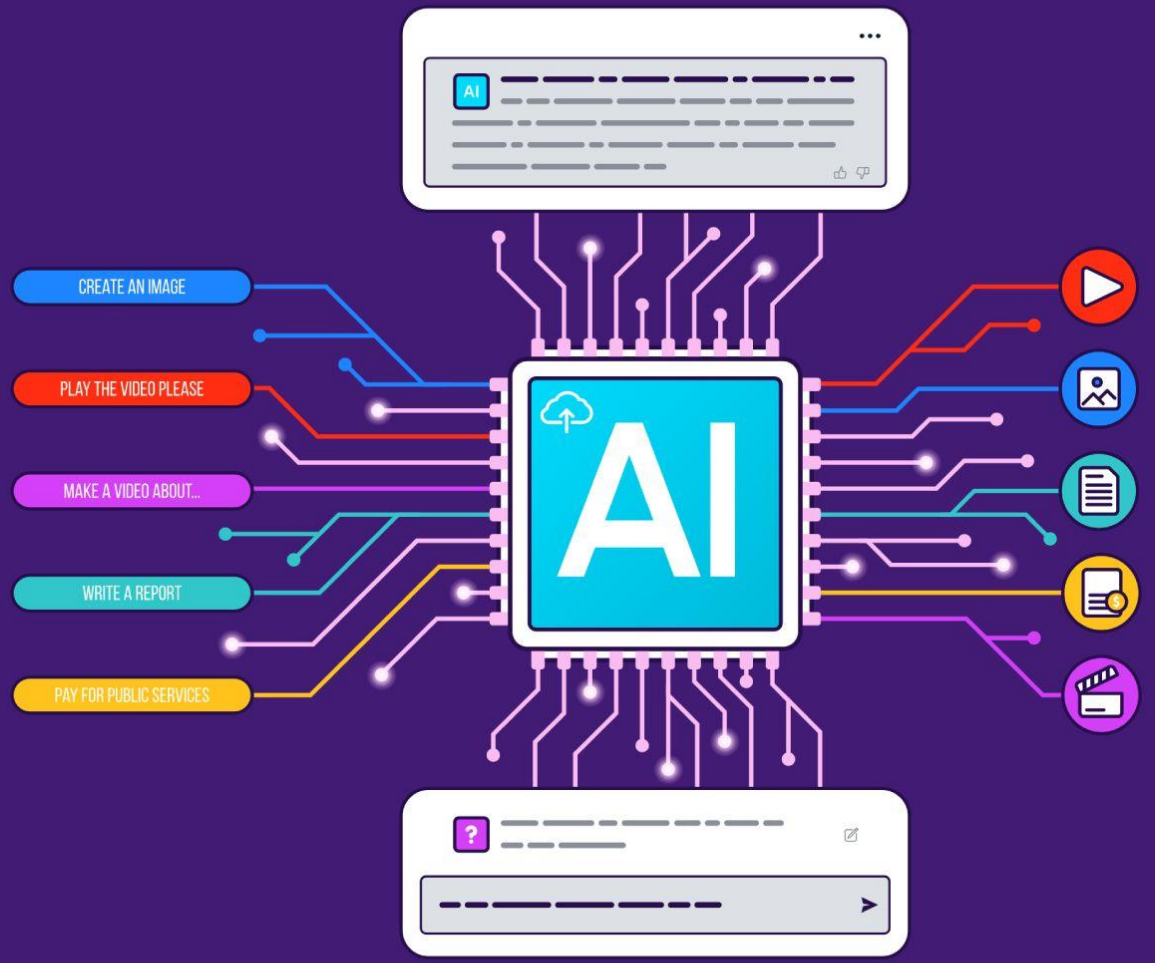
# Ethics in the Age of Generative AI



# Distinguishing responsible tech from human behavior

Distinguishing responsible tech from human behavior
Generative AI is already transforming every aspect of the human experience, and I'm the kind of optimist who believes that these tools will make us more human, more creative, inspired and connected. In the banking sector, AI and machine learning are helping us identify people who might have opportunities to be more financially secure, to save more, to increase retirement contributions and to have better pathways to economic opportunity. In agriculture, AI models are helping to predict large weather events so farmers and producers can understand when additional insurance is wise or help them know exactly when to plant or harvest to best maximize their economic returns. And maybe closer to home inside of organizations, AI is transforming human resources, helping managers understand how to inspire better performance from their teams, correct potential biases or discrimination and make sure that promotions are truly merit based. But even as I'm excited about these potential opportunities ahead, I'm also convinced that we need to make sure we build these tools with positive intention with a grounding in ethical and responsible reasoning. And I'm not alone. From the very beginning, developers of artificial intelligence have known the incredible power of these tools and consider the ethical quandaries that might arise when we deploy them. In recent years, these apprehensions have reemerged with a sharper edge. Researchers and now policymakers are exceptionally concerned by the potential ways that AI could perpetuate bias, could make the world more unequal and could do so in ways that are invisible to us. Many of the ethical concerns that AI researchers have worried about are coming to light in a very real way. Consider the idea of deep fakes, tools that might create a persona or an avatar that's pretending to be a trusted person in your organization, delivering fraudulent information to your customers, or even advising them to take dangerous or potentially risky courses of action. We've seen the advent of chatbots everywhere and we know that without ethical design, chatbots might give false advice maybe to medical practitioners or to students. Information that sounds logical, but is in fact based on inaccurate or untruthful information, and has never been audited by a human being. And issues about fundamental ownership, questions of legal and copyright. Who owns the ideas and products that are created by generative AI? We're just beginning to resolve them. Join me for a moment in an example. Let's imagine that your company has deployed a new AI system to support the HR function, scanning resumes of applicants to identify potential interviewees. At first glance, the tool works incredibly well, operating just as quickly and providing the same number of candidates as your human support team. But as you dig deeper, some disturbing patterns emerge. The tools prioritizing a particular gender, folks from a particular address or neighborhood or with a particular pattern to their work history. And you realize that these are the same biases that the humans in the dataset that trained the tool were expressing. We have a challenge here. This algorithm now is providing bad recommendations but humans are the ones to make the decisions about who they will interview. In situations like these, we need to ask ourselves, what is the highest standard of responsible human behavior? What actions can we take to best promote fairness and dignity? And is it possible that we've trained an AI to provide an answer that's lower than that highest standard? When this happens, what can we do to help humans make better decisions based on the algorithm's recommendations? If you're eager to learn more about how to answer these questions, stay with me. Later in this course, we're going to cover step by step how to resolve dilemmas just like these.

# Understanding Vilas' ethical AI framework

Understanding Vilas' ethical AI framework
I'm so excited about how quickly we're building new generations of AI tools, but I know that we need to make sure we're designing tools that support the future we want to create, equitable, sustainable, and thriving. And to do this, we're going to have to come up with new frameworks for ethical creation just as quickly as we advance the frontier of innovation. So how do we translate intuitions and hopes into clear principles for decision making? I'd like to share with you a three part framework that I use for evaluating and advising organizations on the creation of new ethically grounded AI tools and it works equally well for technologists and non technologists. The three pillars of the framework are responsible data practices, well-defined boundaries on safe and appropriate use and robust transparency. Let's start by talking about responsible data practice. This is the starting point for all ethical AI tools. Any new technology is only as ethical as the underlying data that it's trained on. For example, if the majority of our consumers to date have been of a particular race or gender when we train the AI on that data, we'll continue to only design products and services that serve the needs of that population. As you consider building or deploying any new tool you should ask what's the source of the training data? What's been done to reduce explicit and implicit bias in that dataset? How might the data we're using perpetuate or increase historic bias? And what opportunities are there to prevent bias decision making in the future? The second part of the framework is the importance of creating well-defined boundaries for safe and ethical uses. Any new tool or application of AI should begin with a focused statement of intention about the organization's goals and an identification of the population that we're trying to serve. So for example, a new generative AI tool that can write news articles. Well, it could be used to help tell the stories of a wider range of underrepresented voices. We could use it in new languages or it could perpetuate misinformation. When considering ethical use, you should ask who's the target population for this tool? What are their main goals and incentives and what's the most responsible way to make sure we're helping them achieve those goals? The third part of the framework is robust transparency. We need to consider how transparent the recommendations of the tool are, and that includes how traceable those outcomes are. This allows for human auditing of ethical accountability. When it comes to transparency, you should ask how did the tool arrive at its recommendation? And sometimes it's not possible to know, but if so what are other ways we have of testing its fairness? Is it possible for decision makers to easily understand the inputs, analysis, outputs, process of the tool? And finally, have you engaged with a broad range of stakeholders to make sure that this tool promotes equity in the world. As you embark on building and using increasingly more complicated ethical AI tools, this framework of responsible data, well-defined boundaries and robust transparency should provide you with a foundation for making smarter, more informed decisions.

# Applying Vilas' framework in a real world situation

Applying Vilas' framework in a real world situationNow it's time to put the framework we covered in the last video into practice. I want you to consider the following scenario involving the CTO of a technology company. Sarah enters the conference room for an emergency meeting, something serious is happening. She's told that the company's new AI driven chatbot designed to help customers with online orders has been making some inappropriate, inaccurate, and even offensive responses to customers. Sarah knows that this isn't just a product issue, it's an issue that's going to be grounded in ethical decision making. She knows her immediate step is easy. She needs to take the chatbot offline, and she does so, but then she has to figure out what her next step is. As a technologist, she knows to start with data. In other words, how was this tool trained? From talking with her team, she learns that the underlying data set came from an unscrubbed set of internet conversations. In the rush to production, the team didn't run the data set through a set of filters and tools. She knows what the next step is. She directs her team to use a new data set primarily composed of the company's own database of interactions with customers and only after scrubbing the data of any personal information, ahe directs that the model be run through a number of bias detection processes and filters but she knows that data isn't the end of the problem. Sarah finds out as she continues to inquire that customers are using the chatbot to do more than support customer service. They're taking the opportunity to have far ranging conversations on topics that have nothing to do with the company or the product. She knows the technology team should have reviewed other ways that customers might use the tool and considered ethical safeguards. Because the scope of use is widened, the team needs to limit what subjects the chatbot discusses with a customer, and they need to make sure that responses are tailored to the specific expertise that that chatbot is supposed to have. Sarah brings in the customer support team and she engages frontline workers on how they experience conversations with customers. She wants to know what do clients usually want to talk about. Using that information and engaging a shared design process, the team builds entirely new boundary conditions for topics that are relevant for the chatbot to discuss and limit excessive non-business conversation. Finally, Sarah realizes that she queries the tool that she has no way to explain some of the really insensitive outputs that the chatbot is producing. Her team needs to allow for better traceability and evaluation of the tool's outputs. So Sarah encourages the team to build multiple input-output checkpoints, and she encourages the creation of an internal audit process to regularly monitor and check outputs from the chatbot. Accompanying this, she has a risk assessment and response framework to allow a user to flag inappropriate conversations in real time so the team can address issues immediately. Now, let's acknowledge this is a significant effort by the company. It could range from weeks to months, but if the company had followed ethical practices when designing the chatbot, this expense, time, and stress could have been totally avoided. Ethical analysis needs to be intertwined with the initial design of new products and at every phase of deployment. Now there's a happy ending here, acting on our framework, Sarah's company is able to address the ethical dilemma surrounding its chatbot and get back online selling happy widgets to happy customers. How would your organization handle a dilemma like the one faced by Sarah's company? Which steps will you take today to center ethical analysis in your decisions about AI product design?

# Organizing data with ethics in mind

Organizing data with ethics in mind

- I remember standing at the top of the Burj Khalifa, the tallest building in the world. As I looked out on that view, a part of me was wondering just how strong the foundations must be to protect the people inside. These days, I look at generative AI models that can do amazing things and I find that same part of me evaluating the safety, trust, and design that ensures that these models will inspire and protect all of us. AI models are built on top of data. So let's talk about the importance of ethically organizing and using your institution's data. By taking an ethical approach, you'll reduce the risk to your organization and you'll increase the value of that data as an organizational asset. There are three goals in effective and ethical data organization: the first is prioritizing privacy, the second reducing bias, and the third, promoting transparency. The first consideration is prioritizing privacy. Almost every organization collects sensitive data about customers and employees; things like personal healthcare information or financial and banking details. And customers and employees trust the organization with this data so it's important to handle it sensitively and ethically. Failing to uphold this trust can expose a company to liability and reputational harm, and maybe, most importantly, erode trust with your customer. So to test your company's practices, you can lead a privacy audit. During a privacy audit, you build a comprehensive understanding of what data your organization has, how it was collected, how it's stored, and how it's administered. The results of a policy audit inform recommendations to create or adapt your existing privacy policy to protect sensitive data. With a privacy policy in place, the next step is to create a training curriculum for all employees that focuses on understanding why sensitive data must be handled securely and advises them of their responsibilities. The second goal is reducing bias in data collection and in data use. Bias in data can arise from a number of sources and understanding how it makes its way into your dataset requires genuine curiosity in your analysis. To start a bias audit, be curious about whether the data really represents the population you're trying to serve. For example, I recently worked with an organization building AI for cancer screening. And as they tried to deploy this tool, they found that early models exclusively use training data from the global north, requiring a retraining of the model to make it useful for a global population. So does your dataset represent inputs from a diversity of individuals across race, gender, age, and more? Are we asking the right questions when we collect data? You might also consider whether your data collection process was accessible to differently-abled people. And finally, once the data is collected, you might consider whether a team with relevant and diverse lived experience has an opportunity to analyze and interpret this data to reduce the risk of potential bias. Bias is especially important when we attempt to explain how our algorithms make recommendations that have real impacts on people's lives. For example, recent studies have shown that early attempts to automate hiring have propagated existing biases and employment practices. Understanding the bias in the data helps us minimize the negative impacts of bias in the algorithm. After you've completed your privacy and your bias audits, transparency is the final step in the process. You want to be able to explain to all stakeholders, your customers, your employees, your suppliers, your regulators, how data is collected and used. You might consider publishing a data governance framework or a data transparency statement to help your stakeholders understand what you do with their data. And you should also make it clear that individuals can access any data that you might have stored about them and have rights on how you might use it on an ongoing basis. Organizing and understanding your data helps you understand your customers better, ensure they're well represented, weed out biases, and builds a stronger foundation for your AI products and tools.

# Preparing technology teams to make ethical decisions

Preparing technology teams to make ethical decisions
- Technology teams face some special challenges when it comes to ethical decision making. You know, teams are asked to build new technologies for business challenges, but they're not always asked to consider upfront the social expectations and challenges that come with that design. Together we'll explore some ethical dilemmas that might face technology teams and learn together how to create a culture that promotes ethical decision making and accountability. Technology teams are unique in large organizations. First, they're often composed of individuals with really specific skills and expertise, which is frequently not so well understood by others in the organization, and sometimes even by the team's own managers. Second, technology teams often work at an extremely fast pace under tight deadlines. This means they have little extra time and resources to audit or to reflect on the consequences of their decisions on users or on the wider society. And finally, technology teams are often subject to specific regulatory requirements, for example, GDPR and associated laws in Europe. All of these considerations mean that it's important that technology teams have a strong, internal ethical culture, along with external oversight and accountability, to make sure that we are making decisions ethically. Ethical decisions that technology teams might face include: ensuring the security and privacy of data collection; storage, use and reuse; creating and auditing algorithms to ensure that they're fair and they're free from bias; and understanding the data storage and other environmental impact of their decisions and considering opportunities to reduce technology's carbon footprint. Can you think of any others? Every team will face their own unique ethical challenges, which is why it's vital to foster a culture of ethical decision making, one where teams can respond to an array of challenges as they arise, or even, hopefully, prevent them from arising in the first place. Here's a few steps you can take to create a culture of ethical decision making. First, foster a culture of ethical communication within your team. Encourage every team member to openly raise questions and concerns about the ethical use of technology. Here's an idea. Start meetings by focusing on a recent ethical challenge that your team faced and discussing how it was resolved. Reward and celebrate team members who appropriately raise and resolve ethical issues and make it a part of your day to day. Next, you might consider establishing a technology-specific training curriculum for your team, focusing on emerging technologies and the ethical challenges that team members might face when they begin to deploy them. You should explicitly consider ethical challenges at the start of every project. Before launching a new initiative, your team should come together to discuss possible ethical dilemmas and consider potential remediations and decide on a path forward. And when you feel like you need extra support, you should reach out to academics or philosophers, individuals who can help your team understand the full spectrum of ethical challenges that might fit within your work and bring them into a framework focused on advancing both your product and social wellbeing. Giving your technology team the tools to ensure that they can make decisions that align with the company's values and wider social ethics is essential. It sets your team up for success and it empowers them to solve new challenges as they arise in the future.

# Preparing C-Suite in directing responsible AI

Preparing C-Suite in directing responsible AI

- CEOs and the C-Suite play a critical role in building cultures of responsible AI. They set the tone by establishing practices and principles and they ensure that every individual in the organization feels like they're a part of making ethical decisions. Earlier in this course, we discussed the example of Alice Wong and her company that faced a critical dilemma around the deployment of an AI chat bot. In that instance, if we were advising the C-suite of the company, we'd start with the following recommendations. First, to make sure that a responsible AI policy and governance framework is in place. This is a statement from the C-Suite about how the organization should design and manage AI technologies. It should describe how to make ethical decisions. It should protect privacy, and it should focus on the elimination or reduction of bias. For example, the C-suite might mandate that AI tools are trained with diverse data sets or they might require that chatbots are always identified as an AI and not impersonating a human customer support agent. These guiding principles create a shared set of values that everyone from data scientists to supervisor, to field staff, can use to evaluate and guide the deployment of artificial intelligence. Next, we might advise the C-Suite, provide an maybe even mandate responsible AI training and education for every person in the organization. This method of democratizing decision-making around AI tools can be very powerful, can bring business knowledge, present and frontline service fields to help train and develop internal models. For example, in Alice Wong's example, Alice relied on customer service agents with years of direct customer experience, helping solve consumer challenges to validate the recommendations of the AI models. Empowering these agents to understand the limitations of the model can increase the quality of their feedback. Then, C-Suites should insist on building ethical AI elements into all of their technologies and conduct regular audits. The C-Suite can identify specific metrics such as customer satisfaction and create regular reporting mechanisms to ensure that the company's AI practices are aligned with responsible AI principles. Here's an example. That might look like setting up monthly standups, where technology executives join the C-suite and present ethical challenges that have emerged in the past month. This could be the start of a dialogue with C-Suite executives understanding and documenting ongoing interventions and improving the ethical nature of the product. Much like safety practices in other industries, this practice socializes and makes ethical AI development a shared and accountable responsibility. Finally, the C-Suite might consider hiring a chief AI ethics officer. The company might establish a specific senior role, focused on AI ethics that can develop and oversee the use of responsible AI practices and serve as a central audit for other departments. They should understand the intersection of the business the technology and the customer experience and they could provide a check-in balance for technology development to ensure that community voices are also present and ensure that potential risks are identified early in the creation process. The C-Suite sets the tone for responsible AI across the organization, creating strong policies, ensuring that there's appropriate training, establishing monitoring and reporting mechanisms, and potentially creating roles focused on AI Ethics. With the C-Suites primary responsibility for guiding responsible AI, we can build cultures that focus on ethical decisions, even as we deploy great new products.

# Preparing the Board of Directors to manage risk and opportunity in AI

Preparing the Board of Directors to manage risk and opportunity in AI

- I work with a lot of board directors, and they're having trouble sleeping right now because I can tell you, even the smartest AI can't possibly predict all the ethical dilemmas that might arise from these new technologies, and yet it's the board's job to make sure that organizations are prepared, to make sure that new technologies are deployed in the best interests of all stakeholders. Board directors have a legal and an ethical obligation to act in the best interest of an organization and its stakeholders. Board members have different responsibilities from the organization's C-Suite. As you might recall from the previous video, the C-Suite's responsible for the day-to-day operations of the company, making real-time decisions about when and how to use new technologies. For a board of directors to ensure that an organization's use of AI aligns with ethical values and regulatory requirements, they should first make sure that the organization has policies and procedures in place to identify and address ethical concerns that might arise in the use of AI. These policies should be designed to mitigate risks, including bias, privacy, and security, but they should also create opportunities for individuals with ethical concerns to come forward, including directly to the board when and if necessary. The board should make sure that the organization has their resource and the expertise necessary to manage the ethical risk posed by AI effectively. Do you know if there's a board policy on the ethical use of AI at your organization? If not, it may be worth it to find out. Board members also have specific responsibilities and obligations to regulators. They're accountable for ensuring that the organization complies with statutory requirements related to the use of AI as well as monitoring all of the new regulations and understanding the possible impact for the organization. To be effective at these responsibilities, the board of directors should establish a dedicated AI committee to oversee ethical AI practices within the organization. This committee should receive outside advice from experts in the fields of AI, ethics, and law, and should be responsible for providing guidance on issues such as bias, transparency, and accountability. They should, in particular, be available to advise the C-Suite on significant decisions about these matters. The board of directors has a critical role in managing ethical risks within organizations. They must ensure that the organization has appropriate policies and procedures in place to address ethical concerns, but they also have to provide guidance and oversight to the executive team, ensure compliance with regulatory requirements, and they should establish a dedicated committee to oversee ethical AI practices within the organization. By taking these actions, boards of directors can build trust with stakeholders and they can ensure that organizations are focused on the long-term success of their AI efforts.

# Consulting your customers in building AI

Consulting your customers in building AI
- In the previous videos, we've explored the importance of technology teams, of C-suites, of boards of directors in ensuring responsible and ethical AI practices. But what about the most important stakeholder of all, our customers? Designing great products means that we have to understand and incorporate their preferences, their needs, and their wants into our product design. I'd like to share with you a powerful framework for listening to our customers, an acronym that I call LISA. First, we listen to users before we start to build. Developing and launching new technologies requires a clear understanding of our users' goals, their needs, and their fears. It can be difficult to create a product when we haven't heard what our customers expect. Research has shown that users care deeply about the experience and usability of the technology products they use. In a recent survey conducted by the Nielsen Group, 85% of respondents said they would not return to a website or a product if they had a poor user experience. The second part of the LISA framework, how do we involve our customers in design decisions? We know that our customers want to feel that their opinions matter, and we want to include them in design decisions that can be crucial to building our products to meet their needs. This can be especially helpful when we're seeking to ensure that our decisions reflect the full diversity of our user base. Here's an example. In 2016, Airbnb launched its Community Commitment initiative. They gathered input from users on ways to make their platform more inclusive and welcoming for people from diverse backgrounds. This simple practice led to the creation of brand new features, such as the filters for gender neutral pronouns or the ability to search for wheelchair accessible listings. Another way to involve customers and design decisions is creating a user advisory board, a group of users who are invited to provide feedback and input on new features, designs, and other aspects of product development, even as you're in the design environment. For example, Microsoft has done this with a customer advisory board made up of customers from a range of industries and backgrounds who are invited to provide feedback on Microsoft's products and services, even while they're still in development. By including users from diverse backgrounds and experiences, we get a much wider range of perspectives on how to design and build products that actually meet the needs of all users. The third part of the LISA framework, sharing simple and transparent privacy policies. By prioritizing user privacy, we focus on building trust with our users and we create a more loyal user base. This is so important. According to a survey by Pew Research Center, 79% of adults in the US are concerned about how companies use their data. This can be a barrier that keeps users from engaging with your products, even when these products might actually help them improve their lives. There are ways that we can do this that include using plain language to explain data collection practices, providing customers with clear opt-in and opt-out options, and implementing privacy by design principles into your core technology development process. The final part of the LISA process is auditing our work and inviting outsiders in to help hold us accountable. Every existing and new technology product should be audited on a regular cycle, a process where you review the purpose of the product, potential risk to users, and maybe most importantly, the possible unintended consequences that might happen because of that product. Here's an example where this works well. Google developed an AI principles framework which guides their development and use of AI technology. That framework includes principles like fairness, privacy, and accountability, and it's used as a guide for conducting regular audits of their AI systems, identifying potential risks, and coming up with remediation. There are a number of risks we should be aware of. They could include bias or data privacy concerns, or even security vulnerabilities. And we know that organizations that bring in users and audit these risks do a great job of responding to them. At OpenAI, the trust and safety team is responsible for identifying potential risks associated with AI technologies. The team includes experts in the fields of computer science, law, and philosophy. They work together to ensure that OpenAI's technology is ethical and responsible. Once potential risks are identified, then we have to step back and conduct a risk assessment to evaluate the likelihood and possible impact of those risks. This assessment should consider the potential impact on users as well as the business impact of the risk. For example, at IBM, there's a separate AI governance board that's responsible for conducting risk assessments for AI systems. The board evaluates the potential risks and makes recommendations to mitigate those risks and improve safety for users. Building great products means listening to our customers, and using the framework we've described here, affectionately termed LISA, lets us listen to our customers, lets us involve them in decision making, shares privacy practices, and ensures that we're living a practice of regular audit and accountability. These practices mean that we can build better trust with our customers and ensure that technologies meet the needs and preferences of communities, not just the ones we serve today, but the ones we aspire to serve in the future.

# Communicating effectively organizationally and globally

Communicating effectively organizationally and globally
- In this course, we've talked about the decisions you will make as a leader in defining responsible AI practice and the roles of those around you in firms and organizations. But as we speed into the transformation the generative AI reflects, we know these products will touch every person on the planet, and it's important for us to consider the interests of various stakeholders. I use a convenient acronym called ethics to remind myself of the specific responsibilities of each stakeholder group. Using this mnemonic can help ensure that you're fulfilling your responsibilities to core AI action and including stakeholders across the globe. The ethics framework outlines six key stakeholder responsibilities for responsible AI. First, E for executives and board members. Top management has a responsibility to establish ethical AI cultures across organizations. This includes setting ethical guidelines and standards, ensuring that ethical considerations are integrated into decision-making processes, and allocating resources for ethical AI deployment and development. The T in ethics, technologists, engineers and developers who have a responsibility to design and develop products that are transparent, explainable, and accountable, avoiding bias in data and algorithms, ensuring that systems are secure and safe, and developing AI systems that are compatible with existing ethical frameworks. H, human rights advocates. Human rights advocates have a responsibility to ensure that the systems that technologists build respect human rights and dignity. This includes monitoring how AI systems are being used by vulnerable groups, identifying potential human rights violations, and advocating for the ethical use of AI. Is for industry experts. Industry experts have a responsibility to share their knowledge and expertise on the ethical implications of AI. This might include providing guidance on developing tools, identifying potential risks, and collaborating with other stakeholders to address ethical concerns. C, customers and users. Customers and users have a responsibility to provide feedback and insights. This could include communicating concerns, feedback to relevant stakeholders, participating in user-testing and feedback sessions, and staying informed about the ethical implications of AI. And finally, S for society at large. Let's acknowledge that this is a shared journey that all stakeholders have a responsibility to consider how these tools are changing the ways that we interact as humanity. This includes identifying and mitigating potential risks, promoting and advocating for transparency and accountability, and making sure that AI is used in a way that benefits society broadly. It's vitally important to coordinate these different stakeholders to create new spaces and forums for groups to come together. It's not enough to have each stakeholder playing their part. We have to coordinate across different stakeholder groups. Everybody needs to know what the others are doing and that means, we need to create new forums and new participatory mechanisms to make sure that stakeholders are working together to maintain ethical AI. Here's a few suggestions of what you might be able to do to promote the ethics framework. You could establish new mechanisms of clear communication between the stakeholders in your work, your customers, your executives, and your technology teams. You could develop training programs to educate employees and stakeholders about ethical considerations in AI. You could advocate to create a cross-functional team within your organization, or even a cross-organizations within an industry, bringing folks together from different departments to develop and implement guidelines and standards. You might consider developing a system for collecting and addressing user-feedback particularly around concerns and risks about AI systems. And you might consider engaging formally and informally with external stakeholders, human rights advocates, industry experts in civil society to ensure that we're considering the broadest possible implications of AI. We're at a moment in time where building products feels like the most important way to explore what generative AI can do for humanity. And yet, if we build products without also asking how those products will be used, what needs they serve, and how they'll impact vulnerable people, we miss an opportunity to use AI to make humanity better. The ethics framework gives us a way to encourage and involve stakeholders from across society, to make sure that as we build products, we're also building an AI ecosystem for the future of humanity.