The Asilomar AI Principles

Testimony, as prepared, by Anthony Aguirre, co-founder and Board Member of the Future of Life Institute
Vermont House of Representatives
Committee on Energy and Technology
April 28, 2021

Good morning. Chairman Briglin, Ranking Member Tangerman, and other distinguished Members of the Committee, thank you for inviting me to testify at today's hearing. My name is Anthony Aguirre and I'm a professor at the University of California at Santa Cruz, and a co-founder and co-director of the Future of Life Institute, a nonprofit that brings together luminaries from the academic, corporate, and nonprofit worlds to research, discuss, and share insights regarding the impact of major society-shaping new technologies such as artificial intelligence. In addition to my testimony today, I would like to enter into the record specific comments we have prepared with regard to H.140 and H.263.

Suddenly, starting just a few years ago, machine learning and artificial intelligence systems are everywhere – from driving directions to cars that will someday soon drive themselves, and including systems that recognize faces, translate text, recognized spoken directions, organize newsfeeds, defeat masters in Chess and Go, compose text, and aid in scientific research.

In January of 2017, the Institute called together a meeting in Asilomar, California to bring together prime movers in AI and related fields to discuss an urgent question: given the recent explosion of capability and adoption in machine learning and artificial intelligence, how do we make sure AI is not just powerful, and not just profitable, but beneficial to its consumers, users, and society at large.

We set for this group a specific task: to source, formulate, debate, and hopefully adopt a set of principles that could guide technologists, policymakers, and others toward this goal. This process, started by reading and synthesizing essentially all extant proposals for AI policy — something that could be done then, then composing and debating specific possibilities in the weeks prior to the meeting, then discussing and refining them for several days in-person. This succeeded beyond our expectations. We were able to formulate 23 principles with excellent consensus support across the spectrum.

They have drawn support from the heads of AI research at companies such as Google DeepMind, Facebook, OpenAI, Google Brain, Apple, in all more than 1000 AI researchers, as well as leaders in the tech industry like Elon Musk and Sam Altman, those in academia from the late Stephen Hawking to Stuart Russell and Peter Norvig who literally wrote the textbook on AI, and many in the nonprofit and policy worlds.

We were able to get this level of support and consensus both because the issue is of great importance and, I think, because many of these principles are almost self-evidently desirable. We all want AI systems that are safe (principle 6), that are understandable (principles 7 and 8), that respect our liberty and privacy and rights, that support our shared prosperity, mitigate risks, and keep us humans in the drivers' seat.

But we all know these desires are not self-fulfilling. Safe technologies don't happen by themselves, privacy is not automatically respected, and civic processes are not inherently immune from subversion. Potentially dangerous arms races, literal and figurative, can start even if nobody wants them. Ensuring the benefit of technology takes active effort and participation — from researchers, developers, policymakers such as yourselves, and the public.

Since 2017, such activity has rapidly expanded. The Asilomar principles were officially endorsed by the state of California; and other sets of principles have been devised. Of note are the OECD AI principles, which have been officially adopted by the US. These are strong, overlap heavily with the Asilomar Principles, and FLI endorses them. (Though we do have a reservation, which I'll return to momentarily.) Legislation and other policy addressing AI has been crafted in a number of states, the US recently passed the very significant National AI Initiative Act, and the European Union is contemplating a major piece of legislation.

There are lots of concerns regarding bias, privacy, judicial transparency, and the like, which must be addressed for AI systems to comport with our laws and ideals, and the Asilomar Principles among others clearly point to these. But in my own view there are two particular classes of danger indicated by the Asilomar principles that often get shunted aside because it is not in the interest of AI developers to talk much about them.

The first concerns an Asilomar principle using a word that is surprisingly uncommon nowadays.

> ***Liberty and Privacy:*** *The application of AI to personal data must not unreasonably curtail people's real or perceived liberty.*

Many can clearly see the trajectory many see toward a collection giant corporations awash in personal data with high-powered AI systems designed to gather data, sell products, and potentially manipulate individuals. If we end up in a system where each individual is responsible for protecting themself against the power of these systems working to exploit them, it's going to be an ugly picture. This power disparity will not solve itself, and in my view can only truly be addressed by the power that the government wields in service of the population. Closely related, and to watch out for, is

> ***Non-subversion:*** *The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, the social and civic processes on which the health of society depends.*

Second, very highly capable AI systems, including but not limited to so called "artificial general intelligence" are coming — and we really don't know when — and provide great opportunity but also immense risks. As the Asilomar Principles state, **"Advanced AI could represent a**

profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources." (principle 20). And as they call for, **high-powered AI systems must be subject to planning and mitigation efforts commensurate with their expected impact (principle 21).** We encourage the prospective AI Commission to be established by H.410 to anticipate that AI systems will become increasingly powerful and increasingly generalizable, and as such, present unique risks to society.

We will happily support Vermont in this endeavor, and I look forward to any questions you may have for me today.