Interim Report from the Panel Chairs

AAAI Presidential Panel on Long-Term AI Futures

August 2009

The AAAI 2008-09 *Presidential Panel on Long-Term AI Futures* was organized by the president of the Association for the Advancement of Artificial Intelligence (AAAI) to bring together a group of thoughtful computer scientists to explore and reflect about societal aspects of advances in machine intelligence (computational procedures for automated sensing, learning, reasoning, and decision making). The panelists are leading AI researchers, well known for their significant contributions to AI theory and practice. Although the final report of the panel has not yet been issued, we provide background and high-level summarization of several findings in this interim report.

AI research is at the front edge of a larger computational revolution in our midst—a technical revolution that has been introducing new kinds of tools, automation, services, and new access to information and communication. Efficiencies already achieved via computational innovations are beyond the scope of what people could have imagined just two decades ago. It is clear that AI researchers will spearhead numerous innovations over the next several decades. Panelists overall shared a deep enthusiasm and optimism about the future influence of AI research and development on the world. Panelists expect AI research to have great positive influences in many realms, including healthcare, transportation, education, commerce, information retrieval, and scientific research and discovery.

The panel explored a constellation of topics about societal influences of AI research and development, reviewing potential challenges and associated opportunities for additional focus of attention and research. Several topics were highlighted as important areas for future work; there was a sense that, for these issues, increased sensitivity, attention, and research would help to ensure better outcomes. The panel believed that identifying and highlighting potential "rough edges" that might arise at the intersection of AI science and society would be beneficial for directing ongoing reflection, as well as for guiding new research investments. The study had three focus areas and associated subgroups.

Subgroup on Pace, Concerns, Control

The first focus group explored concerns expressed by lay people—and as popularized in science fiction for decades—about the long-term outcomes of AI research. Panelists reviewed and assessed popular expectations and concerns. The focus group noted a tendency for the general public, science-fiction writers, and futurists to dwell on *radical* long-term outcomes of AI research, while overlooking the broad spectrum of opportunities and challenges with developing and fielding applications that leverage different aspects of machine intelligence.

Popular perspectives on the outcomes of AI research include expectation that there will be one or more disruptive outcomes. These include that notion that the research will somehow lead to the advent of utopia or catastrophe. The utopian perspective is perhaps best captured in the writings of Ray Kurzweil and others, who speak of a forthcoming "technological singularity." At the other end of the spectrum, some people are concerned about the "rise of intelligent machines," fueled by popular novels and movies, that tell stories of the loss of control of robots. Whether forecasting utopian or catastrophic outcomes, the radical perspectives are frightening to people in that they highlight some form of radical change on the horizon—often founded on a notion of the loss of control of the computational intelligences that we create.

The panel of experts was overall skeptical of the radical views expressed by futurists and science-fiction authors. Participants reviewed prior writings and thinking about the possibility of an "intelligence explosion" where computers one day begin designing computers that are more intelligent than themselves. They also reviewed efforts to develop principles for guiding the behavior of autonomous and semi-autonomous systems. Some of the prior and ongoing research on the latter can be viewed by people familiar with Isaac Asimov's Robot Series as formalization and study of behavioral controls akin to Asimov's Laws of Robotics. There was overall skepticism about the prospect of an intelligence explosion as well as of a "coming singularity," and also about the large-scale loss of control of intelligent systems. Nevertheless, there was a shared sense that additional research would be valuable on methods for understanding and verifying the range of behaviors of complex computational systems to minimize unexpected outcomes. Some panelists recommended that more research needs to be done to better define "intelligence explosion," and also to better formulate different classes of such accelerating intelligences. Technical work would likely lead to enhanced understanding of the likelihood of such phenomena, and the nature, risks, and overall outcomes associated with different conceived variants.

The group suggested outreach and communication to people and organizations about the low likelihood of the radical outcomes, sharing the rationale for the overall comfort of scientists in this realm, and for the need to educate people outside the AI research community about the promise of AI for enhancing the quality of human life in numerous ways, coupled with a re-focusing of attention on actionable, shorter-term challenges.

Subgroup on Shorter-Term Challenges

A second subgroup focused on nearer-term challenges, examining potential "rough edges," where AI research touches society, that may be addressed via new vigilance, sensitivity, and, more generally, with investment in additional focused research. Several areas for future research were identified as having valuable payoff in the shorter term. These include the promise of redoubling research on using AI methods to enhance peoples' privacy. There already has been interesting and valuable work in the AI research community on methods for enhancing privacy while enabling people and organizations to personalize services. Other shorter-term opportunities include the value of making deeper investments in methods that enhance interactions and collaborations between people and machine intelligence. The panel's deliberation included discussion of the importance of endowing computing systems with deeper competencies at working in a complementary manner with people on the joint solution of tasks, and in supporting fluid transitions between automated reasoning and human control. The latter includes developing methods that make machine learning and reasoning more transparent to people, including, for example, giving machines abilities to better explain their reasoning, goals, and uncertainties. Another focus of discussion centered on the prospect that people, organizations, and hostile governments might harness a variety of AI advances for malevolent purposes. To our knowledge, such efforts have not yet occurred, yet it is not difficult to imagine how future computer malware, viruses, and worms might leverage richer learning and reasoning, accessing an increasing number of channels of information about people. AI methods might one day be used to perform relatively deep and long-term learning and reasoning about individuals and organizations—and then perform costly actions in a sophisticated and potentially secretive manner. There was a shared sense that it would be wise to be vigilant and to invest in proactive research on these possibilities. Proactive work includes new efforts in security, cryptography, and AI research in such areas as user modeling and intrusion detection directed at this potential threat, in advance of evidence of such criminal efforts.

Subgroup on Ethical and Legal Issues

A third subgroup focused on ethical and legal questions. This subgroup reflected about ethical and legal issues that could become more salient with the increasing commonality of autonomous or semi-autonomous systems that might one-day be charged with making (or advising people on) high-stakes decisions, such as medical therapy or the targeting of weapons. The subgroup's deliberation included reflection about the applicability of current legal frameworks. As an example, the group reviewed potential issues with assignment of liability associated with costly, unforeseen behaviors of autonomous or semi-autonomous decision-making systems. Other reflection and discussion centered on potential ethical and psychological issues with human responses to virtual or robotic systems that have an increasingly human appearance and behavior. For example, the group reflected about potential challenges associated with systems that synthesize believable affect, feelings, and personality. What are the implications of systems that emote, that express mood and emotion (e.g., that appear to care and nurture), when such feelings do not exist in reality? Discussion centered on the value of investing more deeply in research in these areas, and of engaging ethicists, psychologists, and legal scholars.

Meeting at Asilomar

After several months of discussion by email and phone, a face-to-face meeting was held at Asilomar, at the end of Feburary 2009. Asilomar was selected as a site for the meeting primarily because it is simply a fabulous place for a reflective meeting. We also selected the site given the broad symbolism of the location. The AAAI Panel on Long-Term AI Futures resonated broadly with the 1975 Asilomar meeting by molecular biologists on recombinant DNA—in terms of the *high-level goal of social responsibility for scientists*. The AAAI panel co-chairs also alluded to the goal of generating a report on an assessment and recommendations that would be similar to the 1975 recombinant DNA report in terms of the crispness, digestability, and design for consumption by scientists and the public alike. However, the symbolism stops there: The context and need for the AAAI study differs significantly and in multiple ways from the context of the 1975 meeting on recombinant DNA. In 1975, molecular biologists needed urgently to address a fast-paced set of developments that had recently led to the ability to modify genetic material. The 1975 meeting took place amidst a recent moratorium on recombinant DNA research. In stark contrast to that situation, the context for the AAAI panel is a field that has shown relatively graceful, ongoing progress. Indeed, AI scientists openly refer to progress as being somewhat disappointing in its pace, given hopes and expectations over the years. However, we are seeing ongoing advances in the prowess of AI methods and an acceleration in the fielding of real-world applications (some quite large in scale), a natural increase of reliance on automation, the coming availability of sophisticated methods to a wider set of developers, extending well outside the research community (e.g., in the form of a variety of toolkits), and a growing interest and focus among non-experts on radical outcomes of AI research. On the latter, some panelists believe that the AAAI study was held amidst a *perception* of urgency by non-experts (e.g., a book and a forthcoming movie titled "*The Singularity is Near*"), and focus of attention, expectation, and concern growing among the general population.

The panel has identified multiple opportunities for proactive reflection, focused research, and ongoing sensitivity and attention. We believe that focusing effort as a community of AI scientists on potential societal issues and consequences will ensure the best outcomes for AI research, enabling society to reap the maximal benefits of AI advances.

Eric Horvitz and Bart Selman, *Co-chairs*

AAAI Presidential Panel on Long-Term AI Futures