Jürgen Schmidhuber

Philosophers & Futurists, Catch Up!

Response to The Singularity

Abstract: *Responding to Chalmers' The Singularity (2010), I argue that progress towards self-improving AIs is already substantially beyond what many futurists and philosophers are aware of. Instead of rehashing well-trodden topics of the previous millennium, let us start focusing on relevant new millennium results.*

All indented paragraphs of this paper are quotes taken from Chalmers' paper of 2010, who mentions Good's informal speculations (1965) on ultraintelligent self-improving machines:

> The key idea is that a machine that is more intelligent than humans will be better than humans at designing machines. So it will be capable of designing a machine more intelligent than the most intelligent machine that humans can design.

Chalmers speculates that some sort of meta-evolution could be used to build more and more intelligent machines called AI, AI+, AI++...:

> The process of evolution might count as an indirect example: less intelligent systems have the capacity to create more intelligent systems by reproduction, variation and natural selection. This version would then come to the same thing as an evolutionary path to AI and AI++. [...] If we produce an AI by machine learning, it is likely that soon after we will be able to improve the learning algorithm and extend the learning process, leading to AI+. If we produce an AI by artificial evolution, it is

Correspondence:
Jürgen Schmidhuber IDSIA, Galleria 2, 6928 Manno-Lugano, Switzerland
University of Lugano & SUPSI, Switzerland

Journal of Consciousness Studies, **19**, No. 1–2, 2012, pp. 173–82

> likely that soon after we will be able to improve the evolutionary algorithm and extend the evolutionary process, leading to AI+.

Back in 1987 I put forward the first concrete implementation of this informal idea (Schmidhuber, 1987): an evolutionary self-referential meta-learning problem solver that improves itself such that it becomes better at improving itself, by improving the very process of evolution, lifting Genetic Programming (Cramer, 1985; Dickmanns *et al.*, 1987) to the meta-level, the meta-meta-level, and so on, recursively. This was the first in a long string of papers on self-referential self-improvers — compare the recent overview (Schaul & Schmidhuber, 2010).

Roughly at the same time, two science fiction novels by Vernor Vinge (1984; 1986) introduced me to the concept of a *Technological Singularity*, in my opinion one of the few original ideas put forward by SF authors *after* the so-called Golden Age of SF in the 1950s and 60s. The basic idea is that *technological change accelerates exponentially such that within finite time it reaches a transcendent point beyond human comprehension.* True, in the early 1900s Teilhard de Chardin already predicted (from a more religious perspective) that the evolution of civilization will culminate in a transcendent *Omega point*, and Vinge himself pointed out (1993) that Stanislaw Ulam formulated similar thoughts in 1958:

> One conversation centered on the ever accelerating progress of technology and changes in the mode of human life, which gives the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue.

It was Vinge, however, who popularized the technological singularity and significantly elaborated on it, exploring pretty much all the obvious related topics, such as accelerating change, computational speed explosion, potential delays of the singularity, obstacles to the singularity, limits of predictability and negotiability of the singularity, evil vs benign super-intelligence, surviving the singularity, etc.

I am not aware of substantial additional non-trivial ideas in this vein originating in the subsequent two decades, although other futurists and philosophers have started writing about the singularity as well. Many of them, however, are mostly concerned with ancient debates triggered by non-experts such as Lucas, Searle, Penrose and other authors commonly ignored by hardcore AI researchers:

> Various existing forms of resistance to AI take each of these forms. For example, J.R. Lucas (1961) has argued that for reasons tied to Gödel’s theorem, humans are more sophisticated than any machine.

Instead of further spending time on such frequently refuted claims, I'd like to encourage futurists and philosophers to learn about the more recent, in my opinion much more relevant hardcore AI research outlined in the remainder of this paper.

> Perhaps the core sense of the term [singularity], though, is a moderate sense in which it refers to an intelligence explosion through the recursive mechanism set out by I.J. Good [...] I will always use the term singularity in this core sense in what follows. [...] The argument depends on the assumption that there is such a thing as intelligence and that it can be compared between systems. (See also pp. 25ff.)

The scientific way of measuring intelligence involves measuring problem solving capacity. There are mathematically sound ways of doing this, using basic concepts of theoretical computer science (Levin, 1973; Hutter, 2005; Schmidhuber, 2009b), all of them avoiding the subjectivity of the ancient and popular but scientifically not very useful Turing test, which essentially says 'intelligent is what I feel is intelligent.'

> My own view is that the history of artificial intelligence suggests that the biggest bottleneck on the path to AI is software, not hardware: we have to find the right algorithms, and no-one has come close to finding them yet. [...] The Gödel Machines of Schmidhuber (2003) provide a theoretical example of self-improving systems at a level below AI, though they have not yet been implemented and there are large practical obstacles to using them as a path to AI.

I feel that philosophers and futurists should try to become very familiar with what is currently going on in the field of universal problem solvers. The fully self-referential (Gödel, 1931) Gödel machine (Schmidhuber, 2009b) already *is* a universal AI that is at least theoretically optimal in a certain sense. It may interact with some initially unknown, partially observable environment to maximize future expected utility or reward by solving arbitrary user-defined computational tasks. Its initial algorithm is not hardwired; it can completely rewrite itself without essential limits apart from the limits of computability, provided a proof searcher embedded within the initial algorithm can first prove that the rewrite is useful, according to the formalized utility function taking into account the limited computational resources. Self-rewrites may modify/improve the proof searcher itself, and can be shown to be *globally optimal*, relative to Gödel's well-known fundamental restrictions of provability (Gödel, 1931). To make sure the Gödel machine is at least *asymptotically* optimal even before the first self-rewrite, we may initialize it by Hutter's non-self-referential but *asymptotically fastest algorithm for all well-defined problems* HSEARCH (Hutter, 2002), which uses a hardwired brute force proof searcher and (justifiably) ignores the costs of proof search. Assuming discrete input/output domains $X/Y \subset B^*$, a formal problem specification $f : X \rightarrow Y$ (say, a functional description of how integers are decomposed into their prime factors), and a particular $x \in X$ (say, an integer to be factorized), HSEARCH orders all proofs of an appropriate axiomatic system by size to find programs q that for all $z \in X$ provably compute $f(z)$ within time bound $t_q(z)$. Simultaneously it spends most of its time on executing the q with the best currently proven time bound $t_q(x)$. Remarkably, HSEARCH is as fast as the *fastest* algorithm that provably computes $f(z)$ for all $z \in X$, save for a constant factor smaller than $1+ \varepsilon$ (arbitrary real- valued $\varepsilon > 0$) and an f-specific but x-independent additive constant (*ibid.*). Given some problem, the Gödel machine may decide to replace its HSEARCH initialization by a faster method suffering less from large constant overhead, but even if it doesn't, its performance won't be less than asymptotically optimal.

All of this implies that there already exists the blueprint of a Universal AI which will solve almost all problems almost as quickly as if it already knew the best (unknown) algorithm for solving them, because almost all imaginable problems are big enough to make the additive constant negligible. Hence I must object to Chalmers' statement *'we have to find the right algorithms, and no-one has come close to finding them yet'*. The only motivation for *not* quitting computer science research right now is that many real-world problems are so small and simple that the ominous constant slowdown (potentially relevant at least before the first Gödel machine self-rewrite) is *not* negligible. Nevertheless, the ongoing efforts at scaling universal AIs down to the rather few *small* problems are very much informed by the new millennium's theoretical insights mentioned above, and may soon yield practically feasible yet still general problem solvers for physical systems with highly restricted computational power, say, a few trillion instructions per second, roughly comparable to a human brain power.

Simultaneously, our non-universal but still rather general fast deep/recurrent neural networks have already started to outperform traditional pre-programmed methods: they recently collected a string of 1st ranks in many important visual pattern recognition benchmarks, e.g. Graves & Schmidhuber (2009); Ciresan *et al.* (2011): IJCNN traffic sign competition, NORB, CIFAR10, MNIST, three ICDAR handwriting competitions. Here we greatly profit from recent advances in computing hardware, using GPUs (mini-supercomputers normally used for video games) 100 times faster than today's CPU cores, and a million times faster than PCs of 20 years ago, complementing the recent above-mentioned progress in the theory of mathematically optimal universal problem solvers.

> In principle there could be an intelligence explosion without a speed explosion and a speed explosion without an intelligence explosion.

As pointed out above, problem solving ability does depend on speed, hence intelligence and speed are *not* independent. Computer scientists agree, however, that we are far from the physical limits to computation:

> While the laws of physics and the principles of computation may impose limits on the sort of intelligence that is possible in our world, there is little reason to think that human cognition is close to approaching those limits.

In fact, more than 100 additional years of Moore's Law seem necessary to reach Bremermann's (1982) physical limit of more than 10^{51} elementary instructions per kg and second, roughly 10^{20} times the combined raw computational power of all human brains, give or take a few zeroes.

> The history of AI involves a long series of optimistic predictions by those who pioneer a method, followed by a periods of disappointment and reassessment. This is true for a variety of methods involving direct programming, machine learning, and artificial evolution, for example. Many of the optimistic predictions were not obviously unreasonable at the time, so their failure should lead us to reassess our prior beliefs in significant ways.

I feel that after 10,000 years of civilization there is no need to justify pessimism through comparatively recent over-optimistic and self-serving predictions (1960s: 'only 10 instead of 100 years needed to build AIs') by a few early AI enthusiasts in search of funding.

> If we value scientific progress, for example, it makes sense for us to create AI and AI+ systems that also value scientific progress.

But how to formalize this informal idea? Only recently this has become possible through the *Formal Theory of Creativity* (Schmidhuber, 2006; 2010) mathematically concretizing the driving forces and value functions behind creative behavior such as science and art. Consider an agent living in an initially unknown environment. At any given time, it uses one of the many reinforcement learning (RL) methods (Kaelbling *et al.*, 1996) to maximize not only expected future external reward for achieving certain goals, such as avoiding hunger/

empty batteries/obstacles, etc. but also *intrinsic* reward for action sequences that improve an internal model of the environmental responses to its actions, continually learning to better predict/explain/compress the growing history of observations infiuenced by its experiments, actively infiuencing the input stream such that it contains previously unknown but learnable algorithmic regularities which become known and boring once there is no additional subjective *compression progress* or *learning progress* any more. I have argued that the particular utility functions associated with this theory explain essential aspects of intelligence including selective attention, curiosity, creativity, science, art, music, humor, e.g. Schmidhuber (2006; 2010). They are currently being implemented on humanoid baby-like iCub robots. The theory actually addresses the above-mentioned drawbacks of asymptotically optimal universal AIs, allowing learning agents to not only focus on potentially hard-to-solve externally posed tasks, but also creatively invent self-generated tasks that have the property of currently being still unsolvable but easily learnable, given the agent's present knowledge, such that the agent is continually motivated to improve its understanding of how the world works, and what can be done in it. One topic worth of exploration through futurists and philosophers are the potential consequences of self-improving AIs defining their own tasks in this creative, world-exploring way, which may sometimes conflict with goals of humans.

> If we create an AI through learning or evolution, the matter is more complex. [...] Of course even if we create an AI or AI+ (whether human-based or not) with values that we approve of, that is no guarantee that those values will be preserved all the way to AI++.

Note that the Gödel machine mentioned above *does* preserve those values. It can rewrite its utility function only if it first can prove that the rewrite is useful according to its previous utility function.

All attempts at making sure there will be only provably friendly AIs seem doomed though. Once somebody posts the recipe for practically feasible self-improving Gödel machines or AIs in form of code into which one can plug arbitrary utility functions, many users will equip such AIs with many different goals, often at least partially conflicting with those of humans. The laws of physics and the availability of physical resources will eventually determine which utility functions will help their AIs more than others to multiply and become dominant in competition with AIs driven by different utility functions. The survivors will define in hindsight what's 'moral', since only survivors promote their values, giving evolutionary meaning to Kant's musings:

> Kant held more specifically that rationality correlates with morality: a fully rational system will be fully moral as well [...] The Kantian view at least raises the possibility that intelligence and value are not entirely independent.

Chalmers writes on AIs in virtual worlds:

> It remains possible that they might build computers in their world and design AI on those computers. [...] If one takes seriously the possibility that we are ourselves in such a simulation (as I do in Chalmers, 2005) [...]

Compare the original papers since 1997 (Schmidhuber, 1997; 2000) that introduced and discussed the set of all computable universes as well as the set of possible computable probability distributions on them, extending Konrad Zuse's (1970) pioneering work on digital physics and the computable universe, using algorithmic probability theory (Solomonoff, 1964; 1978; Li & Vitányi, 1997) to analyse the probability of some observer inhabiting a particular 'simulated' or real universe, given his observations.

Uploading brains into cyberspace (Chalmers, 2010, pp. 41ff.) as well as related topics were discussed not only in the cited works of SF author Egan but also in Gibson's earlier famous cyberspace novels of the 1980s, and possibly first by Daniel F. Galuye (1964) who already went far in exploring the consequences. Gradual uploading is an old concept, too:

> Suppose that 1% of Daves brain is replaced by a functionally isomorphic silicon circuit. Next suppose that after one month another 1% is replaced, and the following month another 1%. We can continue the process for 100 months, after which a wholly uploaded system will result.

I think I first read about this thought experiment in Pylyshyn's (1980) paper. Chalmers also writes on consciousness (p. 44):

> It is true that we have no idea how a nonbiological system, such as a silicon computational system, could be conscious.

But at least we have pretty good ideas where the symbols and self-symbols underlying consciousness and sentience come from (Schmidhuber, 2009a; 2010). They may be viewed as simple by-products of data compression and problem solving. As we interact with the world to achieve goals, we are constructing internal models of the world, predicting and thus partially compressing the data histories we are observing. If the predictor/compressor is an artificial recurrent neural network (RNN) (Werbos, 1988; Williams & Zipser, 1994;

Schmidhuber, 1992; Hochreiter & Schmidhuber, 1997; Graves & Schmidhuber, 2009), it will create feature hierarchies, lower level neurons corresponding to simple feature detectors similar to those found in human brains, higher layer neurons typically corresponding to more abstract features, but fine-grained where necessary. Like any good compressor the RNN will learn to identify shared regularities among different already existing internal data structures, and generate prototype encodings (across neuron populations) or *symbols* for frequently occurring observation sub-sequences, to shrink the storage space needed for the whole. Self-symbols may be viewed as a by-product of this, since there is one thing that is involved in all actions and sensory inputs of the agent, namely, the agent itself. To efficiently encode the entire data history, it will profit from creating some sort of internal prototype symbol or code (e. g. a neural activity pattern) representing itself (Schmidhuber, 2009a; 2010). Whenever this representation becomes activated above a certain threshold, say, by activating the corresponding neurons through new incoming sensory inputs or an internal ‘search light’ or otherwise, the agent could be called self-aware. No need to see this as a mysterious process — it is just a natural by-product of partially compressing the observation history by efficiently encoding frequent observations.

Note that the mathematically optimal general problem solvers and universal AIs discussed above do *not at all* require something like an explicit concept of consciousness. This is one more reason to consider consciousness a possible but non-essential by-product of general intelligence, as opposed to a pre-condition.

Conclusion

Instead of elaborating on worn-out singularity-related topics already dealt with ad nauseam in the previous millennium, perhaps philosophers and futurists should catch up with new millennium results on theoretically optimal universal and creative AIs, and try to analyse their consequences.

References

Bremermann, H.J. (1982) Minimum energy requirements of information transfer and computing, *International Journal of Theoretical Physics*, **21**, pp. 203–217.

Chalmers, D.J. (2010) The singularity: A philosophical analysis, *Journal of Consciousness Studies*, **17** (9–10), pp. 7–65.

Ciresan, D.C., Meier, U., Masci, J., Gambardella, L.M. & Schmidhuber, J. (2011) Flexible, High Performance Convolutional Neural Networks for Image Classification. International Joint Conference on Artificial Intelligence (IJCAI-2011, Barcelona), 2011.

Cramer, N.L. (1985) A representation for the adaptive generation of simple sequential programs, in Grefenstette, J.J. (ed.) *Proceedings of an International Conference on Genetic Algorithms and Their Applications*, Carnegie-Mellon University, 24–26 July, Hillsdale, NJ: Lawrence Erlbaum Associates.

Dickmanns, D., Schmidhuber, J. & Winklhofer, A. (1987) Der genetische algorithmus: Eine Implementierung in Prolog, *Fortgeschrittenenpraktikum*, Institut für Informatik, Lehrstuhl Prof. Radig, Technische Universität München, [Online], http://www.idsia.ch/˜juergen/geneticprogramming.html

Galouye, D.F. (1964) *Simulacron 3*, New York: Bantam Books.

Gödel, K. (1931) Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I, *Monatshefte für Mathematik und Physik*, **38**,pp. 173–198.

Graves, A. & Schmidhuber, J. (2009) Offiine handwriting recognition with multidimensional recurrent neural networks, in *Advances in Neural Information Processing Systems 21*, Cambridge, MA: MIT Press.

Hochreiter, S. & Schmidhuber, J. (1997) Long short-term memory, *Neural Computation*, **9** (8), pp. 1735–1780.

Hutter, M. (2002) The fastest and shortest algorithm for all well-defined problems, *International Journal of Foundations of Computer Science*, **13** (3), pp. 431–443. On J. Schmidhuber's SNF grant 20-61847.

Hutter, M. (2005) *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*, Berlin: Springer. On J. Schmidhuber's SNF grant 20-61847.

Kaelbling, L.P., Littman, M.L. & Moore, A.W. (1996) Reinforcement learning: A survey, *Journal of AI Research*, **4**, pp. 237–285.

Levin, L.A. (1973) Universal sequential search problems, *Problems of Information Transmission*, **9** (3), pp. 265–266.

Li, M. & Vitányi, P.M.B. (1997) *An Introduction to Kolmogorov Complexity and its Applications*, 2nd ed., Berlin: Springer.

Pylyshyn, Z.W. (1980) Computation and cognition: Issues in the foundation of cognitive science, *Behavioral and Brain Sciences*, **3**, pp. 111–132.

Schaul, T. & Schmidhuber, J. (2010) Metalearning, *Scholarpedia*, **6** (5), p. 4650.

Schmidhuber, J. (1987) Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-… hook, *Institut für Informatik*, Technische Universität München, [Online], http://www.idsia.ch/˜juergen/diploma.html

Schmidhuber, J. (1992) A fixed size storage O(n^3) time complexity learning algorithm for fully recurrent continually running networks, *Neural Computation*, **4** (2), pp. 243–248.

Schmidhuber, J. (1997) A computer scientist's view of life, the universe, and everything, in Freksa, C., Jantzen, M. & Valk, R. (eds.) *Foundations of Computer Science: Potential -Theory -Cognition*, vol. 1337, pp. 201–208, Lecture Notes in Computer Science, Berlin: Springer.

Schmidhuber, J. (2000) Algorithmic theories of everything, *Technical Report IDSIA-20-00, quantph/0011122, IDSIA*, Manno (Lugano), Switzerland.

Schmidhuber, J. (2002) Hierarchies of generalized Kolmogorov complexities and nonenumerable universal measures computable in the limit, *International Journal of Foundations of Computer Science*, **13** (4), pp. 587–612.

Schmidhuber, J. (2002) The Speed Prior: A new simplicity measure yielding near-optimal computable predictions, in Kivinen, J. & Sloan, R.H. (eds.) *Proceedings of the 15th Annual Conference on Computational Learning Theory (COLT 2002)*, Lecture Notes in Artificial Intelligence, pp. 216–228, Sydney: Springer.

Schmidhuber, J. (2006) Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts, *Connection Science*, **18** (2), pp. 173–187.

Schmidhuber, J. (2009a) Simple algorithmic theory of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes, *SICE Journal of the Society of Instrument and Control Engineers*, **48** (1), pp. 21–32.

Schmidhuber, J. (2009b) Ultimate cognition à la Gödel, *Cognitive Computation*, **1** (2), pp. 177–193.

Schmidhuber, J. (2010) Formal theory of creativity, fun, and intrinsic motivation (1990–2010), *IEEE Transactions on Autonomous Mental Development*, **2** (3), pp. 230–247.

Solomonoff, R.J. (1964) A formal theory of inductive inference: Part I, *Information and Control*, **7**, pp. 1–22.

Solomonoff, R.J. (1978) Complexity-based induction systems, *IEEE Transactions on Information Theory*, **24** (5), pp. 422–432.

Vinge, V. (1984) *The Peace War*, New York: Bluejay Books Inc.

Vinge, V. (1986) *Marooned in Real Time*, New York: Bluejay Books Inc.

Vinge, V. (1993) The coming technological singularity, *VISION-21 Symposium sponsored by NASA Lewis Research Center*, and *Whole Earth Review*, Winter issue.

Werbos, P.J. (1988) Generalization of backpropagation with application to a recurrent gas market model, *Neural Networks*, **1**.

Williams, R.J. & Zipser, D. (1994) Gradient-based learning algorithms for recurrent networks and their computational complexity, in Chauvin, Y. & Rumelhart, D.E. (eds.) *Back-propagation: Theory, Architectures and Applications*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Zuse, K. (1970) *Rechnender Raum*, Braunschweig: Friedrich Vieweg & Sohn, 1969. English trans. *Calculating Space*, MIT Technical Translation AZT-70-164-GEMIT, Massachusetts Institute of Technology (Proj. MAC), Cambridge, MA.