# AI Integration Task Force (AIITF)

Dr. Ron Dolin

ron.a.dolin@aiitf.org www.aiitf.org

Version 1.8, Sep 26, 2024

Executive Summary	1
Introduction	2
Tension	2
Resolution	7
Mission	7
Why This Approach?	7
Work Product	9
Structure	9
Related Organizations (Structural Elements)	
The Sedona Conference (TSC)	9
Internet Engineering Task Force (IETF)	10
Kavli Institute for Theoretical Physics (KITP)	
Spirit Rock Meditation Center	
AIITF	
PIPE-lines	
Pragmatic	
Interdisciplinary	
Public	
Expeditious	
Working Group Oversight Committee	
Conferences	
How You Can Help	
Leadership	16
My Background	
Officers & Directors	16
Working Group Oversight Committee	
Advisors	
Funding	17
Structurally-Related Organizations Budget Estimates	17
AIITF Budget Estimate	18
Potential Funders and Collaborators – Shared Goals	

MacArthur Foundation	18
Schmidt Sciences	19
Stanford's Center on Human-Centered Al	
Harvard's Berkman Klein Center	20
Oxford's Institute for Ethics in AI	20
Appendix A: AI-Generated short and long term legal topics	21
Appendix B: Creativity, AI, and IP Law	24
Appendix C: Examples of Immediate Real-World AI Problems	26
National Security	26
National Security Commission on Artificial Intelligence (2021)	26
Autonomous Weapons Systems	27
Facial Recognition	
Voting, Disinformation, and Voter Profiling	31
Brain-Computer Interfaces	32
Medicine	33
Outcome Research	
Insurance	
Use of AI within the Insurance Industry	
AI Risk Analysis	
Employment & Labor	
International Generally	
Entertainment Industry	
AI-Generated Music	
Motion Picture and Television Production	
Hiring, Recruiting, and Performance Evaluation	
Education	
K-12 In-Class Use of Al	38
College Admissions	39
Courts	40
Analysis of AI Creativity within IP Law	40
Use of AI by Judges	40
Legislation	41
Appendix D: Legal Pro Bono Compliance	42

"Men who have created new fruits in the world cannot create a system whereby their fruits may be eaten. And the failure hangs over the State like a great sorrow."

- John Steinbeck, Grapes of Wrath

# **Executive Summary**

The mission of the AI Integration Task Force is to provide immediate guidance to any and all components of society to *integrate* AI in a beneficial way that minimizes negative consequences. This guidance is intended to be interdisciplinary, practical, short-term, and non-partisan – geared toward users, not developers, of AI.

The development of sophisticated Artificial Intelligence (AI) is easy compared to figuring out how to *integrate* AI into society. This is not an issue of AI capability, nor of building safety features within various AI systems. The challenge is with constructing the plethora of applications – defense, education, labor, the arts, etc. There is a need to guide fields being disrupted by AI in how to revamp workflow, monetization, creativity, etc., so as to maximize the benefits of AI while minimizing the harms. The AI tools of today are not the AI tools of yesterday, and nothing compared to the AI tools of tomorrow – there have been, and will continue to be, quantum leaps in capabilities for many reasons. Throughout this proposal, we turn to AI to tell us about AI, and its descriptions are compelling:

The impact of this inflection point is still unfolding, and the future of AI holds both immense promise and potential challenges. As AI technology continues to advance, it is crucial to address ethical considerations and ensure that AI is developed and used responsibly for the benefit of society.<sup>1</sup>

The rapid rate of changes upon us due to AI would challenge any normal means by which we typically evolve as societies through legislation, courts, business strategies, etc. Given the scope and depth of the realignment being caused by AI, most, if not all, sectors of society would benefit from real-world, pragmatic, immediate, and continual guidance. Guidance might include positive and negative consequences of possible strategic decisions, new metaphors and paradigms to frame analyses, regulatory suggestions, economic and labor models, etc., all with an eye toward, say, the next 1-2 years. This proposal describes one approach to the rapid development of needed guidance, through the recently founded non-profit, the AI Integration Task Force.

<sup>&</sup>lt;sup>1</sup> From Gemini, Google's AI chatbot.

# Introduction

Al tells us that Al is a problem.

# Tension

The development of sophisticated Artificial Intelligence (AI) is easy compared to figuring out how to *integrate* AI into society. This is not an issue of AI capability, nor of building safety features within various AI systems. The challenge is with constructing the plethora of applications – defense, education, labor, the arts, etc. There is a need to guide fields being disrupted by AI in how to revamp workflow, monetization, creativity, etc., so as to maximize the benefits of AI while minimizing the harms. The AI tools of today are not the AI tools of yesterday, and nothing compared to the AI tools of tomorrow – there have been, and will continue to be, quantum leaps in capabilities for many reasons. Throughout this proposal, we turn to AI to tell us about AI, and its descriptions are compelling.<sup>2</sup>

According to Alphabet's (Google's) Al tool Gemini, there has been a significant change in the development of Al over the last few years:

There has been a significant inflection point in the development of AI over the last 10 years. This inflection point is marked by several key advancements:

1. The rise of deep learning: Deep learning, a subset of machine learning that uses artificial neural networks, has seen tremendous progress in recent years. Breakthroughs in algorithms, the availability of large datasets, and increased computing power have enabled deep learning models to achieve unprecedented performance in tasks like image recognition, natural language processing, and game playing.

2. The emergence of large language models (LLMs): LLMs like GPT-3 and GPT-4, trained on massive amounts of text data, have demonstrated remarkable capabilities in generating human-like text, understanding complex questions, and performing various language-based tasks. This has led to the development of powerful AI applications like ChatGPT, which have captured public attention and sparked widespread interest in AI.

3. Increased accessibility of AI tools: Cloud computing and the development of user-friendly AI platforms have made AI tools more accessible to businesses and individuals. This has

<sup>&</sup>lt;sup>2</sup> AI (Gemini) interactions are denoted with the Courier font.

democratized AI, allowing more people to experiment with and utilize AI technology for various purposes.

4. Growing investment in AI research and development: There has been a significant increase in investment in AI research and development, both from private companies and governments. This has fueled innovation and accelerated the pace of AI development.

The inflection point in AI development is characterized by a shift from narrow AI, designed to perform specific tasks, to more general-purpose AI systems that can handle a wider range of tasks. This has opened up new possibilities for AI applications in various industries, including healthcare, finance, education, and entertainment.

The impact of this inflection point is still unfolding, and the future of AI holds both immense promise and potential challenges. As AI technology continues to advance, it is crucial to address ethical considerations and ensure that AI is developed and used responsibly for the benefit of society.

Integrating AI into society is complex. The future impact of AI in the near term is likely immense and potentially as impactful as the Industrial Revolution<sup>3</sup>:

AI is poised to bring about significant changes across various aspects of our lives in the next decade or two. The scope and depth of these changes are likely to be far-reaching, impacting industries, economies, and societies on a global scale. Here's an overview of the potential areas, scope, depth, and measure of changes that could result from AI:

#### Areas of Change:

• Workforce and Employment: AI and automation could displace jobs in sectors like manufacturing, customer service, and transportation. While new jobs may emerge, significant workforce restructuring and upskilling initiatives may be required.

<sup>&</sup>lt;sup>3</sup> For a more detailed analysis by the insurance industry of concrete, near-term risks of the use of AI, see the <u>AI Risk Analysis</u> section under <u>Insurance</u> included in the examples in <u>Appendix C</u>.

- Healthcare: AI-powered diagnostics, drug discovery, and personalized medicine could revolutionize healthcare delivery, leading to improved patient outcomes and more efficient healthcare systems.
- Education: AI-driven personalized learning platforms could tailor education to individual student needs, enhancing learning outcomes and access to quality education.
- Scientific Research: AI could accelerate scientific discovery in fields like materials science, climate modeling, and drug development, leading to breakthroughs that address global challenges.
- **Transportation:** Autonomous vehicles could reshape transportation systems, improving road safety and reducing congestion.
- Finance: AI-powered algorithms could transform financial markets, enhancing trading strategies, risk assessment, and fraud detection.
- Entertainment and Media: AI could create personalized content recommendations, enhance virtual and augmented reality experiences, and even generate creative works like music and art.
- Environmental Sustainability: AI could optimize energy use, monitor environmental conditions, and model climate change scenarios, helping address environmental challenges.

#### Scope of Change:

- Global Impact: AI's impact is likely to be felt across the globe, with developing and developed countries alike experiencing both the benefits and challenges of AI adoption.
- **Cross-Industry Disruption:** AI's transformative potential extends across multiple industries, with few sectors likely to remain untouched by its influence.

Depth of Change:

- Fundamental Shifts: AI could lead to fundamental shifts in how we work, learn, communicate, and interact with technology.
- Societal Implications: The widespread adoption of AI could raise ethical, social, and economic questions about issues like privacy, bias, inequality, and human-machine collaboration.

#### Measure of Change:

- Economic Growth: AI could contribute significantly to economic growth, but its impact on income inequality and job displacement will need to be carefully managed.
- Quality of Life: AI has the potential to improve quality of life through enhanced healthcare, personalized services, and increased convenience, but its impact on social interaction and mental well-being warrants careful consideration.
- Global Challenges: AI could be a powerful tool for addressing global challenges like climate change, disease, and poverty, but its responsible development and deployment are crucial.

In conclusion, AI's potential impact is vast and multifaceted. It holds the promise of transformative advancements in numerous fields, but also raises important considerations about its ethical and societal implications. Careful planning, regulation, and collaboration will be essential to ensure that AI is developed and used responsibly for the benefit of all.

Similarly, Gemini tells us, there exists great concern among many thought leaders about how AI might impact humanity (with possible "hallucinations" included):

"The development of full artificial intelligence could spell the end of the human race... It would take off on its own, and re-design itself at an everincreasing rate. Humans, who are limited by slow biological evolution, couldn't compete, and would be superseded." - Stephen Hawking, BBC Interview, December 2014. "AI is likely to be either the best or worst thing to happen to humanity." - Max Tegmark, MIT physicist, Edge.org, January 2016.

"I'm increasingly inclined to think that there should be some regulatory oversight, maybe at the national and international level, just to make sure that we don't do something very foolish." - Bill Gates, Reddit AMA, March 2015.

"We are rapidly heading towards digital superintelligence that far exceeds any human. I think it's very important that we have a serious discussion about how we manage the transition." - Sam Altman, CEO of OpenAI, The Information, May 2023.

"One of the biggest risks to the future of civilization is AI." - Geoffrey Hinton, 'Godfather of AI', MIT Technology Review, May 2023.

"I think the dangers of AI are real...I worry about an arms race in lethal autonomous weapons." - Stuart Russell, Computer scientist and AI researcher, TED Talk, April 2017.

"AI doesn't have to be evil to destroy humanity - if AI has a goal and humanity just happens to come in the way, it will destroy humanity as a matter of course without even thinking about it, no hard feelings." - Eliezer Yudkowsky, AI researcher and writer, Machine Intelligence Research Institute, 2008.

"If people are worried about AI taking over the world, they should be worried about corporations that are not held accountable to anyone taking over the world through AI." -Yoshua Bengio, Turing Award winner and AI pioneer, Wired, June 2023.

"I am concerned about the potential for misuse of AI, particularly in the areas of autonomous weapons systems and deepfakes." - Demis Hassabis, CEO and co-founder of DeepMind, Financial Times, July 2023.

Data presented in the annual <u>AI Index</u> from Stanford University's <u>HAI</u> ("Human-Centered Artificial Intelligence") gives a detailed account of the data behind, and confirmation of, Gemini's assessments. Certainly sci fi has explored for years the possibility of AI taking over the world, in stories such as <u>Colossus</u> (book – 1966, movie – 1970). For a detailed discussion of the potential for large-scale or even catastrophic harm from improperly managed AI, by Turing award winner Yoshua Bengio, see "<u>Reasoning through arguments against taking AI safety seriously</u>" (2024).

Given the stakes, simply stating the need for guidance, by, uh, *someone* – is wholly insufficient.

Where guidance is needed, and quickly, how can we provide it?

# Resolution

The rapid rate of changes upon us due to AI would challenge any normal means by which we typically evolve as societies through legislation, courts, business strategies, etc. Given the scope and depth of the realignment being caused by AI, most, if not all, sectors of society would benefit from real-world, pragmatic, immediate, and continual guidance. Guidance might include positive and negative consequences of possible strategic decisions, new metaphors and paradigms to frame analyses, regulatory suggestions, economic and labor models, etc., all with an eye toward, say, the next 1-2 years. This proposal describes one approach to the rapid development of needed guidance, through the recently founded non-profit, the AI Integration Task Force.

This approach, described in detail below, allows for an accelerated framework for the identification, prioritization, development, and public dissemination of needed guidance. The guidance needed for one sector may be wholly different than for another one. For example, the legal system needs to consider AI creativity from an intellectual property perspective – what are the ways to consider AI as an inventor, or how might training infringe on copyright? National security considerations include the degree to which autonomous systems should be allowed to attack people in war zones without human confirmation. Businesses and labor organizations need to consider tradeoffs between automation and employment. All of these issues arise in the context of quickly advancing AI capabilities, such that guidance today might be wholly obsolete tomorrow. Thus, the proposed framework provides for the rapid development of short-term guidance addressing the issues identified on a case-by-case basis.

# Mission

The mission of the AI Integration Task Force is to provide immediate guidance to any and all components of society to *integrate* AI in a beneficial way that minimizes negative consequences. This guidance is intended to be interdisciplinary, practical, short-term, and non-partisan – geared toward users, not developers, of AI.

# Why This Approach?

"It is a field of fields ... it holds the secrets which will reorganize the life of the world."

-Thomas Edison (speaking about the coming impact of electricity)

There is not one singular AI. AI integration is many techniques applied to many fields. It's herding cats – toward benevolence and away from harm in the choice between utopia and dystopia. Given the broad range of impact across presumably all components of society, there is a need for interdisciplinary teams to help guide decisions:

- Computer and data scientists who understand relevant aspects of the technology
- Domain experts who understand the subject matter
- Business analysts to explore impacts on monetization and labor
- Other stakeholders impacted by potential choices

As an example, taken from several listed in <u>Appendix C</u>, I was part of an ad hoc working group organized by an IP attorney to explore human and AI creativity. The purpose of the working group was to guide the legal system in the analysis of attempts to allow AI to be included alongside human copyright holders or patent inventors. To properly address this, we needed a team that included not only lawyers and computer scientists, but also philosophers and psychologists (and, ideally, cognitive scientists), in order to compare and contrast human vs. artificial "creativity". The result, obtained in about two months, is included in <u>Appendix B</u>. In that case, the team found that AI does not yet exhibit the type of creativity required by law to allow the creator to be granted such exclusive rights. The paper gives guidance regarding how to interpret existing regulations, what their purpose is, why they might (likely) need to be changed in the future, and which AI capabilities to watch for that would necessitate such regulatory changes. The resulting 40-page white paper is just one of the hundreds, if not thousands, that are needed quickly (and repeatedly, due to the fast-changing nature of AI) to help guide the integration of AI into multiple fields.

The earlier one sets direction, the easier it is to reach the desired destination. In this case, given the scale and scope of the impact of AI, many industries are struggling to figure out how to deal with pressing problems in the near term. Thus, it is imperative that guidance comes quickly, that it is practical, that it deals with immediate, short-term dilemmas, and that it is done transparently and includes all major stakeholders.

The components of this work that are required for a successful result include:

- 1. A leader who
  - a. Proposes (or accepts) a topic and thesis
  - b. Helps select and organize a team
  - c. Assumes editorial ownership and authorship as needed
- 2. An interdisciplinary team that can prioritize this effort and travel as needed
- 3. Focus on a short term need (1-2 years out)
- 4. Produce work product with fast turnaround (2 months from kickoff meeting to final paper)
- 5. Work from a single location for some short period of time (e.g. one week)

# Work Product

The work product from this effort will be a series of publicly available guidance documents (e.g. 20+ pages) of core problem(s), possibly with 3-5 practical, immediate proposals. This output will help stakeholders identify the range of options to address issues of concern, and to explain potential pros and cons of various approaches.

The immediate goal is to produce one such document per week on an ongoing basis. An example of such a guidance document is given in <u>Appendix B</u>, which was produced by an ad hoc team at Oxford University in approximately 2 months, as a test case of this methodology. This particular white paper frames how judges, lawyers, and litigants might frame "creativity" from a legal perspective, and how to compare notions of human vs. AI creativity with the context of intellectual property law. Some exemplary problems in need of this type of immediate guidance are given in <u>Appendix C</u>.

# Structure

The structure of AIITF is designed to facilitate fast, pragmatic guidance in a manner that scales to the size of the problem. While there are some unique aspects of the workflow, several proposed elements are well-established mechanisms that have proven to be very successful in other organizations. Therefore, before describing AIITF, we present some of these specific procedures that highlight particular aspects of what we hope to implement.

# **Related Organizations (Structural Elements)**

Various components of this mechanism can be seen in other, very successful and long-established, organizations. These structural elements are highlighted here for comparison, as well as to demonstrate ways in which they have been shown to work well in other contexts. (Where not clearly stated on the relevant websites, the approximate founding dates were provided by Gemini.)

# The Sedona Conference (TSC)

The Sedona Conference (TSC), founded in 1997, is an organization that informs courts and the legal system generally about ways in which complex technology like e-discovery tools should be integrated into the legal process. Its <u>mission</u> is "to move the law forward in a reasoned and just way."

"The Sedona Conference believes that the combined knowledge of experts representing varied viewpoints, focused through the "magic" of dialogue outside an adversarial setting, can achieve critically important consensus on the most difficult leading edge issues. TSC brings together the brightest minds in a think-tank sett to ing with the goal of creating practical solutions and recommendations. Their findings are developed and enhanced through a substantive peer-review process and the resulting content widely published in conjunction with educational programs for the bench and bar, so that it can swiftly drive the reasoned and just advancement of law and policy in the areas under study." [...]

"The Sedona Conference accomplishes its mission primarily through the activities of its Working Groups, which publish nonpartisan consensus commentaries, primarily in the form of principles, guidelines, and best practices, designed to offer practical solutions on tipping-point issues. As these commentaries are recognized, cited, and followed, they move the law forward in a reasoned and just way."

By combining a multidisciplinary group of experts and public dissemination, The Sedona Conference has been instrumental in guiding the legal system through the adoption of complex tools and protocols, including how judges should consider, allow, and even require their use in appropriate circumstances.

## Internet Engineering Task Force (IETF)

The Internet Engineering Task Force, founded in 1986, is responsible for many of the well-known standards and protocols used to run the internet, such as TCP/IP, DNS, email, etc. It's <u>mission</u> is to continue to develop practical, party-neutral (though not value-neutral) standards:

"The Internet Engineering Task Force (IETF), founded in 1986, is the premier standards development organization (SDO) for the Internet. The IETF makes voluntary standards that are often adopted by Internet users, network operators, and equipment vendors, and it thus helps shape the trajectory of the development of the Internet. But in no way does the IETF control, or even patrol, the Internet." [...]

"There is no membership in the IETF. Anyone can participate by signing up to a working group mailing list (more on that below), or registering for an IETF meeting. All IETF participants are considered volunteers and expected to participate as individuals, including those paid to participate.

"The IETF welcomes all interested individuals and participants come from all over the world and from many different parts of the Internet industry. In any one year, over 7000 people actively participate in the IETF either by authoring a document, engaging in a mailing list discussion, or attending a meeting."

The IETF publishes their protocols in what they call "RFC's" (Request For Comments), in an open and public process. New protocols are originated by any participant via the drafting of an initial document, and comments, development, and adoption are left to interested parties on a voluntary basis. Additionally, of note is the way that IETF handles conferences:

"The IETF holds three week-long meetings a year. Onsite participation averages between 1000 and 1500 participants. Every effort is made to integrate the 600+ remote participants into the overall meeting experience.

"IETF meetings are very different from standard industry conferences as they are working meetings with the primary goal of helping Working Groups (WGs) get their tasks done, and the secondary goal of promoting mixing among the WGs. For that reason, most of the agenda is comprised of multiple simultaneous sessions for Working Groups. There is no exposition hall and only a small section of the agenda is set aside for tutorials, industry presentations, panel sessions, and opportunities for the whole community to come together."

Thus, the IETF utilizes a very public mechanism to produce highly useful guidance in the development of technology of crucial importance the world over.

# Kavli Institute for Theoretical Physics (KITP)

The <u>mission</u> of the Kavli Institute for Theoretical Physics at UCSB, founded in 1979, is to bring together select physicists and related scientists from around the world for short and focused sessions to address key problems.

"The Kavli Institute for Theoretical Physics (KITP) is the first and foremost scientific research facility where theorists in physics and allied fields congregate, for sustained periods of time, to work together intensely on a broad range of questions arising from investigations at the leading edges of science."

"Most programs include a 4-day conference, particularly attractive to experimentalists preferring short trips away from their laboratories."

"The number of participants in KITP programs and conferences averages 1,000 a year."

"Initially conceived principally as a national center for theoretical physics, the KITP has evolved a model for collaboration that has been widely imitated both by other disciplines (mathematics, for example) and by other countries."

The KITP highlights the benefits of gathering experts together for an intense, short period to work through initial details, prior to returning to their home organizations to finish the work. Progress is hampered when participants are juggling other priorities, or working across international timezones, especially in the initial phase of the discussion.

## Spirit Rock Meditation Center

The <u>mission</u> of Spirit Rock Meditation Center, founded in 1987, is to promote, teach, and facilitate the practice of meditation and mindfulness. The Center runs <u>continuous week-long</u> <u>retreats</u> involving several hundred participants on an ongoing basis.

"Since ancient times all wise cultures have known the value of retreat. Time in retreat allows us to step out of the complexity of our life, to listen deeply to our body, heart, and mind. For 2,600 years, meditation retreats have been a central part of the Buddhist path of awakening."

The relevant aspect of this program is the manner in which these retreats are planned and executed, such that the dormitories are basically never idle. There is continuous training and practice involving the coordination of trainers, center staff, and participants. Planning starts well in advance in order to maximize the benefit of the resources of the location and the time and effort of those involved.

# AIITF

AIITF (aiitf.org) has requested non-profit status at the end of July, pending IRS approval. The proposed structure of AIITF is designed to facilitate a PIPE-line of guidance (see below), moving several working groups along in parallel. Topic proposals are either submitted or solicited, and approved by a Working Group Oversight Committee (described below). Proposals must include a team lead, who is responsible for coordinating the interdisciplinary team members' participation (with assistance from, and in cooperation with, AIITF staff). Teams are intended to consist of 5-10 members. The goal is to bring together team members for several consecutive

days (e.g. one week) in an intensive workshop, ideally with travel, accommodations, and meeting logistics arranged and funded by AIITF. By the end of the in-person session, the team should have a rough outline of a guidance document and possibly a sketch of 3-5 concrete proposals.

Depending on the subject matter, proposals may be issues and trade-offs to consider, factors that must be included, or a framework within which stakeholders can work through solutions particular to their own organizations. The working group is expected to complete a final draft of the guidance within two months, with AIITF facilitating online weekly meetings through completion. AIITF would then disseminate the final guidance publicly. The goal is to run several working groups in parallel, such that each week a new working group would meet in person to kick off a guidance process. Once the PIPE-line gets going, this could facilitate the creation of approximately one guidance document per week. The mechanism itself can scale up depending on interest, availability of experts, and funding. If funding is available, conferences would be quite helpful to promote and improve the guidance in coordination with a broader range of stakeholders.

#### **PIPE-lines**

What we're calling the PIPE-line method is a framework for workflow of the guidance. PIPE stands for Pragmatic, Interdisciplinary, Public, and Expeditious. In addition to the team descriptions below, AIITF staff would assist in the kickoff meetings to help facilitate interdisciplinary dialog and to promote <u>design thinking and methodology</u>.

#### Pragmatic

**Pragmatic** means that the guidance should be something that matters to the day-to-day operations and issues for whatever topic is being addressed. As Example 1, how should judges view creativity in terms of assigning any rights like copyright or patent inventorship to AI that contributes to the development of intellectual property? Should they allow it or not, why, and what would need to happen to change the current legal perspective, from both the technology and law components? As Example 2, what might musicians and studios want to consider, in the present day, in terms of artist compensation and the use of their music in training AI, and what should they be looking for in terms of encouraging and protecting human-based creativity?

#### Interdisciplinary

*Interdisciplinary* means that the small teams that are addressing issues should be aware of the major factors and fields that inform the guidance. In Example 1, the team might need to consist of a computer scientist, a law professor, a psychologist, a philosopher, and a cognitive scientist, in order to understand what creativity is, how humans do it, whether AI follows similar processes, and what current law is based on. In Example 2, the team might need to include a

musician, a couple of studio executives from different organizations, a business analyst, and a computer scientist.

#### Public

*Public* means that, at a minimum, the guidance is made easily available to the general public. AITF seeks to disseminate guidance widely. Ideally, if conferences are included, there might be a draft version pending feedback from a wider audience (similar to the way that the IETF handles RFC's within their conferences).

#### Expeditious

*Expeditious* means that at least an initial public draft of the guidance should be made available within 2-3 months of the initial team kickoff. Furthermore, the guidance should focus on the near term, say 1-2 years.

Since the model entails a short, in-person gathering of each team, from 2-7 days, as the kick-off to a 2-month drafting process, it is envisioned that these teams will be working in parallel, moving through the following steps:

- 1. Topic suggestion or solicitation (1-3 month prior to kick-off)
- 2. Topic approval
- 3. Team formation
- 4. Kick-off meeting
- 5. Publication of guidance document (within 2 months of kick-off)
- 6. Conference participation in the first conference scheduled after publication

As envisioned, assuming a full PIPE-line, there could be dozens of guidance documents per year. An initial stretch goal would be one document per week, with up to four conferences per year.

# Working Group Oversight Committee

A Working Group Oversight Committee will be responsible for processing potential AIITF topics. There will be a submission process, and, likely, an outreach and solicitation effort. The criteria under which proposals for working group teams will be selected will need to be established by the oversight committee, but will likely include the following factors:

- 1. The importance of the topic
- 2. The urgency of the topic

- 3. The proposed topic's working group team or, possibly, the likelihood that a reasonable team can be formed
- 4. The qualifications of the team leader
- 5. Availability and level of commitment of the team members

As an example, a list of known AI problems within the legal system, both short-term and long-term, is included as <u>Appendix A</u>. A similar list of problems could easily be generated for most, if not all, components of society. Given the nature of the proposed conferences, described below, one could imagine themed tracks, such that a handful of topics from a single sector might be selected, with a few sectors running in parallel (e.g. law, music, education).

AIITF has started reaching out to various organizations to solicit topics and team participants. As an example of including legal considerations within a team, <u>Appendix D</u> highlights the reasons that working group participation by attorneys qualifies as pro bono time as defined by state bars.

# Conferences

An important component in the success of guidelines is their salience to existing problems, their practicality, and, ultimately, their adoption. Therefore, merely posting guidance established by a select committee, regardless of the quality, is not likely to address the full range of issues needed to be addressed by the array of stakeholders not immediately participating in the working group. As with the IETF, an effective means with which to engage a broader audience of stakeholders, and to facilitate bi-directional communication and dissemination of the guidance, is via a conference open to the public. If a PIPE-line of topics are developed, the conferences would be partitioned along the sectors within which guidance has been developed. Obviously, this is wholly dependent on the degree to which this methodology is successful.

# How You Can Help

There are several ways you might be able to help this effort, depending on your contacts, professional position, etc.

- Identification of issues and potential Working Group team members
- Volunteer:
  - Working Group team participation
  - AIITF staffing (website development, Working Group organization, outreach)
- Funding

Currently Active Working Group: Can/should judges use AI? (see Appendix C for details).

Contact us at info@aiitf.org.

# Leadership

# My Background

I have a broad interdisciplinary <u>background</u> that includes science, technology, business, philosophy, law, design, and more. I have a BA in math and physics, a Ph.D. in computer science, and a JD. I was one of the first 50 employees at Google, working as a software developer, team leader, and project/program manager for years, working across multiple departments. I started the quality evaluation team, and have written and spoken extensively about quality metrics. Prior to grad school, I worked at CERN in Geneva, Switzerland for several years, with people from across the world. I am a licensed attorney in California. I have taught at Stanford, Notre Dame, and Harvard law schools. I co-edited and co-wrote *Legal Informatics* (Cambridge University Press, 2021). I have had training as a communicator, and have managed multi-million dollar projects to completion involving multiple diverse teams. I have co-led design workshops. Finally, I've been an angel investor for over a decade, often focusing on legal tech startups.

Upon funding, it is my intention to dedicate full-time to AIITF.

# **Officers & Directors**

Ron Dolin, President & CEO Stephanie Kimbro, Treasurer Chris Mammen, Secretary

Directors: Ron Dolin, Mary O'Carrol, Chris Mammen

# Working Group Oversight Committee

Chair: ZZZ Members: TBD

# Advisors

# Funding

# Structurally-Related Organizations Budget Estimates

**Prompt:** What are the approximate or estimated annual budgets for the following non-profits or on-campus centers? KITP (UCSB), IETF, The Sedona Conference, Spirit Rock Meditation Center. Please give your best guess and tell me your source.

I cannot give you the exact annual budgets for these non-profits, but here are some estimates based on publicly available information and general knowledge of similar organizations:

- **KITP (UCSB):** This is a research center, and funding likely comes from a mix of government grants, private donations, and university funds. A reasonable estimate might be in the range of **\$5-10 million** annually. The source for this is their annual reports, which mention funding sources but not exact figures.
- IETF: This is an open standards organization. Their budget likely comes from membership fees and donations. A reasonable estimate might be in the range of **\$1-5 million** annually. The source for this is their website, which lists members but not a detailed budget.
- The Sedona Conference: This is a legal education non-profit. Their budget likely comes from conference fees, publications, and donations. An estimate could be in the range of \$2-5 million annually. The source for this is their website, which provides some information about their programs and funding.
- Spirit Rock Meditation Center: This is a meditation retreat center. Their budget likely comes from retreat fees, donations, and program fees. An estimate could be in the range of \$10-20 million annually. This estimate is based on the size of their facility and the variety of programs they offer. The source for

this is their website, which provides information about their programs and retreats.

```
(These are just educated guesses, and the actual budgets are unknown.)
```

# AIITF Budget Estimate

There is currently no formal plan for budgeting AIITF. Given the scope of the proposal, if it were successful, budget items would roughly include the following:

- 1. 5-10 full-time staff and related office, travel, etc., expenses (e.g. \$2M/year)
- Travel, lodging, and meal expenses for working group team kickoff sessions. (e.g. \$1M/year)
- Conference costs (1-4 per year, depending on the success of filling the PIPE-line note that conferences along the lines of IETF do not include trade shows or the like) (e.g. \$1M/year)
- 4. IT and hosting services for guidance documents (e.g. \$1M/year)

Given the estimated budgets of similarly situated organizations, by comparison, a reasonable annual AIITF budget could be in the range of \$2M to \$10M annually, depending on the level of activity.

# Potential Funders and Collaborators – Shared Goals

Several organizations have aligned interests with the AIITF, some of which might be available for funding, and others for various types of collaborations:

MacArthur Foundation

Technology in the Public Interest:

Central to this work is supporting research, policy development, and practice that aims to uphold public interest considerations in the development and governance of artificial intelligence (AI). [...]

Al is being deployed across sectors with too little oversight and accountability, including high-stakes areas such as healthcare, finance, law enforcement, and education. While often touted as neutral, a growing body of interdisciplinary and intersectional research demonstrates that AI systems can replicate and amplify existing biases in society that uphold racism, sexism, White supremacy, and other forms of structural oppression. Moreover, AI-related technologies play a major role in determining what we read, see, watch, and listen to on digital platforms and search engines, but increasingly powerful technology companies use them to optimize clicks and views to maximize their profits.

Beneath the veneer of new and emerging technology is an old story about power and how it operates. Too often, the changes driven by AI and other technologies create and augment existing power asymmetries in society. Addressing these challenges requires supporting and expanding a collaborative and diverse ecosystem of people, organizations, and networks advancing a different vision for technology. A vision that is rooted in equity, justice, and other public interest considerations. Technology in the Public Interest grantmaking is a response to these dynamics.

## **Schmidt Sciences**

#### AI2050: Working List of Hard Problems in AI:

Drawing on previous work in AI, and through numerous conversations with other experts, the initiative has developed an initial working list of the hard problems for AI2050 to take on. This list is aimed at realizing the opportunity for society from AI and addressing the risks and challenges that could result from it. [Some of these problems are presented in <u>Appendix C</u>.]

While we believe the opportunities and challenges described in the working list are multidisciplinary, they are generally aimed at hard scientific and technical problems and societal challenges of different kinds that represent both opportunities and challenges. The list aims at relatively distinct categories of challenges and opportunities to solve.

This working list makes no claim to being comprehensive, final, or fixed in time. We fully expect such a list to continue to evolve as we learn more and as AI's capabilities progress and our use of it continues to evolve. We plan to update this list over time, revising current categories, including subcategories, and potentially introducing new categories of hard problems to solve guided by the motivating question.

## Stanford's Center on Human-Centered AI

#### Mission:

Al has the potential to affect every aspect of our lives and our civilization, from social bonds and ethics to the economy and healthcare, education and government. The faculty and staff of HAI are engaging not only with leading-edge scientists, but also with scholars trying to make sense of social movements, educators enhancing pedagogy, lawyers and legislators working to protect rights and improve institutions, and artists trying to bring a humanistic sensibility to the world in which we live. Together we're helping build the future of AI.

## Harvard's Berkman Klein Center

#### Ethics and Governance of AI:

The rapidly growing capabilities and increasing presence of AI-based systems in our lives raise pressing questions about the impact, governance, ethics, and accountability of these technologies around the world. How can we narrow the knowledge gap between AI 'experts' and the variety of people who use, interact with, and are impacted by these technologies? How do we harness the potential of AI systems while ensuring that they do not exacerbate existing inequalities and biases, or even create new ones?

## Oxford's Institute for Ethics in AI

#### Ethics in AI – Why it matters:

Philosophers made a major contribution to the development of medical ethics 40 years ago, and we are now at a tipping point where a similar ethical intervention is needed to cope with the questions raised by the rise of AI.

Every day brings more examples of the ethical challenges posed by AI, from face recognition to voter profiling, brain machine interfaces to weaponised drones, and the ongoing discourse about how AI will impact employment on a global scale.

This is urgent and important work that we intend to promote internationally as well as embedding in our own research and teaching here at Oxford.

# Appendix A: AI-Generated short and long term *legal* topics

#### From Gemini:

**Prompt 1:** What topics might be included in the *short*-term legal integration of AI into society generally?

#### Response:

In the short-term legal integration of AI into society, several key topics would likely be addressed:

- Liability and Accountability: Determining who is responsible if an AI system causes harm or damage. This could involve manufacturers, developers, owners, or even the AI system itself.
- **Transparency and Explainability:** Requiring AI systems to be transparent in their decision-making processes and providing explanations for their actions. This is especially important in high-stakes areas like healthcare or criminal justice.
- Bias and Discrimination: Addressing the potential for AI systems to perpetuate or amplify existing biases in data and algorithms. Regulations may mandate regular audits and testing for bias.
- Data Privacy and Security: Establishing guidelines for the collection, storage, and use of data by AI systems. This would involve balancing innovation with individual privacy rights.
- Intellectual Property: Determining ownership of creations generated by AI systems, such as artwork or music. This would have implications for copyright laws and creative industries.
- **Consumer Protection:** Ensuring that AI-powered products and services are safe, reliable, and accurately represented. This could involve labeling requirements or standards for testing and certification.
- Employment and Labor: Addressing the impact of AI on the workforce, including job displacement and the need for retraining programs. This could involve changes to labor laws and social safety nets.

- **Regulation and Oversight:** Establishing regulatory frameworks for AI development and deployment. This could involve new agencies, standards, or certification processes.
- International Cooperation: Collaborating with other countries to develop consistent global standards for AI governance. This would prevent regulatory arbitrage and ensure a level playing field for AI development.
- Alternative Dispute Resolution (ADR): How to integrate Al into conflict resolution processes, including mediation and arbitration. [not included in Gemini response]
- Judicial Applications: How should courts integrate the use of AI into the judicial system, especially regarding its use by judges? [not included in Gemini response]

In addition to these broad topics, specific areas of AI integration would likely require their own tailored regulations. For example, autonomous vehicles would need rules regarding safety, insurance, and liability, while AI in healthcare would need guidelines for patient consent and data privacy.

**Prompt 2:** What topics might be included in the *long*-term legal integration of AI into society generally?

#### Response:

The long-term legal integration of AI into society would likely delve deeper into the ethical and societal implications of widespread AI adoption, building upon the foundational issues addressed in the short term. Key topics that may emerge include:

- AI Personhood and Rights: Exploring the potential for granting legal personhood to highly advanced AI systems, along with corresponding rights and responsibilities. This could involve defining consciousness, sentience, and agency in AI.
- Human-AI Interaction and Relationships: Establishing legal frameworks for interactions between humans and AI, including issues like emotional attachment, dependency, and potential harm. This could involve regulations on AI companionship or caregiving roles.
- Global AI Governance: Developing international agreements and frameworks for AI governance to ensure ethical and responsible AI development and use on a global scale. This could involve

addressing issues like AI arms races, weaponization, and misuse.

- AI and Democracy: Examining the impact of AI on democratic processes, including disinformation, manipulation of public opinion, and potential biases in algorithmic decision-making. This could involve regulations on AI use in political campaigns and elections.
- AI and Economic Inequality: Addressing the potential for AI to exacerbate economic disparities by automating jobs and concentrating wealth in the hands of a few. This could involve policies on universal basic income, wealth redistribution, and retraining programs.
- AI and Existential Risk: Considering the potential risks of highly advanced AI systems surpassing human intelligence and posing a threat to humanity's existence. This could involve developing safety protocols and regulations on AI research and development.
- AI Ethics and Morality: Deepening the discourse on AI ethics and morality, addressing questions about values, consciousness, and the role of AI in society. This could involve creating ethical guidelines for AI development and use.

In addition to these broader issues, long-term legal integration would also likely involve ongoing adaptation and refinement of existing regulations to address the evolving capabilities and applications of AI. As AI technology continues to advance, new legal challenges and opportunities will inevitably arise, requiring ongoing collaboration between policymakers, technologists, ethicists, and society as a whole to ensure the responsible and beneficial integration of AI into our lives.

# Appendix B: Creativity, AI, and IP Law

<u>Creativity, Artificial Intelligence, and the Requirement of Human Authors and Inventors in</u> <u>Copyright and Patent Law</u> (Mammen et al., Oxford, 2024)

(https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=4892973)

Abstract:

Copyright and patent law require the identification of an author or inventor, and further require the author or inventor to be human. We explore this requirement primarily with reference to U.S. law and provide additional illustrations from U.K. and E.U. law. A key rationale underlying the requirement of a human author or inventor is that there is something special and important about human creativity.

As AI, particularly generative AI, becomes more capable of producing outputs that look like they could have been human-created, arguments have increasingly been raised that the AI-generated outputs should be afforded copyright and patent protection, on the same basis as those made by human authors and inventors. And there have been arguments that these AI-generated outputs exhibit sufficient creativity, novelty, or innovativeness, to satisfy the laws' underlying creativity rationale.

We examine the concept of creativity from a multidisciplinary perspective, and identify three conceptually distinct components, all of which are necessary for a complete account of creativity. The external component refers to whether an artifact (or idea, or other thing) exhibits the qualities of being novel, valuable, and (on some accounts) surprising. The subjective component focuses on the psychological process of a creative act, which appears to involve a dance between task-focused and mental-wandering states, mediated by a salience functionality, where the person recognizes and selects novel, appropriate ideas. Third, embedded in the analysis of both the external and subjective components is a (largely-implicit) recognition that the social context is integral to creativity; it plays a role in determining whether an artifact has value (or is "appropriate"), and influences the subjective psychological process of plucking certain ideas or conceptions out of the flow of mental activity.

With this enriched account of creativity, we examine how copyright and patent law value not only the creativity of the artifact, but also (to varying extents) the subjective role and social context as part of creativity. We then consider some ways in which arguments that AI-generated artifacts should be eligible for IP protection (e.g., because they are "just as good as" at least some human-generated and IP-eligible artifacts) are insufficient to satisfy the enriched understanding of the creativity requirement underlying the IP laws.

Our investigation additionally reveals some themes that may warrant further and deeper examination. First, copyright has not historically concerned itself with the quantum or quality of

creativity in a human-created work; it just has to be non-zero. Some scholars have raised concerns that this standard has been too indulgent of copyright protection and have argued that the bar of creative quality should be raised. Placed in the context of generative AI's influence, such an argument raises intriguing possibilities, insofar as there is evidence that AI tends to narrow the band of creativity (both human creative thought and expression as well as the "creativity" of AI-generated outputs), hindering or suppressing what might otherwise be a wider range of creative possibility. Second, as AI becomes ever-more-capable, displacing not only traditionally human activities but also traditional human decisions and judgments, it may be necessary to consider fundamental revisions to some of our legal frameworks to accommodate this change.

# Appendix C: Examples of Immediate Real-World AI Problems

The <u>Schmidt Sciences</u> organization has described a set of <u>hard problems</u> in its <u>Al2050</u> initiative (James Manyika, Schmidt Sciences, 2024). These are core problems that they believe need to be addressed over the next couple of decades to unlock the benefits of Al and minimize the potential harms. It addresses "Al's continually evolving safety and security, robustness, performance, output challenges and other shortcomings that may cause harm or erode public trust of Al systems, especially in safety-critical applications and uses where societal stakes and potential for societal harm are high."

Examples include bias and fairness, toxicity of outputs, factuality/accuracy, information hazards including misinformation, reliability, security, privacy and data integrity, misapplication, intelligibility, and explainability, social and psychological harms.

It goes on to discuss "challenges of safety and control, human alignment and compatibility with increasingly powerful and capable AI and eventually AGI."

Examples include risks associated with tool-use/connections to physical systems, multi-agent systems, goal misspecification/drift/corruption, risks of self-improving/self-rewriting systems, gain of function risks and catastrophic risks, alignment, provably beneficial systems, human-machine cooperation, challenges of normativity and plasticity.

Below, we list only a few examples of the many currently active AI issues, ranging from national security, human-computer interfaces, labor, and more. They all exhibit a need for pragmatic, interdisciplinary, public, expeditious guidance.

# **National Security**

## National Security Commission on Artificial Intelligence (2021)

"The Department of Defense (DoD) should [...] *establish the foundations for widespread integration of AI by 2025.* This includes building a common digital infrastructure, developing a digitally-literate workforce, and instituting more agile acquisition, budget, and oversight processes." [emphasis added]

## Autonomous Weapons Systems

Losing Humanity: The Case against Killer Robots (Human Rights Watch and Harvard Law School's International Human Rights Clinic, 2012):

A relatively small community of specialists has hotly debated the benefits and dangers of fully autonomous weapons. Military personnel, scientists, ethicists, philosophers, and lawyers have contributed to the discussion. They have evaluated autonomous weapons from a range of perspectives, including military utility, cost, politics, and the ethics of delegating life-and-death decisions to a machine. According to Philip Alston, then UN special rapporteur on extrajudicial, summary or arbitrary executions, however, "the rapid growth of these technologies, especially those with lethal capacities and those with decreased levels of human control, raise serious concerns that have been almost entirely unexamined by human rights or humanitarian actors." It is time for the broader public to consider the potential advantages and threats of fully autonomous weapons.

Pros and Cons of Autonomous Weapons Systems (Amitai Etzioni & Oren Etzioni, Military Review, 2017):

We find it hard to imagine nations agreeing to return to a world in which weapons had no measure of autonomy. On the contrary, development in AI leads one to expect that more and more machines and instruments of all kinds will become more autonomous. Bombers and fighter aircraft having no human pilot seem inevitable. Although it is true that any level of autonomy entails, by definition, some loss of human control, this genie has left the bottle and we see no way to put it back again.

<u>The Moral Case for the Development and Use of Autonomous Weapon Systems</u> (Erich Riesen, Journal of Military Ethics, 2022):

In this article, I provide the positive moral case for the development and use of supervised and fully autonomous weapons that can reliably adhere to the laws of war. Two strong, *prima facie* obligations make up the positive case. First, we have a strong moral reason to deploy AWS (in an otherwise just war) because such systems decrease the psychological and moral risk of soldiers and would-be soldiers. Drones protect against lethal risk, AWS protect against psychological and moral risk in addition to lethal risk. Second, we have a *prima facie* obligation to develop such technologies because, once developed, we could employ forms of non-lethal warfare that would substantially reduce the risk of suffering and death for enemy combatants and civilians alike. These two arguments, covering both sides of a conflict, represent the normative hill that those in favor of a ban on autonomous weapons must overcome. Finally, I demonstrate that two recent objections to AWS fail because they misconstrue the way in which technology is used and conceptualized in modern warfare.

# **Facial Recognition**

Racial Discrimination in Face Recognition Technology (Alex Najibi, Harvard's Science in the News, 2020):

Police use face recognition to compare suspects' photos to mugshots and driver's license images; it is estimated that almost half of American adults – over 117 million people, as of 2016 – have photos within a facial recognition network used by law enforcement. This participation occurs without consent, or even awareness, and is bolstered by a lack of legislative oversight. More disturbingly, however, the current implementation of these technologies involves significant racial bias, particularly against Black Americans.

<u>Face recognition struggles to recognise us after five years of ageing</u> (Matthew Sparkes, New Scientist, 2022):

As we age, our faces change, becoming slowly different from images used for facial recognition – so new photos may be needed more frequently to maintain accuracy and security.

Police surveillance and facial recognition: Why data privacy is imperative for communities of color (Nicol Turner Lee and Caitlin Chin-Rothmann, Brookings Institution, 2022):

In this paper, we present the case for stronger federal privacy protections with proscriptive guardrails for the public and private sectors to mitigate the high risks that are associated with the development and procurement of surveillance technologies. We also discuss the role of federal agencies in addressing the purposes and uses of facial recognition and other monitoring tools under their jurisdiction, as well as increased training for state and local law enforcement agencies to prevent the unfair or inaccurate profiling of people of color. We conclude the paper with a series of proposals that lean either toward clear restrictions on the use of surveillance technologies in certain contexts, or greater accountability and oversight mechanisms, including audits, policy interventions, and more inclusive technical designs.

<u>Clearview AI scraped 30 billion images from Facebook and other social media sites and gave</u> <u>them to cops: it puts everyone into a 'perpetual police line-up'</u> (Katherine Tangalakis-Lippert, Business Insider, 2023): A controversial facial recognition database, used by police departments across the nation, was built in part with 30 billion photos the company scraped from Facebook and other social media users without their permission, the company's CEO recently admitted, creating what critics called a "perpetual police line-up," even for people who haven't done anything wrong.

#### What is facial recognition and how does it work? (Clare Stouffer, Norton, 2023):

#### Facial recognition pros

As a relatively new technology, we're still understanding the pros and cons of facial recognition for everyday people. Here are some of the main advantages:

- Fly safer: Airports use facial recognition to identify criminals or potential threats.
- Identify criminals: Facial recognition can identify suspects from photos or videos.
- **Find missing persons:** Missing persons have been identified using facial recognition technology.
- **Keep your phone secure:** Many phones today use facial recognition to unlock or verify your identity before purchasing.

#### Facial recognition cons

It's important to understand the limits of facial recognition AI. For example, Randal Reid was arrested and jailed for a week in 2022 after being falsely identified by facial recognition technology. He had never even been to Louisiana, where the crime occurred.

Here are some of the top disadvantages of facial recognition:

- **Mistaken identity:** Relying on facial recognition alone can lead to falsely identifying criminal suspects.
- **Inaccuracies with older adults:** Facial recognition becomes less accurate when people age.
- **Racial and gender bias:** Studies have shown that facial recognition is less effective in identifying people of color and women.
- It can be tricked: Wearing a mask, sunglasses, or even certain makeup can make facial recognition less accurate.
- It can violate privacy: Many critics worry that facial recognition is one more erosion of personal privacy. With more accurate algorithms, it's becoming more of a risk that someone can take a photo of you in public and use AI to find more information.

- It poses security risks: Your facial data can be collected and stored, often without your permission. Hackers could access and steal that data.
- There are potential ownership issues: You may have given up your right to ownership over images of your face when you agreed to social media privacy policies.

Face Recognition: Privacy Concerns and Social Benefits (Jacob Gannon, Security Informed, 2024):

What are the facts about face recognition? Which side is right? Well, there is no definitive answer because, as with all powerful tools, it all depends on who uses it. Face recognition can, in fact, be used in an immoral or controversial manner. But, it can also be immensely beneficial in providing a safe and secure atmosphere for those in its presence.

Exploring the Impact of AI on Facial Recognition Technology (Fredrik Filipsson, Redress Compliance, 2024):

#### The Potential Impact on Society

As AI-powered facial recognition technology advances, its impact on society is twofold. On the one hand, it promises enhanced security, convenience, and personalization, potentially making our social interactions and daily transactions smoother and more secure.

On the other hand, the widespread adoption of facial recognition raises profound questions about privacy and the nature of public spaces.

It necessitates a societal dialogue on the values we wish to uphold and the kind of technological future we envision.

The key to navigating these future directions lies in striking a balance between leveraging Al's benefits in facial recognition and safeguarding individual rights and freedoms.

Ethical guidelines, transparent practices, and inclusive policies will be essential in shaping a future where facial recognition technology serves the greater good, enhancing our lives while respecting our privacy and dignity.

# Voting, Disinformation, and Voter Profiling

How AI Puts Elections at Risk — And the Needed Safeguards (Mekela Panditharatne & Noah Giansiracusa, Brennan Center for Justice, 2023):

Al advances have prompted an abundance of generalized concerns from the public and policymakers, but the impact of Al on the field of elections has received relatively little in-depth scrutiny given the outsize risk. This piece focuses on disinformation risks in 2024. Forthcoming Brennan Center analyses will examine additional areas of risk, including voter suppression, election security, and the use of Al in administering elections.

[...]

Al has the potential to dramatically change elections and threaten democracy. A whole-of-society response is needed.

<u>Preparing to Fight Al-Backed Voter Suppression</u> (Mekela Panditharatne, Brennan Center for Justice, 2024):

Generative AI introduces the possibility of more sophisticated methods of deception, capable of being deployed more cheaply and swiftly on a wider scale. Al's persuasive potential may increase over time as current technological limitations are quickly surpassed and different forms of AI are coalesced in new ways. Some kinds of AI systems will allow election deniers and other discontents to submit mass private challenges to voters' registration statuses more expediently — possibly with even less transparency and with a novel patina of faux legitimacy.

While it remains unclear how much AI will change the face of vote suppression in the 2024 general election, new developments in AI use and capabilities lend fresh urgency to long-standing efforts to abate attempts to subvert elections. Those developments necessitate strong new policy interventions to minimize the dangers on democracy's horizon.

<u>The impact of generative AI in a global election year</u> (Valerie Wirtschafter, Brookings Institution, 2024):

[...] instances of manipulated or wholly generated content have surfaced, posing a threat to democratic discourse and electoral integrity. Addressing this challenge requires a multifaceted response.

Interventions ranging from legislative measures targeting election-specific deepfakes to voter education initiatives are imperative. Tech companies should also play a central role, including through the implementation of imperfect technical solutions to identify the origins of generated media. While these interventions may not eliminate the challenges posed by generative AI, they represent progress toward managing a complex issue during a critical election year.

<u>Election disinformation takes a big leap with AI being used to deceive worldwide</u> (Ali Swenson and Kelvin Chan, AP, 2024):

Artificial intelligence is supercharging the threat of election disinformation worldwide, making it easy for anyone with a smartphone and a devious imagination to create fake – but convincing – content aimed at fooling voters.

It marks a quantum leap from a few years ago, when creating phony photos, videos or audio clips required teams of people with time, technical skill and money. Now, using free and low-cost generative artificial intelligence services from companies like Google and OpenAI, anyone can create high-quality "deepfakes" with just a simple text prompt.

[...]

"A world in which everything is suspect — and so everyone gets to choose what they believe — is also a world that's really challenging for a flourishing democracy," said Lisa Reppell, a researcher at the International Foundation for Electoral Systems in Arlington, Virginia.

# **Brain-Computer Interfaces**

<u>Understanding the Ethical Issues of Brain-Computer Interfaces (BCIs): A Blessing or the</u> <u>Beginning of a Dystopian Future?</u> (Livanis et al., Cureus, 2024):

In recent years, scientific discoveries in the field of neuroscience combined with developments in the field of artificial intelligence have led to the development of a range of neurotechnologies. Advances in neuroimaging systems, neurostimulators, and

brain-computer interfaces (BCIs) are leading to new ways of enhancing, controlling, and "reading" the brain. In addition, although BCIs were developed and used primarily in the medical field, they are now increasingly applied in other fields (entertainment, marketing, education, defense industry). [...] Among the ethical issues that emerged from the thematic data analysis of the reviewed studies included questions revolving around human dignity, personhood and autonomy, user safety, stigma and discrimination, privacy and security, responsibility, research ethics, and social justice (including access to this technology).

[...]

Not surprisingly, developments raise concerns about how far this technology can go and what impact it will have on social, political, and religious norms. In that regard, discussions regarding the commercialization of brain data and the potential for manipulation of the human brain are likely to increase in the coming years. [...] It is therefore imperative that the scientific, political, and business communities cooperate in order to indicate, through bioethics, the safeguards that will ensure the ethical design of the development, use, and management of BCIs. It is everyone's ethical duty to ensure that technological development benefits humanity.

# Medicine

# **Outcome Research**

Large Language Models for Outcomes Research: A Targeted Review (Dolin, O, et al, ISPOR, 2024):

Interest in using large language models (LLMs) for outcomes research has increased in recent years; however, feasibility of LLM integration within research workflows remains unclear. [...]

LLM performance and reliability concerns remain a roadblock to implementation. [...] Usage guidance and standardized validation approaches are currently lacking; both would facilitate LLM use for research.

# Insurance

The role of AI in insurance is multifaceted. Within the insurance industry itself, AI can impact not only customer interactions and data processing, but also the methodology of risk and pricing analysis. In addition, insurance companies need to assess the risks of using AI in general situations in order to offer coverage.

## Use of AI within the Insurance Industry

Regarding the use of AI within the insurance industry itself, some risk factors, though potentially relevant to outcomes, may not be legal for public policy reasons (similar to, say, <u>hiring</u>). Since a combination of otherwise legal inputs may be highly correlated to illegal inputs (e.g. income and address may correlate to race), there is a trade-off between the increased accuracy of AI with the decreased explainability. According to one industry report, the use of AI cuts to the heart of insurance models, in its ability to discern a multitude of risk causes and correlations, across various groups – down to the individual level:

#### <u>REGULATION OF ARTIFICIAL INTELLIGENCE IN INSURANCE: Balancing consumer</u> <u>protection and innovation</u> (Noordhoek, The Geneva Association, 2023)

In a recent report, the Dutch Financial Market Authority (AFM) concluded that, while some groups of customers might face higher premiums or become uninsurable, individual risk assessments are generally considered fair and offer opportunities for risk reduction and mitigation. They also determined that governments have a role in supporting those customers who become uninsurable due to individualised pricing. This highlights that AI lays bare issues that would otherwise not be visible and potentially merit a societal discussion.

#### [...]

Finally, yet importantly, though the growing use of AI means that correlation increasingly substitutes causality, existing insurance regulatory practices in pricing and conduct remain rooted in the latter. This limits the use of rating factors to only those that demonstrably influence the risk. This by itself limits the extent to which AI can be used by insurers and supports the main argument of this report – that crosssectoral regulation that covers the use of AI in insurance is less effective than insurance-specific regulation.

## Al Risk Analysis

In regard to the assessment of the use of AI more generally, the degree of risk can be categorized by the sophistication of the AI used, the maturity of the integration, and the type, scale, and frequency of the potential harm. One insurance study has classified the risks associated with the use of AI into six major categories:<sup>4</sup>

- 1. Data bias or lack of fairness: unintended discrimination against a protected group
- 2. Cyber: system vulnerabilities or malignant use
- 3. Algorithmic and performance: failure to meet metrics requirements
- 4. Lack of ethics, accountability, and transparency: failure to adhere to ethics or accountability requirements, possibly obscured by lack of transparency
- 5. **Intellectual property (IP):** use of third-party IP in training data or unintended infringement
- 6. Privacy: unauthorized use or exposure of personal data in training data or output

In terms of AIITF general working group interdisciplinary analysis and guidance, one stakeholder to consider is the insurance industry.<sup>5</sup>

Insurers can also play an important role in reducing risks associated with ethics, accountability and transparency [...]: assessments of AI, machine learning and analytics models for trustworthiness, robustness, accuracy, transparency, ethical use and governance of data and AI. [They] can find application in multiple industries, including manufacturing (to optimise operations, ensuring product quality, worker safety, and mitigating disruptions in the manufacturing process) and mobility (potential risks associated with self-driving vehicles).

A last word on insurance and AI: when we look at projections for AI growth, mentioned in the introduction as "between USD 2.6 trillion and USD 4.4 trillion" per annum, then AI will become ubiquitous across industry lines. This will bring AI into traditional insurance lines, which if it is not specifically included or excluded, could exacerbate losses. This has been described as 'silent AI risk' and has potentially serious consequences for accumulation risks in insurance portfolios. [...] AI may be revolutionary in many ways, but it will sometimes also be fallible. It is for insurers to consider to sustainably provide and create resilience for this emerging technology.

<sup>&</sup>lt;sup>4</sup> <u>Tech-tonic Shifts: How AI could change industry risk landscapes</u> (Woodward, et al, 2024, Swiss Re Institute, 2024). This article further references the "AI Litigation Database", The George Washington University, (2024).

<sup>&</sup>lt;sup>5</sup> Ibid., Conclusion: AI, risk and implications for insurers.

# **Employment & Labor**

## International Generally

# OECD Employment Outlook 2023: ARTIFICIAL INTELLIGENCE AND THE LABOUR MARKET (OECD, 2023):

Al appears to be different from previous digital technological changes in several ways: i) it significantly expands the range of tasks that can be automated beyond just routine, non-cognitive tasks; ii) Al is a general-purpose technology, meaning that nearly every sector and occupation will be affected; and iii) the speed of development is unprecedented.

[...]

Collective bargaining and social dialogue also have an important role to play in supporting workers and businesses in the AI transition. AI adoption tends to result in better outcomes for workers when their representatives are consulted on the matter. Yet, the specific characteristics of AI and the way it is implemented, such as its rapid speed of diffusion, its ability to learn and the greater power imbalance it can create, put further pressure on labour relations. While AI technologies have the potential to assist social partners to pursue their goals and strategies, the lack of AI-related expertise among social partners is a major challenge.

## **Entertainment Industry**

AI-Generated Music

<u>I Told You This Was Going To Happen: The Musician's Biggest Threat</u> (Rick Beato, YouTube, 2024):

Al-generated music: even though a well-trained ear can currently distinguish between human and Al generated music, "in six months I probably won't be able to tell the difference." Why buy music when you can have Al generate exactly the kind of music you want to hear?

[7:19] "What happens to the musicians? Who makes the money? Does the studio make it? Is anyone getting paid for the stuff it's trained on? How do you even know what it's

trained on? I went and testified in front of Congress for one of these, you know, information things. It was in the seventh of nine things, and they've done nothing about it. They don't even know what to do about it. I don't know what there can be done about it but it's going to replace people."

World's biggest music labels sue over Al copyright (Natalie Sherman, BBC News, 2024):

The world's biggest record labels are suing two artificial intelligence (AI) start-ups over alleged copyright violation in a potentially landmark case.

Firms including Sony Music, Universal Music Group and Warner Records say Suno and Udio have committed copyright infringement on an 'almost unimaginable scale'.

They claim the pair's software steals music to 'spit out' similar work and ask for compensation of 150,000 (£118,200) per work.

Motion Picture and Television Production

<u>Will A.I. Upend White-Collar Work? Consider the Hollywood Editor.</u> (Noam Scheiber, NYT, 2024):

In a dozen interviews with editors and other Hollywood craftspeople, almost all worried that A.I. had either begun displacing them or could soon do so. [...]

At a town-hall meeting to discuss the contract — which covers not just editors but also thousands of makeup artists, prop makers, set designers, lighting technicians and camera operators — the union's president advised members to make the best of it. [...]

But to Mr. Moore and his fellow Cassandras, the failure to secure stronger A.I. protections bodes poorly not only for them but for workers across the country. "If a 70,000-member union like IATSE can't protect workers, what does it mean for everybody else?" he said, referring to the number of craftspeople covered under two major contracts. "For society going forward?"

Hiring, Recruiting, and Performance Evaluation

Al in Hiring and Evaluating Workers: What Americans Think (Tyson, et al, Pew Research, 2023):

A new Pew Research Center survey finds crosscurrents in the public's opinions as they look at the possible uses of AI in workplaces. Americans are wary and sometimes worried. For instance, they oppose AI use in making final hiring decisions by a 71%-7% margin, and a majority also opposes AI analysis being used in making firing decisions. Pluralities oppose AI use in reviewing job applications and in determining whether a worker should be promoted. Beyond that, majorities do not support the idea of AI systems being used to track workers' movements while they are at work or keeping track of when office workers are at their desks.

Yet there are instances where people think AI in workplaces would do better than humans. For example, 47% think AI would do better than humans at evaluating all job applicants in the same way, while a much smaller share – 15% – believe AI would be worse than humans in doing that. And among those who believe that bias along racial and ethnic lines is a problem in performance evaluations generally, more believe that greater use of AI by employers would make things better rather than worse in the hiring and worker-evaluation process.

How Artificial Intelligence Might Prevent You From Getting Hired (Akselrod & Venzke, ACLU, 2023):

Moreover, the correlations that an AI tool uncovers may not actually have a causal connection with being a successful employee, may not themselves be job-related, and may be proxies for protected characteristics. For example, <u>one resume screening tool</u> identified being named Jared and playing high school lacrosse as correlated with being a successful employee. Likewise, the amorphous personality traits that many AI tools are designed to measure — characteristics such as positivity, ability to handle pressure, or extroversion — are often not necessary for the job, may reflect standards and norms that are culturally specific, or <u>can screen out candidates with disabilities</u> such as autism, depression, or attention deficit disorder.

# Education

K-12 In-Class Use of AI

What Teachers Told Me About A.I. in School (Jessica Gross, NYTimes, 2024):

Leila Wheless, a North Carolina teacher who has been an educator since 1991, tried to keep "an open heart" about using artificial intelligence in her middle school English and language arts classroom. She reviewed the guidance of her state's generative A.I. "<u>recommendations and considerations</u>" for public schools. But the results of her students' A.I. use were dispiriting. [...]

The biggest issue isn't just that students might use it to cheat — students have been trying to cheat forever — or that they might wind up with absurdly wrong answers, like confusing Moses with Mr. Clean. The thornier problem is that when students rely on a generative A.I. tool like ChatGPT to outsource brainstorming and writing, they may be losing the ability to think critically and to overcome frustration with tasks that don't come easily to them.

## **College Admissions**

An interesting aspect of College Admissions is that both sides are using AI – applicants and admissions offices. As in the other use of AI in K-12 classrooms, or potentially with job application resumes, we risk a flattening of creativity. In addition, there is an increase in AI evaluating the output of AI, potentially leading to some form of "collapse" or "self-reinforcement loop". As highlighted by Gemini, the risks include degradation of quality, reinforcement of biases, and loss of originality.

#### Is AI Affecting College Admissions? (Claybourn, US News, 2023):

Fifty percent of higher education admissions offices are using AI, according to a September 2023 <u>survey</u> by Intelligent, an online magazine focused on higher education. That number is expected to rise to more than 80% in 2024, according to the survey, which polled nearly 400 education professionals in both K-12 and higher education.

#### Incorporating AI into the Admissions Process (Boyd, Volt, 2024):

"As AI's role in admissions evolves, its applications will likely expand beyond chatbots and predictive analytics to include more sophisticated analysis of essays and personal statements," Johnson said.

#### [...]

Johnson also stressed that, although AI can streamline the admissions process, it lacks the human capacity to appreciate the depth of personal essays fully, potentially overlooking the unique, intangible qualities that make applicants stand out. "This limitation underscores the importance of maintaining human oversight in the review process," he said. "AI is only as just as the equitable decisions that inform its design."

<u>1 in 3 College Applicants Used AI for Essay Help. Did They Cheat?</u> (Klein, Education Week, 2024):

About half of those students—or roughly one in six students overall—used AI the way Makena did, to brainstorm essay topics or polish their spelling and grammar. And about 6 percent of students overall—including some of Makena's classmates, she said—relied on AI to write the final drafts of their essays instead of doing most of the writing themselves.

Meanwhile, nearly a quarter of students admitted to Harvard University's class of 2027 paid a private admissions consultant for help with their applications.

The use of outside help, in other words, is rampant in college admissions, opening up a host of questions about ethics, norms, and equal opportunity.

Top among them: Which—if any—of these students cheated in the admissions process?

For now, the answer is murky.

# Courts

Analysis of AI Creativity within IP Law

Current court cases in the U.S. and elsewhere have debated whether or not AI could be listed as an author worthy of copyright, or as the inventor of a patent. Guidance necessitates an analysis of human vs. AI "creativity", what the law requires and why, and how this might change in the near future. The work product from a group at Oxford addressing this problem is included in <u>Appendix B</u>.

Use of AI by Judges

Snell v. UNITED SPECIALTY INSURANCE COMPANY, Court of Appeals, 11th Circuit 2024

NEWSOM, Circuit Judge, concurring:

"Here's the proposal, which I suspect many will reflexively condemn as heresy, but which I promise to unpack if given the chance: Those, like me, who believe that 'ordinary meaning' is *the* foundational rule for the evaluation of legal texts should consider —*consider*— whether and how AI-powered large language models like OpenAI's ChatGPT, Google's Gemini, and Anthropic's Claude might —*might*— inform the interpretive analysis." [emphasis in original]

# Legislation

Future Politics (Jamie Susskind, Oxford University Press, 2019):

#### AI Democracy

What role will artificial intelligence come to play in governing human affairs? What role *should* it play? These questions have been floating around since the earliest computing machines. In the twentieth century, reactions to the first question tended to involve dark premonitions of humankind languishing under the boot of its robotic overlords. Reflection on the second question has been somewhat limited and deserves more careful thought. [p. 250]

Sometimes the cause of justice is better served by an open mind than a clenched fist. It can be studied calmly, even clinically. It can even be studied *algorithmically*." [focus on distribution, recognition, employment, and property] [p. 258]

# Appendix D: Legal Pro Bono Compliance

According to Gemini, the ability of attorney work to qualify as pro bono under state guidelines typically includes the factors listed below. Among them, and most relevant to this proposal, is that legal work for the public interest, or to protect civil rights and liberties, generally qualifies as pro bono work. The AIITF is a non-profit organization whose mission includes safeguarding the public interest and various human rights and liberties as AI is integrated into various components of society. It is the intention of AIITF that the development of guidelines within the working group structure is in the public interest. Furthermore, like the IETF, the goal is to present guidelines that are party-neutral, that explore options and their consequences, and help promote informed decision-making. The AIITF seeks to structure some or all of the working groups to be compliant with pro bono legal work requirements in those cases where legal issues are appropriately a component of the discussions, as exemplified in <u>Appendix B</u>.

#### From Gemini:

The specific criteria for attorney volunteer work to qualify as pro bono for state bar requirements can vary slightly depending on the jurisdiction. However, some general principles typically apply across most state bars:

- 1. Free Legal Services: The legal services provided must be free of charge or at a significantly reduced fee for clients who cannot afford to pay the usual rate.
- Eligible Clients: The services should primarily benefit individuals or organizations that are considered low-income, disadvantaged, or otherwise unable to access legal representation.
- 3. Charitable or Public Interest: The volunteer work should be undertaken for a charitable purpose, in the public interest, or to advance a particular legal cause. This could include providing legal aid to non-profit organizations, representing underserved populations in civil matters, or assisting in legal clinics.
- 4. Direct Legal Services: In most cases, pro bono work must involve the provision of direct legal services, such as legal advice, representation in court, drafting legal documents, or engaging in legal research.
- 5. Reporting Requirements: Some state bars require attorneys to report their pro bono hours or activities to maintain compliance with professional conduct rules.

#### Specific Requirements in California:

The State Bar of California has a Pro Bono Practice Program that encourages attorneys to volunteer their legal services. For pro bono work to qualify under this program, it must meet the following criteria:

• The legal services are provided free of charge or at a substantially reduced fee to clients who cannot afford to pay.

- The services benefit persons of limited means or charitable, religious, civic, community, governmental, and educational organizations in matters designed primarily to address the legal needs of persons of limited means.
- The services are provided to organizations seeking to secure or protect civil rights, civil liberties, or public rights.
- The services are provided in furtherance of the provision of legal services to the poor.

You can find more information about the California State Bar's Pro Bono Practice Program on their website:

https://www.calbar.ca.gov/Access-to-Justice/Pro-Bono/FAQ Additionally, it's important to consult the specific rules and guidelines of your state bar association to ensure that your volunteer work meets the necessary criteria for pro bono credit.