MIRI
MACHINE INTELLIGENCE
RESEARCH INSTITUTE

Changing the Frame of AI Futurism: From Storytelling to Heavy-Tailed, High-Dimensional Probability Distributions

Steve Rayhawk, Anna Salamon, Michael Anissimov
Machine Intelligence Research Institute

Thomas McCabe, Rolf Nelson
MIRI Visiting Fellows

Rayhawk, Stephen, Anna Salamon, Thomas McCabe, Michael Anissimov, and Rolf Nelson. 2009. "Changing the Frame of AI Futurism: From Storytelling to Heavy-Tailed, High-Dimensional Probability Distributions." Paper presented at the 7th European Conference on Computing and Philosophy (ECAP), Bellaterra, Spain, July 2–4.

This version contains minor changes.

We introduce an interactive web application, *The Uncertain Future* (Uncertain Future 2009), that uses structured probabilistic models to help users think through possible timelines for strong artificial intelligence. To date, there have been few to no efforts to approach the full, multifaceted problem of forecasting the potential development of strong AI using the best formal tools available. There has been serious forecasting work on individual AI-related aspects of the world, such as the future cost of computing power (Anderson et al. 2002), the development of computer chess players (Kurzweil 1990), or the economic impact of robotic systems that substitute for human labor (reviewed in Peláez and Kyriakou [2008]). Long-range forecasts and models have generally had a narrow focus, such as trend-extrapolation models of "accelerating change" (Kurzweil 2005) or analyses of economic and power dynamics given strong AI (Hall 2008; Yudkowsky 2008). *The Uncertain Future* is an early attempt toward presenting a single, combined model that integrates our best estimates about each of the factors and their possible causal interactions over time—including a formal probabilistic treatment of our uncertainties.

The present gap in modeling the trajectory of AI matters. A number of analysts have argued that: (a) there is a substantial chance that "human-level" AI will be developed during this century; (b) human-level AI would have an impact at least comparable to such historical events as the appearance of sexual recombination, the oxygen transition, human language, or the industrial revolution (Kurzweil 2005; Vinge 1993; Bostrom 2003; Sandberg and Bostrom 2008; Hanson 1998). If it is correct to assign any non-negligible probability that both propositions are true, then it is important to use the best available tools to model the relationship between near-term policy decisions and the possible outcomes (Matheny 2007).

Moreover, strong AI has several features which can be expected to limit the effectiveness of both qualitative scenario analysis by single experts and quantitative trend extrapolation. Prediction around strong AI involves unprecedented phenomena that are difficult to visualize, variables which can take on a wide range of values, large potential impacts that can create emotional biases in both individual judgment and community discourse, and a long timeline during which several background variables relevant to AI development can interact in unexpected ways. Simple quantitative trend extrapolation based on historical data may very easily break because of changes in context from the relevant periods. Detailed qualitative scenario analysis, meanwhile, faces two challenges. First, the many variables involved demand the consideration of great numbers of scenarios to capture the space of plausible outcomes, while the best-known futurists focus on only one. Second, even if the attempt to evaluate other scenarios is made, psychological research in heuristics and biases indicates that in complex domains with large unknowns, even domain experts will tend to attach excessive confidence to specific, easily

visualizable scenarios (Tversky and Kahneman 1983). We do in fact see much published AI futurism confidently proclaiming the likelihood of specific future scenarios in cases where others confidently disagree. In policymaking, the characteristic result is neglect of the broad "everything else" category of events that could blindside us.

The Uncertain Future project is an experimental attempt at avoiding these pitfalls. The project has two faces:

1. As a future-projection tool, *The Uncertain Future* generates probability distributions over scenarios using the formalism of continuous-time Bayes nets (Nodelman 2007). We freeze in a particular Bayes net model structure and use experts' impressions to choose the parameters. This approach is similar to, but easier to make principled extensions within than, Trend Impact Analysis (Agami et al. 2008) or Cross Impact Analysis (Asan and Asan 2007).

2. As an educational tool, *The Uncertain Future* project allows individuals to enter their own beliefs for each parameter (in place of the experts' impressions) and to see the implications of their own causal beliefs, i.e. "the Socratic method meets modern probabilistic reasoning."

Our system's key features:

Separated belief-components: *The Uncertain Future* breaks participants' beliefs about AI timelines into a number of relatively independent components. For example, it requires participants to separately specify their probability distributions for how long Moore's law will continue, for the amount of computation required to model the brain, and for the possibility of nuclear war or other major societal disruption.

This helps participants focus separately on each major component of the world, including several background variables that might affect AI development and might not be part of participants' ordinary views of the future (e.g., nuclear and other major disruptions, or intelligence augmentation of a sort that speeds science).

Probabilities, not "most likely" events: Participants enter each belief visually, with a simple point-and-click interface for specifying their probability distribution. All beliefs are entered as probability distributions; even if participants think a particular parameter value or narrow range of values "most likely", they still must enter how likely, so that non-"mainline" sequences of events can be included in their picture. While this is standard practice in many areas of forecasting, it is not common in long-range AI futurism; for example, Kurzweil (2005) outlines a specific range of future predictions, including timelines of AI development, but does not attach probabilities to the predicted ranges.

The combined use of probabilities and of separated belief-components should help participants move from single, easily visualized storylines about "how the future will go" to the broad range of scenarios in which one or more variables may turn in an unexpected direction. For many users, the user's "mainline" scenario track turns out to have only a minority of their total probability; compounding of multiple probability distributions causes a wide range of future outcomes to emerge as a natural consequence.

Collated access to expert opinions, and to the belief-components of other participants: If a user, Bob, wants to think through AI futures, he can incorporate the risk of nuclear disruption without himself being knowledgeable about nuclear risks. Next to his probability-distribution entry box, he'll find a list of relevant experts' views on the size of nuclear risks over the relevant time-period; Bob can defer to expert consensus on this issue (which perhaps is not his specialty) and can then go on to enter his own, more thought-out parameter-values for belief-components for which he has background enough to reasonably disagree with the consensus.

Also, if Bob disagrees with Jane about AI, they can isolate the belief-components that underlie their disagreement and address them in particular. Science has made progress largely by reducing large, important problems to smaller and more manageable components that can be addressed with specific data and models; systems such as *The Uncertain Future* can help us to move between the sub-problems and complex whole.

The Uncertain Future is an early trial project, for which many simplifications were made. In the medium term we would like to capture additional parameters and effects, by building a platform for modular collaboration on futurist scenario projection and model-building. To this end, we wish to note that many of the probabilistic and quantitative methods used in futurism, including both trend extrapolation from historical data and strategic projection of future policies, can be naturally understood in a principled fashion as special cases or approximations within the formalism of continuous-time Bayes nets containing decision nodes. (This formalism is the unifying generalization of dynamic and continuous-time Bayes nets [Nodelman 2007], continuous-time and partially observable Markov decision processes, Bayes decision nets, stochastic differential equations [Särkkä 2006], and control theory, and an important special case of differential game theory.)

For example, both the historical curve-fitting used in the "surprise-free future" phase of trend impact analysis and the perturbations used in the "impact" phase can be understood together as an approximation to Bayesian inference under a stochastic-differential-equation model of the measured variable and the potential disrupting factors (Särkkä

2006, § 3.2). The Bayesian discipline of isolating causal dependencies permits model components and constraints from historical data to be added or modified in a modular fashion. While current computational methods and professional experience with this formalism are limited, we anticipate that using the formalism as a *lingua franca* for model-building will make it significantly easier to extend AI modeling to use more variables, datasets, and interactions in a principled and relatively error-tolerant manner.

References

Agami, Nedaa Mohamed Ezzat, Ahmed Mohamed Ahmed Omran, Mohamed Mostafa Saleh, and Hisham Emad El-Din El-Shishiny. 2008. "An Enhanced Approach for Trend Impact Analysis." *Technological Forecasting and Social Change* 75 (9): 1439–1450. doi:10.1016/j.techfore.2008.03.006.

Anderson, Timothy, Rolf Färe, Shawna Grosskopf, Lane Inman, and Xiaoyu Song. 2002. "Further Examination of Moore's Law with Data Envelopment Analysis." *Technological Forecasting and Social Change* 69 (5): 465–477. doi:10.1016/S0040-1625(01)00190-1.

Asan, Seyda Serdar, and Umut Asan. 2007. "Qualitative Cross-Impact Analysis with Time Consideration." *Technological Forecasting and Social Change* 74 (5): 627–644. doi:10.1016/j.techfore.2006.05.011.

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.

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. 1998. "Long-Term Growth as a Sequence of Exponential Modes." Unpublished manuscript. Last revised December 2000. http://hanson.gmu.edu/longgrow.pdf.

Kurzweil, Ray. 1990. *The Age of Intelligent Machines.* Cambridge, MA: MIT Press. http://www.kurzweilai.net/ebooks/the-age-of-intelligent-machines.

———. 2005. *The Singularity Is Near: When Humans Transcend Biology.* New York: Viking.

Matheny, Jason G. 2007. "Reducing the Risk of Human Extinction." *Risk Analysis* 27 (5): 1335–1344. doi:10.1111/j.1539-6924.2007.00960.x.

Nodelman, Uri D. 2007. "Continuous Time Bayesian Networks." PhD diss., Stanford University. http://ai.stanford.edu/~nodelman/papers/ctbn-thesis.pdf.

Peláez, Antonio López, and Dimitris Kyriakou. 2008. "Robots, Genes and Bytes: Technology Development and Social Changes Towards the Year 2020." *Technological Forecasting and Social Change* 75 (8): 1176–1201. doi:10.1016/j.techfore.2008.01.002.

Sandberg, Anders, and Nick Bostrom. 2008. *Whole Brain Emulation: A Roadmap.* Technical Report, 2008-3. Future of Humanity Institute, University of Oxford. http://www.fhi.ox.ac.uk/Reports/2008-3.pdf.

Särkkä, Simo. 2006. "Recursive Bayesian Inference on Stochastic Differential Equations." PhD diss., Helsinki University of Technology. http://lib.tkk.fi/Diss/2006/isbn9512281279/isbn9512281279.pdf.

Tversky, Amos, and Daniel Kahneman. 1983. "Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment." *Psychological Review* 90 (4): 293–315. doi:10.1037/0033-295X.90.4.293.

The Uncertain Future. 2009. "Visualizing 'The Future According to You.'" The Uncertain Future. Accessed August 3, 2012. http://www.theuncertainfuture.com/.

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`.

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.