Looks, 388–393. Lecture Notes in Computer Science 6830. Berlin: Springer. doi:10.1007/978-3-642-22887-2_48.

Armstrong, Stuart, Anders Sandberg, and Nick Bostrom. 2012. “Thinking Inside the Box: Using and Controlling an Oracle AI.” *Minds and Machines.* doi:10.1007/s11023-012-9282-2.

Berglas, Anthony. 2012. “Artificial Intelligence Will Kill Our Grandchildren (Singularity).” Unpublished manuscript, draft 9, January. Accessed December 31, 2012. http://berglas.org/Articles/AIKillGrandchildren/AIKillGrandchildren.html.

Bostrom, Nick. 2012. “The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents.” In “Theory and Philosophy of AI,” edited by Vincent C. Müller. Special issue, *Minds and Machines* 22 (2): 71–85. doi:10.1007/s11023-012-9281-3.

Goertzel, Ben. 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. http://ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00006.

Hanson, Robin. 2012. “Meet the New Conflict, Same as the Old Conflict.” *Journal of Consciousness Studies* 19 (1–2): 119–125. http://www.ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00008.

Heylighen, Francis. 2012. “Brain in a Vat Cannot Break Out.” *Journal of Consciousness Studies* 19 (1–2): 126–142. http://www.ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00009.

Miller, James D. 2012. *Singularity Rising: Surviving and Thriving in a Smarter, Richer, and More Dangerous World.* Dallas, TX: BenBella Books.

Muehlhauser, Luke. 2012. "The Human's Hidden Utility Function (Maybe)." *Less Wrong* (blog), January 28. http://lesswrong.com/lw/9jh/the_humans_hidden_utility_function_maybe/.

Muehlhauser, Luke, and Louie Helm. 2012. "The Singularity and Machine Ethics." In *Singularity Hypotheses: A Scientific and Philosophical Assessment,* edited by Amnon Eden, Johnny Søraker, James H. Moor, and Eric Steinhart. The Frontiers Collection. Berlin: Springer.

Müller, Jonatas. 2012. "Ethics, Risks and Opportunities of Superintelligences." Unpublished manuscript. Accessed April 16, 2012. http://jonatasmuller.com/superintelligences.pdf.

Omohundro, Stephen M. 2012. "Rational Artificial Intelligence for the Greater Good." In *Singularity Hypotheses: A Scientific and Philosophical Assessment,* edited by Amnon Eden, Johnny Søraker, James H. Moor, and Eric Steinhart. The Frontiers Collection. Berlin: Springer.

Pearce, David. 2012. *The Biointelligence Explosion: How Recursively Self-Improving Organic Robots Will Modify Their Own Source Code and Bootstrap Our Way to Full-Spectrum Superintelligence.* BLTC Research, Brighton, UK. http://www.biointelligence-explosion.com/.

Sotala, Kaj. 2012. "Advantages of Artificial Intelligences, Uploads, and Digital Minds." *International Journal of Machine Consciousness* 4 (1): 275–291. doi:10.1142/S1793843012400161.

Tipler, Frank. 2012. "Inevitable Existence and Inevitable Goodness of the Singularity." *Journal of Consciousness Studies* 19 (1–2): 183–193. http://www.ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00013.

Yampolskiy, Roman V. 2012. "Leakproofing the Singularity: Artificial Intelligence Confinement Problem." *Journal of Consciousness Studies* 2012 (1–2): 194–214. http://www.ingentaconnect.com/content/imp/jcs/2012/00000019/F0020001/art00014.

Yampolskiy, Roman V., and Joshua Fox. 2012. "Artificial General Intelligence and the Human Mental Model." In *Singularity Hypotheses: A Scientific and Philosophical Assessment,* edited by Amnon Eden, Johnny Søraker, James H. Moor, and Eric Steinhart. The Frontiers Collection. Berlin: Springer.

Yampolskiy, Roman V., and Joshua Fox. 2012. "Safety Engineering for Artificial General Intelligence." *Topoi.* doi:`10.1007/s11245-012-9128-9`.

Muehlhauser, Luke. 2013. "Intelligence Explosion FAQ." The Machine Intelligence Research Institute. Accessed March 8, 2013. `http://intelligence.org/ie-faq/`.

Bostrom, Nick, and Eliezer Yudkowsky. Forthcoming. "The Ethics of Artificial Intelligence." In *Cambridge Handbook of Artificial Intelligence,* edited by Keith Frankish and William Ramsey. New York: Cambridge University Press.