JDSR 6(4) (2024) 107-117

JOURNAL OF DIGITAL SOCIAL RESEARCH

10.33621/jdsr.v6i440477

Getting democracy wrong

How lessons from biotechnology can illuminate limits of the Asilomar AI principles

Gwendolyn Blue[1], Mél Hogan[2]

[1] University of Calgary, Calgary, Alberta, Canada
[2] Queen's University, Kingston, Ontario, Canada

✉ ggblue@ucalgary.ca

Abstract

Recent developments in large language models and computer automated systems more generally (colloquially called 'artificial intelligence') have given rise to concerns about potential social risks of AI. Of the numerous industry-driven principles put forth over the past decade to address these concerns, the Future of Life Institute's Asilomar AI principles are particularly noteworthy given the large number of wealthy and powerful signatories. This paper highlights the need for critical examination of the Asilomar AI Principles. The Asilomar model, first developed for biotechnology, is frequently cited as a successful policy approach for promoting expert consensus and containing public controversy. Situating Asilomar AI principles in the context of a broader history of Asilomar approaches illuminates the limitations of scientific and industry self-regulation. The Asilomar AI process shapes AI's publicity in three interconnected ways: as an agenda-setting manoeuvre to promote longtermist beliefs; as an approach to policy making that restricts public engagement; and as a mechanism to enhance industry control of AI governance.

Keywords: Principles, Asilomar, Governance, Artificial Intelligence, Biotechnology, Longtermism

1. Introduction

In 2017, the Future of Life Institute (FLI) unveiled the Asilomar AI principles aimed at guiding AI research, development, and commercialization toward the benefit of all (Stirling, 2018). These principles were crafted at an invitation-only workshop at the Asilomar Conference Grounds in Pacific Grove, California which brought together prominent figures in industry and academia with a shared objective to create a unified vision for AI governance.

Founded in 2014 by tech entrepreneurs and academic researchers,[1] the FLI supports and promotes longtermism, an influential belief system, social movement, and policy narrative that focuses on ensuring

[1] FLI founders include Skype co-founder Jaan Tallin, DeepMind's Viktoriy Krakovna, MIT cosmologist Max Tegmark, and physicist Anthony Aguirre at University of California Santa Cruz. Advisors include entrepreneurs such as Elon Musk, AI researchers such as Stuart Russell, physicists such as Steven Hawking (now deceased), geneticists such as George Church, and science communicators such as Alan Alda, among others (Labaree, 2014).

the well-being of inhabitants of the distant future by preventing catastrophic risks of technological innovation in the present (Torres 2021, Crary 2023). Drawing inspiration from philosopher Nick Bostrom's concept of effective altruism, a moral belief that advocates for wealth creation as the best way to promote the public good (Bostrom 2014), William MacAskill coined the term longtermism to describe approaches that seek to protect future humans from existential risks that threaten humanity's long-term potential, including risks from AI (MacAskill, 2022). The FLI fulfils its longtermism mission, in part, by organizing invite-only conferences with AI experts to promote principles to guide the future of AI. As stated on the FLI website, "we believe that meetings of experts from industry and academia, and from across the globe, can significantly improve common understanding and improve our collective future" (FLI, n.d.).

The FLI's Asilomar AI principles are one illustration of broader AI ethics initiatives, which are principle-based approaches to governance led by industry, government, academia, and not-for-profit groups that have gained momentum over the past decade (Jobin et al., 2019; Marchant et al., 2020; Milmo, 2023; Munn, 2023). Examples of AI ethics statements include civil society-driven initiatives such as Amnesty International's Toronto Declaration, state-driven initiatives such as the European High Level Expert Group on AI's Ethics Guidelines for Trustworthy AI, and stakeholder initiatives such as the Montréal Declaration on Responsible AI, among many others (for comprehensive overview, see Jobin et al., 2019). Ethical guidelines are a response to concerns about the social implications of emerging technologies, and are often presented as a flexible alternative to, or prototype for, government regulation to guide the trajectory of technological research, development, and implementation in ways that do not stymie innovation. Critics argue that principle-based ethical initiatives are a strategic manoeuvre on the part of technology developers and private industry to sidestep government regulation and avoid public scrutiny (Wagner, 2018; Bender et al., 2021; McKelvey and Roberge, 2023; Munn, 2023). The inherent ambiguity of principles further compounds the challenge of their implementation.

The Asilomar AI principles were among the first public-facing principles to guide the development of AI and played an agenda-setting role for subsequent principle-based frameworks (Fjeld et al., 2020). To date, the principles have been adopted and signed by thousands of individuals and organizations. This paper positions the Asilomar AI principles as a noteworthy dimension of AI's contemporary publicity. Situating the Asilomar AI principles within a longer history of development of the Asilomar model, a well-established principle-based approach to science policy, highlights the FLI's emphasis on self-regulation and particularly the assumption that the people who develop technology and put it to use are best positioned to define the terms of public debate and government regulation. Our concern is not with principle-based governance of emerging technologies per se, but rather with use of the Asilomar model by the FLI to promote its guiding principles. First developed in the early 1970s in response to innovations in genetic engineering, the Asilomar model – as will be discussed further - has since been applied to other publicly contentious technologies such as geoengineering, synthetic biology, human gene editing, and most recently, AI. While proponents laud the Asilomar model as a successful example of expert solidarity in anticipating social consequences of technology, critics highlight problematic implications including positioning publics and government regulators in a reactive and reactionary position, forever playing catch-up with technological advances (Hurlbut, 2015a; Hurlbut, 2015b; Jasanoff and Hurlbut, 2018; Taylor and Dewsbury, 2019).

What lessons can be learned about AI's current publicity by situating the Asilomar AI principles in a longer history of Asilomar-style approaches to scientific and technological governance? The FLI's initiatives share many similarities with Asilomar approach to technology governance, including an invitation-only conference which resulted in the release of guiding principles, followed by calls for a temporary pause on AI research to provide time for experts to develop guidelines and regulation. A closer look reveals that, in focus and intent, the Asilomar AI principles, and the process to develop them, co-opted problematic aspects of the Asilomar model. In what follows, we first trace a brief history of debates about experiments in expert-driven technical regulation dating back to the initial Asilomar conferences

https://doi.org/10.33621/jdsr.v6i440477

on recombinant DNA which set a precedent for technological elites to anticipate and address potential social consequences of research, and to forward their own research and commercial interests. Next, we examine how the FLI's Asilomar AI process, which included the Beneficial AI conference, the Asilomar AI principles, and a follow-up letter calling for a pause on AI research, drew on the original Asilomar model and, in so doing, shaped AI's publicity in three interconnected ways: as an agenda-setting manoeuvre to promote longtermist beliefs; as an approach to policy making that restricts public engagement; and as a mechanism to enhance industry control of AI governance.

2. Setting the stage: The Asilomar conferences on recombinant DNA

In the early 1970s, scientists were among the first to sound the alarm about the potential social consequences of emerging genetic engineering technologies, concerns which led to Asilomar-style approaches to science policy. Since details about the Asilomar conferences are well-documented in the academic literature, we present here a brief overview of key events.

In 1972, biochemist Paul Berg, along with other colleagues at Stanford University, proposed the creation of a new DNA molecule by inserting genetic material from a monkey virus (SV40) into a virus that infects *Escherichia coli*, a bacteria found in the human gut (Jackson et al., 1972). The new hybrid molecule, and the technique used to develop it, are commonly referred to as recombinant DNA (rDNA).[2] In January 1973, in response to scientific and public concerns about the potential consequences of rDNA research, Berg organized a scientific conference at the Asilomar Hotel and Conference Grounds in Pacific Grove, California to discuss potential biohazards with a focus on laboratory safety protocols. In the same year, Stanley Cohen (Stanford) and Herbert Boyer (UCSF) published results of the first successful cloning experiment which introduced rDNA into bacteria, for which they would later receive the first biotechnology patent which inaugurated the biotechnology industry. In June 1973, further concerns about rDNA research were raised at the Gordon Conference on Nucleic Acids, resulting in a letter by scientists at the conference to the National Academies of Science (NAS) and the National Academy of Medicine to implement guidelines for biotechnology research (Singer and Soll, 1973). In response, NAS created an ad hoc study group chaired by Berg called the Committee on Recombinant DNA Molecules, which penned an open letter calling for a moratorium on rDNA research until the hazards were better understood (Berg et al., 1974). This letter had two important outcomes: it resulted in the first voluntary moratorium in the history of molecular biological research; and, it raised awareness among scientists and broader publics of potential harmful implications of biotechnology.

The second, and more well-known, Asilomar conference took place at the same location in February 1975. Lasting over four days, this meeting - officially called the International Congress on Recombinant DNA Molecules - convened an international group of 140 scientists primarily in molecular biology, along with a handful of lawyers and journalists, to discuss potential consequences of rDNA, and to develop guidelines to lift the voluntary moratorium. Following the conference, the organizers filed a report to NAS outlining recommendations and conditions under which rDNA research could proceed (Berg et al., 1975). These conditions included the establishment of appropriate safeguards, namely efforts to ensure the containment of newly created organisms through the implementation of biological and physical barriers. In 1976, these containment principles were formulated into National Institute of Health (NIH) guidelines for the funding of rDNA research.

Twenty-five years after the initial Asilomar conferences on rDNA, lawyer and medical ethicist Alexander Capron organized a follow up workshop - Asilomar 3 - to discuss whether Asilomar's legacy had contemporary relevance for emerging technology innovations (Barinaga, 2000). Attended by many

[2] Recombinant DNA is now a routine and commonly used laboratory technique.

https://doi.org/10.33621/jdsr.v6i440477

organizers of the original Asilomar conferences, Asilomar 3 convened a more diverse gathering including scientists but also philosophers, historians, ethicists, legal scholars, and government officials (Capron and Shapiro, 2001). The attendees agreed that the original Asilomar model was no longer appropriate for the following reasons. Views about biotechnology had become polarized in ways that made consensus difficult, if not impossible, to achieve. Controversies over genetically modified foods in the 1990s revealed that scientific proclamations of safety-through-containment were insufficient to quell public concerns about commercial, ethical, and social consequences of genetic engineering. In turn, powerful commercial forces in genomic science created conflicts of interest that undermined the legitimacy of scientific self-governance (Barinaga, 2000). Conference attendees concluded that these limitations signalled the need to extend the Asilomar model to include a broader range of perspectives, values, and expertise beyond scientists and technology developers to ensure that social, ethical, and legal aspects of emerging technologies were addressed in advance of implementation (Rufo and Ficorilli, 2019).

Following the Asilomar 3 meeting, several Asilomar + ("self-regulation plus") models were applied to contentious technologies such as synthetic biology (Ferber, 2004), geo-engineering (ASOC, 2010), environmental algorithms (Galaz, 2015), and human germline editing (Baltimore et al., 2015). At one high-profile Asilomar+ meeting in 2015, three years before He Jiankui used genome editing on human embryos, scientists, bioethicists, and legal experts convened to discuss the technical and social implications of human germline editing made possible by advances in gene editing technologies. Berg was instrumental in organizing this workshop which was modelled in the style of the original Asilomar conference but with attention to the need for a more open dialogue with ethicists about the steps to guarantee the safe development of human genome modification (Baltimore et al., 2015). Concurrently, editorials in academic venues such as *Nature* (Editorial, 2015) and public-facing venues such as the *New York Times* (Capron, 2015) highlighted the continued relevance of the Asilomar model for science policy and underscored the necessity of engaging with broader stakeholders beyond the scientific community.

The Asilomar model continues to play a symbolic role in science and technology policy as shorthand for responsible self-regulation, where scientists and technology developers take the lead in publicly acknowledging potential harmful consequences, and in developing principles to guide technology development and implementation. Asilomar's continued relevance stems in part from the successful outcomes of the original Asilomar conferences, outcomes which include improved attention to laboratory safety, and, from the perspective of scientists, an avoidance of stringent government regulation (Berg, 2004, 2006, 2008). In Berg's estimation, the success of the Asilomar model resulted specifically from focused discussion of plausible risks that could be feasibly addressed by technical experts (Berg, 2004).

The Asilomar model has also generated significant critique, particularly about the democratic implications of limiting input to technical experts and restricting the scope of issues to technical matters (Rogers, 1975; Krimsky, 1982; Gottweis, 2005; Gregorowius et al., 2017; Hurlbut, 2015a; Hurlbut, 2015b; Jasanoff et al., 2019; Parthasarathy, 2015; Rufo and Ficorilli, 2019; Taylor and Dewsbury, 2019). In essence, the Asilomar model enables technology developers to powerfully shape the trajectory of regulation and public opinion. The Asilomar model – even the improved Asilomar + approach – does not empower diverse constituencies to shape technological research and innovation agendas, but rather secures support for expert-driven governance and technology commercialization (Rufo and Ficorilli, 2019). According to Ben Hurlbut, the Asilomar model gets democracy wrong: arguably, from a democratic perspective, technologies should be shaped by collectively articulated visions of the public good and the public interest, and not solely by the people who stand to profit from technology development and implementation (Hurlbut, 2015a). Asilomar-style models, including Asilomar+, position technical experts as the most qualified to regulate the social aspects of technology, and in turn position publics and regulatory agencies as passive recipients of principles generated by scientific and technical communities. "The public role that the Asilomar story celebrates is one of dependence," as Taylor and Dewsbury note, "with the public passively learning — and deferring to — science's authoritative judgment about what is at stake and when a democratic reaction is warranted" (2019: 12).

https://doi.org/10.33621/jdsr.v6i440477

Many critical commentators have subsequently concluded that Asilomar and Asilomar+ models are a missed opportunity to anticipate and address social concerns in a democratic fashion (Parthasarathy, 2015). Proposals for improvement include the development of procedures to enhance public engagement in developing ethical principles (Rufo and Ficorilli, 2019), and, more substantively, the creation of new global institutions to bring diverse views to inform the governance of emerging technologies (Jasanoff and Hurlbut, 2018).

3. Asilomar AI: Process and principles

The Asilomar AI Principles were developed during an FLI conference on Beneficial AI that took place in January 2017 over two and a half days at the Asilomar Conference Center, as a sequel to a 2015 conference held in Puerto Rico. The California-based Beneficial AI conference was attended by over 1000 participants, including several leading AI researchers and entrepreneurs, in addition to experts from economics, law, ethics, and philosophy. Many of the speakers at the conference are well-known promoters of longtermism such as Nick Bostrom, Ray Kurzweil, and Elon Musk. The conference schedule, and all talks, are publicly available on the FLI website (FLI, n.d.).

Session descriptions provide an indication of the focus of the Beneficial AI conference, much of which centred on the permissibility of technological innovation, ways to circumvent regulation, and the inevitability of super-intelligent AI. Sessions ranged from economics ("How can we grow our prosperity through automation without leaving people lacking income or purpose"), creating human-level AI ("What remaining obstacles can be identified?"), superintelligence ("What can we do now to maximize the probability of a positive outcome"), law, policy, and ethics ("How can we update legal systems, international treaties and algorithms to be more fair, ethical, and efficient and to keep pace with AI?").

A key outcome of the Beneficial AI workshop was the development of a consensus document consisting of 23 voluntary principles organized into three categories (Research Issues, Ethics and Values, Long-Term Issues) (FLI, 2017). These principles represent a voluntary commitment on the part of its signatories for the research, development, and commercialization of AI. Overall, the Asilomar AI principles reflect the techno-optimist, longtermist beliefs of the FLI. Consider, for instance, the opening statement of the Asilomar AI principles: "Artificial intelligence has already provided beneficial tools that are used every day by people around the world. Its continued development, guided by the following principles will offer amazing opportunities to help and empower people in decades and *centuries* ahead" (FLI, 2017: italics added). Underpinning these principles is the assumption that AI will develop superintelligence, and that AI is inherently beneficial to all humans (if governed correctly).

As others have observed, many of principles are vague and inconsistent, and we present here select examples for illustrative purposes (see Stapf-Fine et al., 2018 for more extensive overview). For instance, the first principle states that the goal of AI research should be to create beneficial but not undirected intelligence. This connection is weak: something can be undirected (i.e. uncontrolled) but still be beneficial, and something can be controlled and still be harmful. Principles three and four call for a culture of cooperation, marked by "constructive and healthy exchanges" among AI researchers and policy-makers. What makes for a "healthy exchange" of views between researcher and policy-makers? Cooperation is not inherently healthy, as conflict and conflicting opinions also play a significant role in policy development. It is questionable whether policy makers should be expected to cooperate and align harmoniously with the interests of industry. Principle 12 requires the right to personal data protection but is restricted to data generated by users and does not address data collected about users. Principle 14 aims to empower and benefit as many people as possible, suggestive of an instrumental goal of maximized self-interest consistent with longtermism, but with no indication of how such benefits might be realized in an inequitable world. Overall, the Asilomar AI principles paint a utopian picture of the inevitability and desirability of AI, suggestive of the possibility of a dystopian future if AI is not developed in alignment with longtermist values. Beyond questions of logical coherence, the principles sidestep

https://doi.org/10.33621/jdsr.v6i440477

important questions such as what AI is, what constitutes a risk or benefit (and for whom), and what kind of futures various and diverse people might find desirable.

The Asilomar AI process, similar to the original Asilomar model, included a call for a moratorium on research, although the timing was significantly different in the Asilomar AI case. In the original Asilomar model, the goal of the moratorium was to provide time for experts to convene to discuss potential responses to possible biohazards associated with rDNA (Berg and Singer, 1995). In the Asilomar AI process, calls for a moratorium emerged five years following the initial release of the Asilomar AI principles. In March 2023, one week after the public release by Open AI of the large language model ChatGPT-4, the FLI issued an open letter calling for a six-month pause on the training of AI systems more powerful than GPT-4 (FLI, 2023). The rationale for the pause, according to the FLI letter, was to provide time for AI labs and experts to develop and implement safety protocols. The letter began with the following statement "AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research and acknowledged by top AI labs," followed by a reference to the "widely-endorsed" Asilomar AI principles. and particularly the principle that "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" (FLI, 2023). In the letter, the FLI lamented that current levels of planning and management had yet to adequately address the risks of AI. As such, a temporary pause was necessary, not for the development of AI in general, but to prevent a "dangerous race to ever-larger unpredictable black-box models with emergent capabilities" (FLI, 2023). The FLI claimed that AI development should be focused on ensuring that systems are developed in accordance with principles such as accuracy, safety, interpretability, transparency, robustness, alignment, trustworthiness, and loyalty. In the letter's conclusion, the FLI stated that adherence to these principles can ensure that "humanity can enjoy a flourishing future with AI…(h)aving succeeded in creating powerful AI systems, we can now enjoy an AI summer: which we reap the rewards, engineer these systems for the clear benefit of all, and give society a chance to adapt" (FLI, 2023).

The FLI pause letter was well-received and to date has amassed over 30 000 signatures. However, the letter's core assumption, that AI technologies will be developed and implemented with industry labs and developers at the helm in shaping ethical and regulatory pathways, was met with harsh criticism by AI ethicists and scientists. Immediately following the letter's public release, Timnit Gebru, Emily Bender, Angelina McMillan-Major, and Margaret Mitchell issued a statement denouncing the letter's misleading and fear-mongering rhetoric (Gebru et al., 2023). Terms used in the FLI pause letter, such as "powerful digital minds," "human-competitive intelligence" and "artificial general intelligence," alongside leading questions, such as "Should we develop nonhuman minds that might eventually outnumber, outsmart, obsolete and replace us," oversell the capacities of automated computer systems, and, in so doing, deceive people by misattributing agency to synthetic media. This has the effect of shifting conversations about accountability and responsibility away from the companies that develop and profit from the commercialization of automated computer systems. Additionally, the pause letter positioned social actors as passive, where society is given a "chance to adapt" to AI rather than having opportunities to actively shape AI's trajectory. In turn, calls for a pause are significantly at odds with the profit-seeking motives of the tech industry, making the prospects of implementation of an actual moratorium slim to none (see Richards et al., 2023).

4. Asilomar AI's publicity

With the historical development of the Asilomar model as a backdrop, what are the implications for AI's publicity of the FLI's use of the Asilomar model? The Asilomar AI principles and process shape AI's publicity in three interconnected ways: as an agenda-setting manoeuvre to promote longtermist beliefs; as an approach to policy debates that restricts public engagement; and as a mechanism to enhance industry interests and control of AI governance. We address each of these points in turn.

https://doi.org/10.33621/jdsr.v6i440477

Promotion of longtermist beliefs: The FLI's longtermist perspective, which prioritizes the enhancement and maximization of humanity's potential in a distant future, is grounded in a concern with existential risk. No longer a fringe philosophy, longtermism is an influential policy agenda for AI governance that informs lawmakers and public concern about AI's future apocalyptic potential (McKelvey and Roberge, 2023). Longtermism provides a moral compass for well-endowed institutions such as the FLI, as well as Open Philanthropy, Open AI, and the now defunct Future of Humanities Institute (a sister institute to the FLI). As the coffers of longtermist organizations grow, so do their influence on policy and governance. The FLI's ability to secure millions of dollars in funding is due, in part, to longtermism adherents' uncritical belief in the transformative power of technology, and unwavering support of the political economic systems that generate the very problems its followers claim to address (Crary 2023, Naughton 2022). A key aspect of longtermist thinking is support for status quo institutions and practices: "longtermism calls on us to safeguard humanity's future in a manner that both diverts attention from current misery and leaves harmful socioeconomic structures critically unexamined" (Crary, 2023: 50).

Consistent with longtermist beliefs, the Asilomar AI principles promote both utopian and dystopian narratives about AI in ways that shift attention away from existing harms caused by automated systems and from the companies that create, endorse, and profit from these systems. Consider, for instance, differences between the FLI framing of its Asilomar AI principles and the original Asilomar framings of rDNA. Whereas the original Asilomar conferences focused on issues such as laboratory safety that were understood by scientific experts, the Asilomar AI conference focused on existential risks and doomsday speculations about abstract and largely unknown future dangers of artificial intelligence. Indeed, throughout the FLI's Asilomar AI principles, the phrase 'artificial intelligence' is invoked as if its meaning were self-evident yet its risks unknown. Whereas rDNA was a novel technology in the 1970s with yet-unrealized benefits and risks, autonomous computerized systems and algorithms are already deployed and existing harms have been well-documented by academic researchers, ethicists, and journalists, including labor issues (Williams et al., 2022), deepening inequality (Eubanks 2017), surveillance and privacy (Zuboff 2019), entrenched racism and sexism (Noble 2018, Buolamwini 2023), environmental degradation (Crawford 2021, Hogan 2015), and the consolidation of corporate power (Lewis-Kraus 2016), among numerous other issues, with solutions, policies, and frameworks available to address these problems.

Many high-profile AI researchers and tech developers follow the ideas set in place by Asilomar AI principles, and use public platforms to generate both fear and hype that AI will soon develop superhuman capabilities and pose existential risks to humans in the distant future. In an interview in *MIT Technology Review,* Max Tegmark, MIT professor and founder and president of FLI, stated that the goal of the FLI pause letter was, in part, to enable public conversations about the existential risks that AI poses to future generations (Heikkilla 2023). Two months after the FLI released its pause letter, the Center for AI Safety – along with Sam Altman of Open AI, Yoshua Bengio, and Geoffrey Hinton - released a short statement about the risks of rogue AI: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war" (Center for AI Safety, 2023). In the same month, Altman and colleagues wrote a blog calling for the "governance of superintelligence" (Altman et al., 2023). In line with longtermist rhetoric informing the Asilomar AI principles, debate focused on how to govern the eventual emergence of harmful AI to ensure its safety in the future, rather than on existing and well-known harms.

In addition to longtermist concerns about a distant future serving as a distraction from existing threats and very real risks, proponents of longtermist beliefs draw on and perpetuate discriminatory beliefs and attitudes, including racism, sexism, classism, and xenophobia, with roots in long-discredited eugenics ideologies (Gebru and Torres, 2024). Longtermism and its associated beliefs contribute the perpetuation

https://doi.org/10.33621/jdsr.v6i440477

and exacerbation of social inequalities alongside the undermining of existing efforts to address ongoing injustices against marginalized communities.[3]

Limited opportunities for public engagement: Similar to the Asilomar model, Asilomar AI provided limited opportunities for participation beyond the circle of invited attendees at the Beneficial AI conference. Conference proceedings and principles were made publicly available, but diverse publics were not given the opportunity to shape the agenda or the content of discussion. Although it is well-known that current applications of AI disproportionately harm groups already disadvantaged by race, gender and socio-economic background, perspectives that account for these disadvantages were not represented in the Beneficial AI conference nor reflected in the Asilomar AI principles. Conspicuously absent from Beneficial AI conference, and from signatories on the Asilomar AI principles and FLI pause letter, were leading critical scholars and thought leaders on AI ethics such as Emily Bender, Meredith Broussard, Joy Buolamwini, Timnit Gebru, Meredith Whittaker, and Safiya Noble (to list only a few) and human-rights focused organizations such as the Distributed AI Research Institute (DAIR) (Prabhakaran et al, 2022). The insularity and lack of diversity in Asilomar AI is particularly problematic given that the FLI claims to speak about benefits of AI for all humanity yet does so from the limited perspective of an economically privileged, predominantly white constituency.

Entrenchment of industry interests and control: Whereas the original Asilomar was led by publicly-funded research scientists concerned about potential social implications of laboratory research, the Asilomar AI was coordinated by an interconnected group of wealthy people and well-funded technology institutions, all with vested financial interests in AI development. As Berg cautioned, the Asilomar model of scientific self-governance does not work if industry interests are at play:

> …the best way to respond to concerns created by emerging knowledge or early-stage technologies is for scientists from publicly-funded institutions to find common cause with the wider public about the best way to regulate — as early as possible. Once scientists from corporations begin to dominate the research enterprise, it will simply be too late (Berg, 2008: 291).

Concern about the role of tech corporations in reshaping social and political orders at national and global scales has been a recurring theme of critical AI scholarship for some time (Zuboff, 2019; Dyer-Witheford et al., 2019). In a review of AI ethics principles, Attard-Frost and colleagues concluded that important principles – such as fairness, accountability, and transparency – are undermined in practice by existing business norms centered on competitiveness and secrecy (Attard-Frost et al., 2023). The firing by Google of AI ethics researchers Timnit Gebru and Margaret Mitchell serves as a case in point of the inadequacy of well-meaning principles in transforming actual industry practices (Schiffer, 2021).

5. Beyond Asilomar

Since its inception five decades ago, the Asilomar model has taken on a symbolic role in science and technology policy. Often hailed as a success story of scientific solidarity in addressing social responsibility, Asilomar's legacy is seen in the privileged positioning of technical experts, particularly those with vested interests in the deployment of the technologies they develop, to render political judgements about how technology should be regulated and governed. The Asilomar model shapes technology regulation by narrowing the focus of public discussion to an examination of safety, impacts, and consequences, while sidelining broader questions about the purposes and imaginaries that animate technological research and innovation. Asilomar-style approaches, even revised ones that acknowledge the importance of diverse expertise, position publics and regulatory agencies as passive recipients of, and by-standers to, expert-driven self-governance.

[3] Although beyond the scope of this paper to discuss in detail, the links between longtermism and eugenics ideologies warrant critical scrutiny, as do intersections between the ideologies driving discourses of safe A.I. and safe biotechnologies.

https://doi.org/10.33621/jdsr.v6i440477

The FLI's Asilomar AI approach echoes and repeats many aspects of the original Asilomar model such as expert-driven principles, calls for a moratorium on research, and limited opportunities for public engagement. Unlike the original Asilomar's narrow focus on risks scientists knew best, the Asilomar AI model focused on existential risks for yet-to-be born humans, while downplaying known harms of current AI deployment. The FLI's Asilomar AI principles served to garner legitimacy for the longtermist agenda by linking AI products with a higher moral purpose, and by evoking Asilomar-style approaches to technology governance.

We see this paper as a contribution to growing literature on principle-based governance of AI. In contextualizing the Asilomar AI principles in a longer history, we draw connections with earlier efforts to govern biotechnology in the 1970s. Our concern rests primarily with the FLI's use of the Asilomar model to advance longtermist principles. We agree with critics who claim that the Asilomar model itself is a "poor and political dangerous model for governing emerging technologies" because of its limited scope and circumscribed opportunities for diverse participation (Parthasarathy, 2015: 305).

Here, we emphasize that it matters for AI's publicity how principles to govern AI are developed, by whom, and in line with which guiding ideas (Ferri and Gloerich 2023). Moving beyond the Asilomar model could entail proving opportunities to diversify engagement in the shaping of AI governance, yet engagement risks becoming a procedural end-in-itself if it serves as a surrogate and substitute for more radical visions and enactments of political judgement and politicized struggle over the future of AI (Barney, 2008). Enhancing democratic politics in AI's publicity should also entail directing attention to the purposes of AI research, development, and implementation; redistributing the profits of AI development from private interests to public commons; strengthening regulatory mechanisms to halt harmful applications of AI; and enacting broader institutional and structural changes to challenge existing power hierarchies and systems of oppression. We view the Asilomar AI principles as a harmful diversion that advances the misguided longtermist ideologies of the FLI and enhances the interests of those who stand to profit from the continued development of AI in the short term.

References

ASOC (2010) The Asilomar Conference Recommendations on Principles for Research into Climate Engineering Techniques. Climate Institute. Washington, D.C. Available at: https://climateviewer.com/downloads/Asilomar-Conference-Report-2010.pdf (accessed December, 2023)

Altman S, G B and Sutskever I (2023) Governance of Superintelligence. In: Open AI. Available at: https://openai.com/blog/governance-of-superintelligence (accessed December 1, 2023).

Attard-Frost B, De los Ríos A and Walters DR (2023) The ethics of AI business practices: a review of 47 AI ethics guidelines. *AI and Ethics* 3(2): 389-406. https://doi.org/10.1007/s43681-022-00156-6

Baltimore D, Berg P, Botchan M, et al. (2015) A prudent path forward for genomic engineering and germline gene modification. *Science* 348(6230): 36-38. https://doi.org/10.12998/wjcc.v11.i3.528

Barinaga M (2000) Asilomar Revisited: Lessons for Today? *Science (American Association for the Advancement of Science)* 287(5458): 1584-1585. https://doi.org/10.1126/science.287.5458.1584

Barney D (2008) Politics and emerging media: The revenge of publicity. *Global Media Journal -- Canadian Edition* 1(1): 89-106.

Bender EM, Gebru T, McMillan-Major A, et al. (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event, Canada, pp.610 - 623. Association for Computing Machinery.

Berg P (2004) *Asilomar and recombinant DNA*. Available at: https://www.nobelprize.org/prizes/chemistry/1980/berg/article/. (accessed December 1, 2023)

Berg P (2006) Brilliant science, dark politics, uncertain law. *Jurimetrics* 46(4): 379-389. http://www.jstor.org/stable/29762947

Berg P (2008) Asilomar 1975: DNA modification secured. *Nature* 455(7211): 290-291. https://doi.org/10.1038/455290a

Berg P, Baltimore D, Boyer HW, et al. (1974) Biohazards of Recombinant DNA. *Science* 185: 3034.

Berg P, Baltimore D, Brenner S, et al. (1975) Summary statement of the Asilomar Conference on recombinant DNA molecules. *Proc. Nat. Acad. Sci* 72: 1981-1984.

Buolamwini J. (2023) *Unmasking AI: My mission to protect what is human in a world of machines*. New York: Random House.

Bostrom N. (2014) *Superintelligence: Paths, dangers, strategies*. Oxford UK: Oxford University Press.
Capron A (2015) The lessons of Asilomar for today's science. *New York Times*, May 28.
Capron A and Shapiro R (2001) Remember Asilomar? Re-examining science's ethical and social responsibilities. *Perspectives in Biology and Medicine* 44(2): 162 - 169. https://doi.org/10.1353/pbm.2001.0022
Center for AI Safety (2023) AI Extinction Statement Press Release. Available at: www.safe.ai/press-release (accessed December 1, 2023).
Crary A (2023) The toxic ideology of longtermism. *Radical Philosophy* 214: 49 - 57.
Crawford K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.
Dyer-Witheford N, Mikkola Kjosen A and Steinhoff J (2019) *Inhuman Power: Artificial Intelligence and the Future of Capitalism.* New York: Pluto Press.
Editorial (2015) After Asilomar. *Nature* 526(7573): 293-294. https://doi.org/10.1038/526293b
Eubanks V. (2017) *Automating inequality: how high-tech tools profile, police, and punish the poor*. New York: St. Martin's press.
Ferber D (2004) Time for a Synthetic Biology Asilomar? *Science* 303(5655): 159-159. https://doi.org/10.1126/science.303.5655.159
Ferri G, Gloerich I. (2023) Risk and harm: Unpacking ideologies in the AI discourse. *Proceedings of the 5th International conference on Conversational User Interfaces* 23: 1 - 6
Fjeld J, Achten N, Hilligoss H, et al. (2020) Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. Berkman Klein Center. January 15.
Future of Life Institute (FLI) *Our mission*. Available at: https://futureoflife.org/our-mission/ (accessed December 1, 2023).
Future of Life Institute (FLI) (2017) *AI Principles*. Available at: https://futureoflife.org/open-letter/ai-principles/ (accessed December 1, 2023).
Future of Life Institute (FLI) (2023) *Pause giant AI experiments: An open letter*. Available at: https://futureoflife.org/open-letter/pause-giant-ai-experiments/(accessed December 1, 2023).
Future of Life Institute (FLI) (n.d.) *Beneficial AI 2017*. Available at: https://futureoflife.org/event/bai-2017/ (accessed December 1, 2023).
Galaz V (2015) A manifesto for algorithms in the environment. *Guardian*, October 5. Available at: https://www.theguardian.com/science/political-science/2015/oct/05/a-manifesto-for-algorithms-in-the-environment. (accessed December 1, 2023)
Gebru T, Bender E, McMillan-Major A, et al. (2023) Statement from the listed authors of Stochastic Parrots on the "AI Pause" letter. DAIR. Available at: https://www.dair-institute.org/blog/letter-statement-March2023/ (accessed December 1, 2023).
Gebru T, Torres E. (2024). The TESCREL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday* 29 (4).
Gottweis H (2005) Transnationalizing recombinant-DNA regulation: Between Asilomar, EMBO, the OECD, and the European Community. *Science as Culture* 14: 325 - 338. https://doi.org/10.1080/09505430500369020
Grace K (2015) The Asilomar conference: A case study in risk mitigation. *Technical report 2015–9. Machine Intelligence Research Institute.* Berkeley, CA. Available at: https://intelligence.org/files/TheAsilomarConference.pdf
Gregorowius D, Biller-Andorno N and Deplazes-Zemp A (2017) The role of scientific self-regulation for the control of genome editing in the human germline: The lessons from the Asilomar and the Napa meetings show how self-regulation and public deliberation can lead to regulation of new biotechnologies. *EMBO Rep* 18(3): 355-358. https://doi.org/10.15252/embr.201643054
Heikkila M (2023). What's changed since the 'pause AI' letter six months ago? *MIT Technology Review* September 26. Available at: https://www.technologyreview.com/2023/09/26/1080299/six-months-on-from-the-pause-letter/. (accessed August 25, 2024).
Hogan M (2015) Data flows and water woes: The Utah Data Center. *Big Data & Society* (July–December): 1–12 https://doi.org/10.1177/2053951715592429
Hurlbut JB (2015a) Limits of Responsibility: Genome Editing, Asilomar, and the Politics of Deliberation. *Hastings Center Report* 45(5): 11-14.
Hurlbut JB (2015b) Remembering the future: Science, law, and the legacy of Asilomar. In: Jasanoff S and Kim S-H (eds) *Dreamscapes of Moderity: Sociotechnical imaginaries and the fabrication of power*. Chicago, IL: University of Chicago Press, pp.126 - 151.
Jackson DA, Symons RH and Berg P (1972) Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of Escherichia coli. *Proc Natl Acad Sci U S A* 69(10): 2904-2909.
Jasanoff S and Hurlbut JB (2018) A global observatory for gene editing. *Nature* 555: 435-437.
Jasanoff S, Hurlbut JB and Saha K (2019) Democratic Governance of Human Germline Genome Editing. *Crispr j* 2(5): 266-271. https://doi.org/10.1089/crispr.2019.0047

Jobin A, Ienca M and Vayena E (2019) The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1(9): 389-399. https://doi.org/10.1038/s42256-019-0088-2

Krimsky S (1982) *Genetic Alchemy: The Social History of the Recombinant DNA Controversy.* Cambridge, Mas: MIT Press.

Labaree A (2014) Our science-fiction apocalypse: Meet the scientists trying to predict the end of the world. *Salon*, May10. Available at: https://www.cser.ac.uk/news/our-science-fiction-apocalypse-meet-scientists-try/ (accessed August 25, 2024).

Lewis-Kraus, G. (2016) The great A.I. awakening. The New York Times Magazine. Dec. 14. Available at: https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html (accessed August 25, 2024)

MacAskill W (2022) *What we owe the future.* New York: Basic Books.

Marchant G, Tournas L and Gutierrez CI (2020) Governing new technologies through soft law: Lessons for artificial intelligence. *Jurimetrics* 61(1): 1 - 18.

McKelvey F and Roberge J (2023) Recursive Power: AI Governmentality and Technofutures. In: Lindgren S (ed) *Handbook of Critical Studies of Artificial Intelligence*. London: Elgar, pp.21 - 32.

Milmo D (2023) Google, Microsoft, Open AI and start up form body to regulate AI development. *The Guardian*, July 16. Available at: https://www.theguardian.com/technology/2023/jul/26/google-microsoft-openai-anthropic-ai-frontier-model-forum (accessed August 25, 2024).

Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1, 501–507

Munn L (2023) The uselessness of AI ethics. *AI and Ethics* 3(3): 869-877. https://doi.org/10.1007/s43681-022-00209-w

Naughton J. (2022) Longtermism: how good intentions and the rich created a dangerous creed. *The Guardian* Dec. 4. Available at: https://www.theguardian.com/technology/commentisfree/2022/dec/04/longtermism-rich-effective-altruism-tech-dangerous. (accessed August 25, 2024).

Noble S. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York: NYU Press.

Parthasarathy S (2015) Governance Lessons for CRISPR/Cas9 from the Missed Opportunities of Asilomar. *Ethics in Biology, Engineering and Medicine: An International Journal* 6: 305-312.

Prabhakaran V, Mitchell M, Gebru T, Gabriel J. 2022. A human rights-based approach to responsible AI. *arXiv: 2210.02667* https://doi.org/10. 48550/arXiv.2210.02667

Richards B, Aguera Y, Arcas B, et al. (2023) The illusion of AI's Existential Risk. *Noema*, July 18. Available at: https://www.noemamag.com/the-illusion-of-ais-existential-risk/ (accessed August 25, 2024).

Rogers M (1975) The pandora's box congress. *Rolling Stone*, June 19. Available at: https://web.mit.edu/endy/www/readings/RollingStone(189)37.pdf. (accessed December 1, 2023)

Rufo F and Ficorilli A (2019) From Asilomar to Genome Editing: Research Ethics and Models of Decision. *NanoEthics* 13(3): 223-232.

Schiffer Z (2021) Google fires second ethics researcher following internal investigation. *The Verge*, February 19. Available at: https://www.theverge.com/2021/2/19/22292011/google-second-ethical-ai-researcher-fired. (accessed December 1, 2023)

Singer M and Soll D (1973) Guidelines for hybrid DNA molecules. *Science* 181: 1114.

Stapf-Fine H, Bartosch U, Bauberger S, et al. (2018) Policy Paper on the Asilomar Principles on Artificial Intelligence. June. Berlin: Federation of German Scientists. Available at: https://vdw-ev.de/wp-content/uploads/2019/05/Policy-Paper-on-the-Asilomar-principles-on-Artificial-Inteligence_end.pdf (accessed December 1, 2023)

Stirling B (2018) The Asilomar AI Principles. *Wired*, August 13. Available at: https://www.wired.com/beyond-the-beyond/2018/06/asilomar-ai-principles/ (accessed December 1, 2023)

Taylor C and Dewsbury B (2019) Barriers to inclusive deliberation and democratic governance of genetic technologies at the science-policy interface. *Journal of Science Communication* 18(3): Y02. https://doi.org/10.22323/2.18030402

Torres É (2021) Against Longtermism. *Aeon*. Available at: https://aeon.co/essays/why-longtermism-is-the-worlds-most-dangerous-secular-credo. (accessed December 1, 2023)

Wagner B (2018) Ethics As An Escape From Regulation.: From "Ethics-washing" To Ethics-shopping? In: Bayamlioğlu E, Baraliuc I, anssens J, et al. (eds) *Being Profiled*. Amsterdam: Amsterdam University Press, pp.84–89.

Williams A, Miceli M, and Gebru T. (2022) The exploited labor behind artificial intelligence. *Noema*. https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/

Zuboff S (2019) *The age of surveillance capitalism: The fight for a human future at the new frontier of power.* New York: The Hatchett Group.

