Asilomar AI Principles: A framework for human value alignment

February 6 2017, by Nancy Owano

Credit: Piotr Siedlecki/Public Domain

(Tech Xplore)—This site and other tech-watching portals give you ample news throughout the week on enhancements to technology making use of artificial intelligence—new awakenings, new feats. The time, artificial intelligence drew a crowd of technologists and thought leaders for a discussion about AI principles.

Step back to last month when the Future of Life Institute's second conference on the future of artificial intelligence experts at the Beneficial AI conference took place in Asilomar, California.

Sam Shead in *Business Insider* reported that 23 principles were developed "off the back of the Beneficial AI conference." He said attendance included "some of the most high profile figures in the AI community, including DeepMind CEO Demis Hassabis and Facebook AI guru Yann LeCun."

The Institute describes its work: "We are currently focusing on keeping artificial intelligence beneficial and we are also exploring ways of reducing risks from nuclear weapons and biotechnology."

Rappler reported on Sunday that we now have a newly developed 23 Asilomar AI Principles.

The list has three components, serving as the topic umbrellas for the principles stated: Research Issues, Ethics and Values and Longer-Term Issues.

Regarding the group's thoughts on research, "The goal of AI research should be to create not undirected intelligence, but beneficial intelligence."

Regarding superintelligence, it should only be developed "in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization."

As for recursive self-improvement: "Systems that were designed to recursively self-improve or self-replicate in a manner that could lead to rapidly increasing quality or quantity must be subject to strict safety and control measures."

Why is this important and not just an opportunity for AI experts to get away for the day on a talkfest?

AI is getting smarter and the question becomes who's in charge now and who or what may be in charge tomorrow.

As the Future of Life Institute put it, "a major change is coming, over unknown timescales but across every segment of society."

Sharing thoughts, concerns and suggestions is an important exercise, as experts want to make sure AI remains beneficial and not going rogue.

One discussion point made at the event is that you can have the most efficient AI system but politics and social policy are considerations too. You cannot solve the problems just technologically.

Part of the struggle is in the area of politics and social policy, because we're talking about human AI. So if we are blessed with a technically sound AI system, a question remains if we do have the political will and ethics to use it to people's benefit.

Business Insider pointed out that AI has two sides of the coin; it can cure cancer and slow global warming or it can destroy us.

So will the principles effort carry any weight? George Dvorsky in *Gizmodo* pointed out that at the meeting the FLI organizers only accepted a principle "if at least 90 percent of the attendees agreed with it."

Dvorsky likened the group's output to a "kind of best practices rulebook as it pertains to the development of AI."

He quoted physicist Anthony Aguirre: "Our hope was—and will be

going forward—to involve a very diverse set of stakeholders, but also to ground the process in the expert knowledge of the AI developers who really understand what systems are being developed, what they can do, and where they may be going."

Principles are not edicts. As Dvorsky also pointed out, "There's no provision to have these guidelines enforced or updated, nor is there any call for institutional or governmental oversight," he said. "As it stands, developers can violate these best-of-breed recommendations, and face no consequences."

He further commented that as idealistic as the principles may be, "we're better off having them."

More information: futureoflife.org/ai-principles/