# A DCI Deliberation Guide

# Artificial Intelligence:

*How Should We Relate to it as Individuals, Communities, and Policymakers?*

## Format for Deliberation

**Before the Deliberation**

I. Read this Deliberation Guide (required)

II. Watch these two videos to learn about the artificial intelligence systems that are the focus of this guide (optional but recommended):

a. What are Generative AI Models?

b. How Large Language Models Work

**During the Deliberation**

I. Setting Expectations (5 min.)

II. Getting to Know Each Other (5 min.)

III. Concerns & Hopes about AI (10 min.)

IV. The Transformative Impact of AI (15 min.)

V. Navigating AI's Impact Together (15 min.)

VI. Reflections (10 min.)

## Background

> "To make a difference in how we deploy AI calls for a deeper, more prolonged engagement, one that arouses a society's ethical and political intelligence. We need to bring AI back onto the agenda of deliberative democracy. That project will take more than six months, but it will be wholly worth it."
>
> – Sheila Jasanoff, Pforzheimer Professor of Science and Technology Studies at Harvard University

### Section 1: Introduction

In November of 2022, the now-famous tech startup OpenAI released ChatGPT to the public. ChatGPT was a *chatbot*—a computer model with which users can interact in a conversational way. One surprising thing about ChatGPT is just how successful it was at interacting with us *successfully*—the model could provide not just grammatical but appropriate conversational responses via text. It could respond to text-based inputs not only grammatically, but also appropriately. It seemed to be able to follow the rules of conversation we can expect humans to follow. It could generate novel responses in conversation just like a person would. It even had some capacity to seemingly understand jokes and write poetry.[1] It could pass the Uniform Bar Examination, more commonly known as "the bar," (the professional examination required before a person is allowed to practice law in any state in the U.S.) in nearly the 90th percentile.[2]

Although ChatGPT's release generated public excitement about its novelty and successes, skeptics warned about a dark side. The model was criticized for generating inaccurate claims, especially because predecessors to cutting edge models like ChatGPT had already been used to spread misinformation online.[3] It was also criticized for having content that was socially or politically biased.[4] The data labeling industry which is required for training large generative models came under fire for exploitative and harmful labor practices.[5] Another concern is that the model was (and remains) *uninterpretable*. This means that its inner workings in principle cannot be discovered by humans. It is therefore not possible for us to "trace back" how the model generated an output based on a certain input and its initial settings. Therefore, skeptics warn that we should not trust such models in high stakes situations.[6]

But the history of skepticism about artificial intelligence (also known as *AI*) dates back to at least the mid-20th century. Alan Turing, the British mathematician and logician responsible for the invention of digital computing, predicted that "intelligent machines" would replace human workers, as long as the work they did could be specified by a set of rules (and hence, could be captured in an algorithm used by a computer). This Deliberation Guide provides introductory background that can help the public weigh the promise of AI against the risks associated with it. To that end, this section includes a brief introduction to how AI systems work.

[1] Bonos, Lisa. "ChatGPT Might Kill Us All…With Dad Jokes." Washington Post. 2023.
Cushman, Jack. "ChatGPT: Poems and Secrets." Library Innovation Lab. 2022.
Mair, Victoria. "GPT Has a Sense of Humor (Sort Of)." Language Log. 2023.

[2] Kimmel, Lara. "ChatGPT Passed the Uniform Bar Examination: Is Artificial Intelligence Smart Enough to be a Lawyer?" International and Comparative Law Review. 2023.
It is worth noting that other chatbots similar to ChatGPT existed at the time; however, ChatGPT is the example we use here because of its widespread use. According to Reuters in February 2023, it was the fastest growing computer application in history. See: Hu, Krystal. "ChatGPT Sets Record for Fastest Growing User Base." Reuters. 2023.

[3] Hsu, Tiffany and Stuart A. Thompson. "Disinformation Researchers Raise Alarms About A.I. Chatbots." New York Times. 2023.

[4] Motoki, Fabio, Valdemar Pinho Neto and Victor Rodrigues. "More Humand than Human: Measuring ChatGPT Political Bias." Public Choice. 2023.

[5] Williams, Adrienne, Milagros Miceli and Timnit Gebru. "The Exploited Labor Behind Artificial Intelligence." Noema. 2022.

[6] Rudin, Cynthia. "Stop Explaining Black Box Models for High Stakes Decisions and Use Interpretable Models Instead." Nature Machine Intelligence. 2019.

So, how do these systems actually work? ChatGPT is part of a class of computational models called *large language models* (or *LLMs*). LLMs belong in the class of *foundation models*, a type of program that can generate novel outputs based on its computer architecture and body of content on which the model was trained.[7] These models are trained on extremely large amounts of information—in the case of ChatGPT, about 300,000 billion words.[8] But in what sense are the outputs of foundation models like LLMs really novel?

The answer is that the output generated by LLMs is new in the sense that it *does not appear in the training data*, or the body of content that was used to create the model's ability to interact using test, images, voice, and so on. Foundation models do not merely "regurgitate" the content that was used to create them. Rather, they generate genuinely creative results, since they create new content that is informed by but importantly different from the content they were trained on. Since these systems actually generate new content based on their training data, they are classified as *generative artificial intelligence*, or *genAI* for short.

There is another sense in which the outputs of these systems are novel. Once the set of training data gets large enough, foundation models exhibit fundamentally unpredicted and unpredictable capabilities. The kinds of things that these models can do fundamentally changes once they learn enough from the training data. But we cannot predict what these emergent capabilities will be, and so the models are not only *uninterpretable* but also *unpredictable*, in the sense that we cannot know, based on our understanding of the models' workings and the training data set, what the models will be able to do.

At this point, any of the following questions might strike you as significant:

- What good can this technology bring about, and what harm can it do?
- What will be the social, political, and economic impact of this technology?
- To what extent could these models transform the world we live in, and what are domains in which they will likely have the greatest impact?
- What should individuals, communities, and policymakers do about this technology, if anything?

This guide will touch on each of the above questions. Because this technology is advancing so rapidly, materials about it can become dated quickly. While this deliberation guide will address some current issues in the global AI landscape, it is primarily targeted at more longstanding questions about this technology. In this sense, this guide may continue to be useful in the future to those interested in the public impact of AI and the governance of this technology.

[7] *Training* is the process by which foundation models gain their capacities. See: Lutkevich, Ben. "Foundation Models Explained: Everything You Need to Know." TechTarget. 2023.

[8] Nolan, Beatrice. "Google's Researchers Say They Got OpenAI's ChatGPT to Reveal Some of Its Training Data With Just One Word." Business Insider. 2023.

## Section 2: Risks to Human Wellbeing and other Drawbacks

Popular culture is rife with examples of artificially intelligent systems or robots intentionally causing destruction to human welfare—just think of films like *2001: A Space Odyssey*, *Terminator*, *The Matrix,* or the TV show *Battlestar Galactica*. But the potential threat AI poses to human welfare and our social fabric is neither mere science fiction nor is the only threat an agent bent on the destruction of humans. In this section, we explore several potential threats that genAI might pose to us, as well as the sources of those risks.

### 2.1 The Alignment Problem

GenAI systems create novel outputs based on their training data sets. This means that whatever outputs gets created—whether it is text, sound, visual information, or all of the above—it did not appear in the data that was used to create the model's abilities. This idea dates back to the work of Alan Turing, who argued that intelligent machines will surpass humans in intelligence, learning ability, and control.[9] In 1960, the computer scientist and philosopher Norbert Wiener argued that artificially intelligent machines could develop unforeseen abilities and that our control of these machines would eventually be "nullified."[10]

These ideas give rise to what has become known as *the alignment problem*. The problem is that humans may not be able to prevent AI systems, including genAI systems, from generating outputs that run counter to human values, well-being, and safety. If this in fact occurs, these systems will become *misaligned* with the interests of humanity. This problem has been articulated by contemporary high-profile figures like Elon Musk (who helped to start OpenAI) and scholars like Stephen Hawking and Nick Bostrom.[11] They are concerned that AI could redesign itself at a rate that we cannot control, that AI systems could usurp resources to pursue their goals and leave humans with insufficient resources for survival, and that AI systems will generally pursue goals that harm human interests.

Many AI researchers are working on finding practical solutions that will make AI systems more aligned with our values and ensure that we can test how aligned they really are.[12] However, it is not yet clear that a technical solution can be found, and therefore we may never know how safe these systems are. One reason why is that AI capabilities can emerge at faster rates than our alignment research can keep up with.[13] But another and much deeper reason is that, when we train AI systems that are highly capable, uninterpretable, and unpredictable—systems like the

[9] Turing, A.M. "Intelligent Machinery, A Heretical Theory." 1951.

[10] "Norbert Wiener." Wikipedia. 2024.
Wiener, Norbert. "Some Moral and Technical Consequences of Automation." Science. 1960.

[11] Bostrom, Nick. *Superintelligence: Paths, Dangers, and Strategies*. 2014.
Cellan-Jones, Rory. "Stephen Hawking Warns Artificial Intelligence Could End Mankind."
Gohd, Chelsea. "Elon Musk Claims We Only Have a 10 Percent Chance of Making AI Safe." Futurism. 2017.

[12] See the work of organizations like NIST and METR.

[13] Leike, Jan. "What Could a Solution to the Alignment Problem Look Like?" Musings on the Alignment Problem. 2022.

current GPT-4 from OpenAI, or Alphabet's Gemini, or Meta's LLaMa 2, we cannot know exactly what dispositions and capabilities we are training into them.

### 2.2 Transparency, Trust, and Bias

In addition to the alignment problem, which may put human welfare at risk in a post-genAI world (and according to some experts, at great risk), there are other concerns that may be justified about this technology.  One of these has to do with the fact that, since these models are black boxes, we cannot understand the basis on which an output was generated.

This is not a problem in certain contexts. Asking a chatbot to generate a grocery list based on one's nutritional goals is not a task for which we need to have transparency. But some scholars argue that we should not use black box models for all purposes in which we might want to rely on machine learning.[14] In the context of high-stakes decisions, these critics argue that we should use AI models whose functioning we are able to interpret. This allows us to understand the *basis* on which a particular output was generated.

A related concern about AI when it is applied in practical settings is bias. Consider the case of prison sentencing. AI systems are frequently used to determine prison sentences for criminal convictions. In the UK, a system called Offended Assessment System (Oasys) has been used for about twenty years. It is a black box system and human beings therefore cannot precisely glean the basis on which the prison sentences were determined. It is also not clear whether the sentences generated by Oasys lead to more or fewer repeat offenses. If systems like this are biased in ways that are not desirable or just and these biases are not detectable by humans, then our criminal justice decisions may be unjustified.[15] To understand whether a sentence is justified, critics argue that we must use transparent models rather than black box models.

Moreover, if the public cannot gain access to the basis on which a decision about a prison sentence was made—as we can do when we get access to a court decision written up by a judge—then the public may lose trust in the criminal justice system. Yet trust in the court system is a key part of our public's trust in democratic institutions generally.[16] A supporter of AI use might respond, however, that trust in our court system is already degraded, and we actually

[14] Rudin, Cynthia. "Stop Explaining Black Box Models for High Stakes Decisions and Use Interpretable Models Instead." Nature Machine Intelligence. 2019.
Rudin, Cynthia and Joanna Radin. "Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition." Harvard Data Science Review. 2019.

[15] Hamilton, Melissa and Pamela Ugwudike. "A 'Black Box' AI System has been Influencing Criminal Justice Decisions for Over Two Decades—It's Time to Open It Up." The Conversation. 2023.

[16] "Public Trust and Confidence." National Association for Court Management. 2024.
"Issue 2: Preserving Public Trust, Confidence, and Understanding." United States Courts. Accessed 2024.
Sherman, Lawrence W. "Trust and Confidence in Criminal Justice." Office of Justice Programs. 2001.

don’t know what goes into a judge’s decision-making process.[17] An AI system might ultimately be less biased than one that only relies only on humans.[18]

### 2.3 Other Risks & Drawbacks

There is a wide variety of risks posed by AI and genAI in particular. AI has the potential to massively amplify the spread of manipulated or falsified information (including deepfakes) around the world, which the World Economic Forum ranks as the most critical short-term risk to democracy and social cohesion around the globally.[19] Severe job loss, misuse of AI in military contexts, and increased threats to business and national security as a result of cyberattacks also rank highly among the most likely risks.[20] The concentration of technological power is also a global concern, because the financial and intellectual capital required to create the most powerful AI systems is concentrated in a relatively few states.[21]

## Section 3: The Potential for Good

It’s safe to say that AI systems, especially generative AI systems, come with some significant risks whose likelihood we cannot fully understand yet. But these systems also have enormous potential for good. And not only good with respect to human welfare and decreased suffering, but also good understood more broadly. This section surveys a number of AI applications that have transformative potential to improve and increase wellbeing, understanding, convenience, economic growth, access to education, and other dimensions of human life.

One challenge for harnessing AI’s potential for good is that it requires coordinated human effort on a large scale.[22] But AI has already been used to decrease online abuse of women, predict conflict zones, and increase our knowledge of climate informatics so that we may combat climate and change.[23]

### 3.1 Discoveries in the Humanities and Sciences

AI systems have the potential to uncover heretofore undiscoverable truths. For example, texts that have been carbonized and buried in the volcanic eruption that destroyed Pompeii have been able to be read for the first time since AD 79, after an AI model gained the capacity to

[17] Rachlinski, Jeffrey J., Andrew J. Wistrich and Chris Guthrie. “Can Judges Make Reliable Numeric Judgments? Distorted Damages and Skewed Sentences.” Indiana Law Journal. 2015.
[18] Berman, Robby. “Algorithms Identify Repeat Offenders Better than Judges.” Big Think. 2020.
[19] Torkington, Simon. “These are the 3 Biggest Emerging Risks the World Is Facing.” World Economic Forum. 2024.
[20] Ibid.
[21] Ibid.
Rotman, David. “How to Solve AI’s Inequality Problem.” MIT Technology Review. 2022.
[22] Tomašev, Nenad, Julien Cornebise and Frank Hutter, et.al. “AI for Social Good: Unlocking the Opportunity for Positive Impact.” Nature Communications. 2020.
[23] Ibid.

detect textural differentiations in the carbonized scrolls.[24] Discoveries like this may help scholars in the humanities make sense of intellectual and artistic history in a new way.
AI systems have also been shown to solve heretofore unsolved mathematical problems.[25] As a result of some of these solutions, some AI systems can drastically outperform preexisting solutions for debugging computer programs and optimizing code performance.

The AI model AlphaFold developed by DeepMind can get a much more accurate understanding we have of proteins, which are essential to all living things. AlphaFold helps accurately predict protein structure, and understanding protein structure accurately has been one of the most important and yet most difficult challenges in modern biology. This discovery is revolutionary in biology, and it could help us make significant and rapid progress toward understanding diseases and making new drug discoveries.[26] Though some critics argue that these advances have been overblown in the press and that there are still significant limitations to our understanding of protein folding, DeepMind's discovery will likely speed up other scientific discoveries.[27]

AI impacts on healthcare, especially in the realm of diagnostic testing, can be tremendously positive. In some settings, medical practitioners tend to agree 90-95% of the time with AI diagnostic tools.[28] In a study reported on by *Nature* but has not yet been peer-reviewed, a chatbot developed by Google provided more accurate diagnoses of respiratory and cardiovascular diseases than board-certified primary-care doctors – and was better at conversing with simulated patients.[29] AI systems can be used to detect conditions ranging from cancers, to Alzheimer's disease, to acute appendicitis.[30]

### 3.2 Economic Productivity

The potential economic benefits of AI systems are vast. In 2023, it was predicted that over 85 million skilled labor jobs could go unfilled because the skilled workforce was not large enough to meet demand. AI systems coupled with advancements in robotics could fill this need and fix existing issues in a variety of supply chains.[31] The idea is not that these systems would replace existing human workers but that they would fill a gap in the labor market where there are not

[24] Merchant, Jo. "AI Reads Text from Ancient Herculaneum Scrolls for the First Time." Nature. 2023.
[25] Sample, Ian. "AI Scientists Make 'Exciting' Discovery Using Chatbots to Solve Math Problems." The Gudardian. 2023.
Romera-Paredes, Bernadino, Mohammadamin Barekatain and Alexander Novikov, et. al. "Mathematical Discoveries from Program Search with Large Language Models." Nature. 2024.
[26] Callaway, Ewen. "'It Will Change Everything': DeepMind's AI Makes Gigantic Leap in Solving Protein Structures." Nature. 2020.
[27] Curry, Stephen. "No, DeepMind Has Not Solved Protein Folding." Reciprocal Space. 2020.
[28] Kennedy, Shania. "AI Achieves High Diagnostic Accuracy in Virtual Primary Care Setting." Health IT Analytics. 2023.
[29] Lenharo, Mariana. "Google AI Has Better Bedside Manner Than Human Doctors—and Makes Better Diagnoses." Nature. 2024.
[30] Umapathy Vidhya, Suba B. Rajinikanth and Rajkumar Samuel Raj, et al. "Perspective of Artificial Intelligence in Disease Diagnosis: A Review of Current and Future Endeavours in the Medical Field." Cureus. 2023.
[31] McKendrick, Joe. "We Can't Find Enough Skilled Workers: Can Automation Fill the Gaps?" Forbes. 2023.

enough humans to do so. Some predict that labor shortages will continue until 2040.[32] Demand for hands-on workers is predicted to grow, while AI is expected to decrease the demand for knowledge workers.[33]

AI has the potential to impact the education sector and increase access to education broadly.[34] It can also help improve job performance for workers who underperform at their jobs.[35] It can help workers who are less experienced increase their expertise comparatively quickly.[36] As a result, although AI systems may replace many jobs in the short term, they may also lead to increased productivity in the long term. Overall, AI may contribute $15 trillion to the global economy.[37]

### 3.3 Autonomous Vehicles and Increased Safety

One domain in which AI has been used for decades is in the development of autonomous vehicles, sometimes popularly called “self-driving cars.” Autonomous vehicles have the potential to be much safer than our current system because most serious crashes are caused by human error or poor decisions—an estimated 94% of serious crashes.[38] Proponents of autonomous vehicles argue that AI can therefore help prevent the deaths and injuries that result from driver errors. If autonomous vehicles can deliver on this promise, then we have strong reason to think that AI will have a positive impact on human welfare in this domain.

### 3.4 Other Benefits of AI

Because of its wide-ranging capacities, genAI can be applied to solve a vast variety of problems. It can be used as a tutor in an educational setting, personalized to the needs of the learner, and could teach difficult tasks step by step.[39] AI systems can not only be personalized learning companions, but engage with our emotional states, be playmates to young children, and companions for the elderly.[40] They may be able to make more accurate and timely military decisions than humans can, or be used to gain military intelligence, thereby decreasing certain risks in warfare.[41] These systems may increase national security more broadly as well, provided that it is implemented safely.[42] It may help us eliminate global poverty, dramatically reduce

[32] Conerly, Bill. “The Future of Work: AI, Remote Work and the Labor Shortage.” Forbes. 2024.
[33] Ibid.
[34] Saujani, Reshma. “We Don’t Have to Choose Between Ethical AI and Innovative AI.” Time. 2023.
[35] Ito, Aki. “AI Is the Great Equalizer.” Business Insider. 2023.
[36] Georgiva, Kristalina. “AI Will Transform the Global Economy. Let’s Make Sure It Benefits Humanity.” IMF Blog. 2024.
[37] “Artificial Intelligence: A *Real* Game Changer.” Bank of America. Accessed 2024.
[38] Lutkevich, Ben. “Self-Driving Car (Autonomous Car or Driverless Car).” TechTarget. 2023.
[39] Mollick, Ethan and Lilach Mollick. “Part 2: AI as Personal Tutor.” Harvard Business Publishing: Education. 2023.
[40] “Personal Robots.” MIT Media Lab: People. Accessed 2024.
Mole, Beth. “AI Companion Robot Helps Some Seniors Fight Loneliness, but Others Hate It.” Ars Technica. 2023.
[41] King, Anthony. “Artificial Intelligence and Urban Operations.” Journal of Strategic Security. 2023.
[42] Lee, Tony. “Where AI Can Improve National Security.” Government Technology Insider. 2023.

disease, and potentially free many humans from unengaging work.[43] AI can also increase the rate and quality of product development.[44]

## Section 4: AI Governance

To what extent should governments—either independently or collaboratively—regulate the development, release, and testing of generative AI systems? A wide range of AI governance policies have been proposed or implemented by countries around the world. This section will review the governance strategies put forward by the United Nations (U.N.), European Union (E.U.), and the United States. Other countries and intergovernmental organizations, including China, Japan, and the G7, have pursued new regulations as well but due to space considerations are beyond the scope of this guide.[45]

### U.N.

In December 2023, the U.N. published a report that outlines its guidelines for increased AI governance, given that this technology is developing at an unprecedented pace. This report argues that the greatest good from AI will come from its impact on sectors like healthcare, agriculture, and education. However, the report also outlines risks such as the increased potential for surveillance and decreased accountability for public officials.[46]

This report states the following main reasons for global cooperation on AI governance:

- A global approach to AI governance is needed to ensure that this technology is accessible to a variety of states and prevents harm on a global scale.
- A global approach is needed because this technology does not have many bounds—it can be applied in almost any context and to almost any type of task.
- Emerging players in the AI sphere—whether these are public officials, owners of private companies, or researchers—need some way by which they can be held accountable for the harms that may be caused by AI.
- There are pressing global challenges such as climate change that AI is well-positioned to help us solve, but a global effort at governing the use of this technology is the best way for us to solve these pressing problems.

### E.U.

The EU AI Act is a piece of legislation that attempts to outline goals and policies for AI development and risk mitigation. The Act outlines certain AI application as outlawed:

- “biometric categorisation systems that use sensitive characteristics (e.g. political, religious, philosophical beliefs, sexual orientation, race);

[43] Anderson, Janna and Lee Rainie. “Artificial Intelligence and the Future of Humans.” Pew Research Center. 2018.
[44] Columbus, Louis. “10 Ways AI is Improving New Product Development.” Forbes. 2020.
[45] Chee, Foo Yun. “Exclusive: G7 to Agree AI Code of Conduct for Companies.” Reuters. 2023.
Deck, Andrew. “Japan’s New AI Rules Favor Copycats Over Artists, Experts Say.” Rest of World. 2023.
Sheehan, Matt. “China’s New AI Regulations and How They Get Made.” Carnegie Endowment for International Peace. 2023.
[46] “Interim Report: Governing AI for Humanity.” UN AI Advisory Board. 2023.

- untargeted scraping of facial images from the internet or CCTV footage to create facial recognition databases;
- emotion recognition in the workplace and educational institutions;
- social scoring based on social behaviour or personal characteristics;
- AI systems that manipulate human behaviour to circumvent their free will;
- AI used to exploit the vulnerabilities of people (due to their age, disability, social or economic situation)."[47]

The justification for this law is the protection of European society and its economy: "The AI Act sets rules for large, powerful AI models, ensuring they do not present systemic risks to the Union and offers strong safeguards for our citizens and our democracies against any abuses of technology by public authorities. It protects our SMEs, strengthens our capacity to innovate and lead in the field of AI, and protects vulnerable sectors of our economy."[48]

**U.S.**
In the domestic context, President Biden has issued an Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence.[49] Unlike the E.U. regulation above, this Executive Order does not outlaw any specific applications of AI tools or development thereof. Rather, the Order sets guidelines and priorities for strategic partnerships.

The priorities set by the Executive Order on Safe, Secure, and Trustworthy AI include:

- Development of standards, tests, and tools by the National Institute of Standards and Technology (part of the U.S. Department of Commerce)
- Development of screening against biomedical weapons (which can be developed by AI technologies)
- Development of detection mechanisms for AI-generated content so as to decrease fraud and deception
- Strengthening of research that focuses on protecting the privacy of Americans
- Evaluation of ways that information and data is used (in particular commercially available data)
- Attention to algorithmic discrimination where civil rights are at issue
- Development of best practices for use of AI in criminal justice applications to ensure justice and equity

## Section 5: Other Ethical Dimensions

This Deliberation Guide has outlines some of the major threats and drawbacks as well as hopes and benefits associated with the recent advents in artificial intelligence. There will be significant difficulties balancing these against one another. For example, it is an open question how the net

[47] "Artificial Intelligence Act: Deal on Comprehensive Rules for Trustworthy AI." European Parliament. 2023.
[48] Ibid.
[49] "FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy AI." The White House. 2023.

benefits related to human welfare (supposing there will be net benefits) might weigh against concerns about autonomy, privacy, and other rights. This is because it is an unsettled ethical question about how to weigh the consequences of our actions against other goods. This is not a conflict that this guide will explore in detail, but it is something to consider when deliberating about these issues.[50] The basic problem is to what extent, if ever, is it permissible to curtail liberties and rights for the purpose of alleviating suffering or increasing wellbeing? As we reason about AI together, we may need to confront this difficult question and others that are as yet unsettled – both generally and in contexts beyond artificial intelligence.

Another one of these unsettled questions has to do with the nature of consciousness and what kinds of systems deserve our moral consideration.[51] If AI systems can perform functions that were heretofore reserved for humans—responding to others with emotional intelligence, solving complex problems, creating art, designing business solutions—then we may have to grapple with the question of how we ought to treat them, morally speaking.[52] As AI systems begin to make decisions regarding what humans used to decide, we may need to confront a confusing lack of moral responsibility for grave outcomes. For example, AI applications in autonomous weapons systems or drones can lead to questions about who can be held accountable for a war crime.[53]

A final set of questions relates to the origin of the data that AI systems are trained on. Do the designers of these systems owe something to the creators of this content? Are AI systems and their designers plagiarizing this content, or are they operating as humans do after they read something and then apply it in their own thinking? This is the subject of ongoing legal disputes and is beyond the scope of this guide, but it is another issue to consider.[54]

This Guide began with a quotation from Sheila Jasanoff, who suggests that AI should be on the agenda of deliberative democracy so that we can arouse our collective "ethical and political intelligence" and thoughtfully engage with both the risks and opportunities it presents. By seriously considering both the concerns and hopes associated with this technology, we will be better equipped to make informed decisions about how it can best be deployed – both now and in the future.

[50] "What is the Difference Between Deontology and Consequentialism?" Pediaa. 2019.
[51] Lenharo, Mariana. "AI Consciousness: Scientists Say We Urgently Need Answers." Nature. 2023.
Finkel, Elizabeth. "If AI Becomes Conscious, How Will We Know?" Science. 2023.
[52] Corbyn. Zöe. "Philosopher Peter Singer Says: 'There's No Reason to Say that Humans Have More Moral Worth than Animals." The Guardian. 2023.
[53] Sparrow, Robert. "Killer Robots." Journal of Applied Philosophy. 2007.
[54] Brynbaum, Michael M. and Ryan Mac. "The Times Sues OpenAI and Microsoft Over Use of Copyrighted Work." New York Times. 2023.
Coffman, Carla. "Does the Use of Copyrighted Works to Train AI Qualify as Fair Use?" Copyright Alliance. 2023.
Rosen, Michael, "Can AI Violate Copyright? A New Lawsuit Argues Yes." AEI. 2023.

## Setting Expectations (5 min)

In this section, we will review the “Expected Outcomes,” Deliberative Dispositions,” and “Conversation Agreements” below.

### Expected Outcomes of the Conversation

The purpose of this deliberation is to deepen our understanding of AI and how we should live with it going forward. Over the course of the deliberation, we will have the opportunity to listen to the perspectives of our fellow deliberators as well as share our own thoughts about AI and then deliberate about AI governance. Finally, we will have reflected on our conversation, our areas of agreement and disagreement, and what we have learned from our time together.

### Deliberative Dispositions

The DCI has identified several “deliberative dispositions” as critical to the success of deliberative enterprises. When participants adopt these dispositions, they are much more likely to feel their deliberations are meaningful, respectful, and productive. Several of the Conversation Agreements recommended below directly reflect and reinforce these dispositions, which include a commitment to egalitarianism, open mindedness, empathy, charity, attentiveness, and anticipation, among others. A full list and description of these dispositions is available at https://deliberativecitizenship.org/deliberative-dispositions/.

### Conversation Agreements

In entering into this discussion, to the best of our ability, we each agree to:

1. Be authentic and respectful
2. Be an attentive and active listener
3. Be a purposeful and concise speaker
4. Approach fellow deliberators’ stories, experiences, and arguments with curiosity, not hostility
5. Assume the best - and not the worst - about the intentions and values of others, and avoid snap judgements
6. Demonstrate intellectual humility, recognizing that no one has all the answers, by asking questions and making space for others to do the same
7. Critique the idea we disagree with, not the person expressing it, and remember to practice empathy
8. Note areas of both agreement and disagreement
9. Respect the confidentiality of the discussion
10. Avoid speaking in absolutes (e.g., “All people think this,” or “No educated people hold that view”)

## Getting to Know Each Other (10 min.)

In this section, we will take less than a minute to share our names and answer one of the questions below.

1. What are your hopes and concerns for your family, community and/or country?
2. What would your best friend say about who you are?
3. What sense of purpose / mission / duty guides you in your life?

## Concerns & Hopes About AI (10 min.)

In this section, we will each take 1-2 minutes to answer the question below, without interruption or crosstalk. After everyone has answered these questions, the group is welcome to take a few minutes for clarifying or follow up questions and responses. Continue exploring the topic as time allows.

> **What are your biggest concerns about the impact of AI? What are your biggest hopes about the impact of AI? Why these particular concerns and hopes?**

## The Transformative Impact of AI (15 min.)

We will now address the potentially transformative impacts of AI and how we should navigate these together. We will each choose one of the questions below to answer, and then together we'll explore our areas of agreement and disagreement. We can also generate additional ideas that may transcend and elicit more support than existing proposals.

> **AI has the potential to be a socially transformative technology—as transformative as the industrial revolution or even as the agricultural revolution.**
>
> **What can we do as communities in order to best imagine what a fundamentally unknown future could look like?**
>
> **Should we be conservative and skeptical or optimistic and entrepreneurial when it comes to this future?**

Once we have all had a chance to address this question, discuss our answers together, and note where we agree and disagree, please move on to the next section.

## Navigating AI's Impact Together (15 min.)

We will now address how we should balance AI's potentially beneficial applications against the potential risks it poses to humanity and society. We will each address one of the questions below individually before moving on to discussion.

> **What are some reasons to regulate AI more and some reasons to regulate AI less?**
>
> **The AI governance proposals reviewed in the guide range from vague to specific. If you were to brainstorm policies for AI regulation right now, what would be your top priorities? Why?**

## Reflections (10 min)

While today's conversation is an important step in the journey, effectively managing the difficulties presented by new technologies is necessarily an ongoing effort. Please reflect on the insights from your discussion with your fellow participants today, and then answer one of the questions below without interruption or crosstalk. After everyone has answered, the group is welcome to continue exploring additional questions as time allows.

1. What was most meaningful or valuable to you during this deliberation?
2. Where are the areas of both agreement and disagreement in your group?
3. Have any new ways to think about this issue occurred to you as we have talked today? Any new ideas that might transcend our current way of conceiving of the problem and its potential solutions?
4. Was there anything that was said or left out from the discussion that you think should be addressed with the group? Are there any perspectives missing from this conversation that you feel would be important to hear?
5. What did you hear that gives you hope for the future of conversations about public issues related to AI?
6. Is there a next step you would like to take based upon the deliberation you just had?

# About This Guide

**Writer:** Sara Copic

**Executive Editor:** Graham Bullock

**The Deliberative Citizenship Initiative**

The Deliberative Citizenship Initiative (DCI) is dedicated to the creation of opportunities for Davidson students, faculty, staff, alumni, and members of the wider community to productively engage with one another on difficult and contentious issues facing our community and society. The DCI regularly hosts facilitated deliberations on a wide range of topics and organizes training workshops for deliberation facilitators. To learn more about these opportunities, visit www.deliberativecitizenship.org.

**DCI Deliberation Guides**

The DCI has launched this series of Deliberation Guides as a foundation for such conversations. They provide both important background information on the topics in question and a specific framework for engaging with these topics. The Guides are designed to be informative without being overwhelming and structured without being inflexible. They cover a range of topics and come in a variety of formats but share several common elements, including opportunities to commit to a shared set of Conversation Agreements, learn about diverse perspectives, and reflect together on the conversation and its yield.  The DCI encourages conversations based on these guides to be moderated by a trained facilitator. After each conversation, the DCI also suggests that its associated Pathways Guide be distributed to the conversation's participants.

**DCI Pathways Guides**

For every Deliberation Guide, the DCI has also developed an associated Pathways Guide, which outlines opportunities for action that participants can consider that are related to the covered topic. These Pathways Guides reinforce the DCI's commitment to an action orientation, a key deliberative disposition. While dialogue and deliberation are themselves important contributors to a healthy democracy, they become even more valuable when they lead to individual or collective action on the key issues facing society. Such action can come in a range of forms and should be broadly understood. It might involve developing a better understanding of a topic, connecting with relevant local or national organizations, generating new approaches to an issue, or deciding to support a particular policy.

If you make use of this guide in a deliberation, please provide attribution to the Deliberative Citizenship Initiative and email dci@deliberativecitizenship.org to tell us about your event. To access more of our growing library of Deliberation Guides, Pathways Guides and other resources, visit www.deliberativecitizenship.org/readings-and-resources.