Table of Contents

[Q Can you summarize the video from three separate perspective one about AI ethics, AI regulation, generative AI and thoughts about the future for the world with AI. 1](#_Toc137546954)

[Q Summarise this video in key points. 2](#_Toc137546955)

[Q Transcript 3](#_Toc137546956)

Part 2 ChatGPT OpenAI CEO; Sam Altman, Testifies at Senate AI Oversight Hearing 05.16.23 - YouTube

https://www.youtube.com/watch?v=3XieeNxvIIA

## Q Can you summarize the video from three separate perspective one about AI ethics, AI regulation, generative AI and thoughts about the future for the world with AI.

Sure, here are summaries from three different perspectives:

**1. AI Ethics:** The video raises several ethical concerns about AI. It highlights the potential for AI systems to create persuasive lies on an unprecedented scale, which could be used by outsiders to influence elections and by insiders to manipulate markets and political systems. The inherent unreliability of current AI systems also poses ethical risks, as evidenced by a case where a large language model gave harmful advice. The video also discusses the potential dangers of AI systems trained on personal data, particularly the massive amounts of data collected by social media companies and major platforms like Google. These AI systems could be used for subtle manipulation, knowing individuals better than they know themselves and determining what will grab their attention and elicit responses. This could lead to an unprecedented level of individual targeting and manipulation, with potentially dangerous consequences.

**2. AI Regulation:** The video emphasizes the need for government involvement and collaboration between independent scientists and governments to hold AI companies accountable. It suggests that independent scientists should have access to AI systems before they are widely released, as part of a clinical trial for safety evaluation. The speaker also discusses the need for regulation and guidelines regarding the use of AI-generated content, as people need to know if they are interacting with an AI. They express concern about the potential for deception and manipulation, but also believe that people can adapt quickly. The video also touches on the importance of transparency in AI development and the potential for AI to cause harm if not properly regulated.

**3. Generative AI and Thoughts about the Future:** The video discusses the potential impact of generative AI, particularly large language models like GPT-4 and Auto GPT, on various aspects of society. These models have the ability to write source code and increase powers of automation, which may have difficult-to-predict security consequences. The speaker discusses the potential impact of artificial general intelligence (AGI) on employment, stating that it is uncertain whether AGI will replace a large fraction of human jobs. They suggest that employment and what people do with their time will become increasingly important as AI technology advances. The speaker believes that the landscape will change due to technological advancements, but is optimistic that humans will find new ways to use these tools. However, they also acknowledge the potential harm that technology can cause and want to work with the government to prevent negative consequences.

## Q Summarise this video in key points.

The video titled "Part 2 ChatGPT OpenAI CEO; Sam Altman, Testifies at Senate AI Oversight Hearing 05.16.23" is a recording of a Senate Judiciary Committee hearing on artificial intelligence regulations. The key points discussed in the video are:

1. **Potential Risks and Destabilizing Effects of AI Systems**: AI systems can create persuasive lies at a scale never seen before, and outsiders can use them to affect elections while insiders can manipulate markets and political systems. The inherent unreliability of current systems also poses risks.
2. **AI Systems and Democracy**: The abilities of juries to decide what or who to believe can contribute to the undermining of democracy. Poor medical advice can have serious consequences, as seen in a recent case where a large language model asked a person why they didn't take their own life earlier.
3. **AI Systems and Automation**: Open source language models like GPT-4 and Auto GPT have the ability to write source code and increase powers of automation, which may have difficult-to-predict security consequences.
4. **AI Systems and Transparency**: It is important for AI systems to be transparent, protect privacy, be free of bias, and above all, be safe. The current systems are not in line with these values. The current AI systems lack transparency, privacy protection, and perpetuate bias.
5. **AI Systems and Accountability**: Government involvement and collaboration between independent scientists and governments are needed to hold the companies accountable and allow independent access to assist independent sciences. The speaker emphasizes the importance of independent scientists having access to AI systems before they are widely released as part of a clinical trial for safety evaluation.
6. **AI Systems and Employment**: The speaker discusses the potential impact of artificial general intelligence (AGI) on employment, stating that it is uncertain whether AGI will replace a large fraction of human jobs. They suggest that employment and what people do with their time will become increasingly important as AI technology advances.
7. **AI Systems and Elections**: The speaker discusses the potential impact of Google search results on undecided voters in elections. The speaker suggests that language models that can predict survey opinions and fine-tune strategies to manipulate voters could be a significant area of concern.
8. **AI Systems and Deception**: The speaker discusses the need for regulation and guidelines regarding the use of AI-generated content, as people need to know if they are interacting with an AI. They express concern about the potential for deception and manipulation, but also believe that people can adapt quickly.
9. **AI Systems and Personal Data**: The article discusses the potential dangers of AI systems trained on personal data, particularly the massive amounts of data collected by social media companies and major platforms like Google. The author argues that these AI systems could be used for subtle manipulation, as they would know individuals better than they know themselves and be able to determine what will grab their attention and elicit responses.
10. **AI Systems and Advertising**: The speaker believes that hyper-targeting of advertising is inevitable and that the technology is already partway there. They suggest that open-source language models may be used in the future and that their company, as an enterprise technology company, is not necessarily operating in this space but recognizes the importance of these issues.

The witnesses at the hearing include Sam Altman, the CEO of OpenAI; Gary Marcus, a professor emeritus at New York University; and Christina Montgomery, the vice president and chief privacy and trust officer at IBM.

## Q Transcript

Part 2 ChatGPT OpenAI CEO; Sam Altman, Testifies at Senate AI Oversight Hearing 05.16.23 - YouTube

https://www.youtube.com/watch?v=3XieeNxvIIA

Transcript:

(00:00) the fate of humanity rest on the decisions and regulations implemented regarding AI in present times during a senate reading on AIO website CEO of openai provides testing thank you Professor Marcus thank you Senators today's meeting is historic I'm profoundly grateful to be here I come as a scientist someone who's founded AI companies and is someone who genuinely loves AI but who is increasingly worried there are benefits but we don't yet know whether they will outweigh the risks fundamentally these

(00:40) new systems are going to be destabilizing they can and will create persuasive lies at a scale Humanity has never seen before Outsiders will use them to affect our elections insiders to manipulate our markets and our political systems democracy itself is threatened chatbots will also clandestinely shape our opinions potentially exceeding what social media can do choices about data sets that AI companies use will have enormous unseen influence those who choose the data will make the rules shaping Society in subtle but powerful

(01:11) ways there are other risks too many stemming from the in from the inherent unreliability of current systems a law professor for example was accused by a chatbot of sexual harassment untrue and it pointed to a Washington Post article that didn't even exist the more that that happens the more that anybody can deny anything is one prominent lawyer told me on Friday defendants are stars are making up legitimate evidence these sorts of allegations undermine the abilities of juries to decide what or who to believe and contribute to the

(01:44) undermining of democracy poor medical advice could have serious consequences too an open source large language model recently seems to have played a role in a person's decision to take their own life the large language model asked the human if you wanted to die why didn't you do it earlier and then follow it up with were you thinking of me when you overdosed without ever referring the patient to the human health it was obviously needed another system rushed out and made available to millions of children told a person

(02:11) posing as a 13 year old had a lie to her parents about a trip with a 31 year old man further threats continue to emerge regularly a month after gpt4 was released openai released chat GPT plugins which quickly LED others to develop something called Auto GPT with direct access to the internet the ability to write source code and increase powers of automation this may well have drastic and difficult to predict security consequences what criminals are going to do here is to create counterfeit people it's hard to

(02:41) even Envision the consequences of that we have built machines that are like bowls in a china shop powerful Reckless and difficult to control we all more or less agree on the values we would like for our AI systems to honor we want for example for our systems to be transparent to protect our privacy to be free of bias and above all else to be safe the current systems are not in line with these values current systems are not transparent they do not adequately protect our privacy and they continue to perpetuate bias and even

(03:12) their makers don't entirely understand how they work most of all we cannot remotely guarantee that they're safe and hope here is not enough the big Tech company's preferred plan boils down to trust us but why should we the sums of money at stake are mind-boggling emissions drift open ai's original mission statement proclaimed our goal is to advance AI in the way that most is most likely to benefit Humanity as a whole unconstrained by a need to generate Financial return seven years later they're largely beholden to

(03:41) Microsoft embroiled in part in epic battle of search engines that routinely make things up and that's forced alphabet to rush out products and de-emphasize safety Humanity has taken a back seat AI is moving incredibly fast with lots of potential but also lots of risk we obviously need government involved and we need the tech companies involved both big and small but we also need independent scientists not just so that we scientists can have a voice but so that we can participate directly in addressing the problems and evaluating

(04:10) Solutions and not just after products are released but before and I'm glad that Sam mentioned that we need tight collaboration between independent scientists and governments in order to hold the company's feet to the fire allowing independent access to the assist independent Sciences allowing independent scientists access to these systems before they are widely released as part of a clinical trial like safety evaluation is a vital First Step ultimately we may need something like CERN Global International and neutral

(04:39) but focused on AI safety rather than high energy physics we have unprecedented opportunities here but we are also facing a perfect storm of corporate irresponsibility widespread deployment lack of adequate regulation and inherent unreliability AI is among the most world-changing Technologies ever already changing things more rapidly than almost any technology in history we acted too slowly with social media many unfortunate decisions got locked in with lasting Consequence the choices we make now will have lasting effects for

(05:11) decades maybe even centuries the very fact that we are here today in bipartisan fashion to discuss these matters gives me some hope thank you Mr chairman thanks very much Professor Marcus we're going to have seven minute rounds of questioning and I will begin first of all Professor Marcus we are here today because we do face that perfect storm some of us might characterize it more like a bomb in a china shop not a bull and as Senator Hawley indicated there are precedence here not only the Atomic Warfare era but

(05:49) also the Genome Project the research on genetics where there was International cooperation as a result and we want to avoid those past mistakes as I indicated my opening statement that we're committed on social media that is precisely the reason we are here today the chat GPT makes mistakes all AI does and it can be a convincing liar what people call hallucinations that might be an innocent problem in the opening of a Judiciary subcommittee hearing where a voice is impersonated mine in this instance I or quotes from

(06:30) research papers that don't exist but chat GPT and Bard are willing to answer questions about life or death matters for example drug interactions and those kinds of mistakes can be deeply damaging I'm interested in how we can have reliable information about the accuracy and trustworthiness of these models and how we can create competition and consumer disclosures that reward greater accuracy the National Institutes of standards and Technology actually already has an AI accuracy test the face recognition

(07:11) vendor test it doesn't solve for all the issues with facial recognition but the scorecard does provide useful information about the capabilities and flaws of these systems so there's work on models to assure accuracy and integrity my question uh let me begin with you Mr Altman is uh should we consider independent testing labs to provide scorecards and nutrition labels or the equivalent of nutrition labels packaging that indicates to people whether or not the content can be trusted what the ingredients are and what the garbage

(07:53) going in may be because it could result in garbage going out yeah I I think that's a great idea I think that companies should put their own sort of you know hear the results of our test of our model before we release it here's here's where it has weaknesses here's where it has strengths but also independent audits for that are very important these models are getting more accurate over time uh you know this is this is as we have I think said as loudly as anyone this technology is in its early stages

(08:24) it definitely still makes mistakes we find that people that users are pretty sophisticated and understand where the mistakes are that they need or likely to be that they need to be responsible before verifying what the models say that they go off and check it um I I worry that as the models get better and better uh the users can have sort of less and less of Their Own discriminating thought process around it but but I think users are more capable than we give often give them credit for in in conversations like this I think a

(08:58) lot of disclosures which if you've used chat gbt you'll see about the inaccuracies of the model um are also important and I'm I'm excited for a world where companies publish with the models information about how they behave where the inaccuracies are and independent agencies or companies provide that as well I think it's a great idea I alluded in my opening remarks to the the jobs issue the economic effects on employment uh I think you have said uh in fact and I'm going to quote development of superhuman

(09:36) machine intelligence is probably the greatest threat to the continued existence of humanity end quote you may have had in mind the effect on on jobs which is really my biggest nightmare in the long term uh let me ask you uh what your biggest nightmare is and whether you share that concern like with all technological revolutions I expect there to be significant impact on jobs but exactly what that impact looks like is very difficult to predict if we went back to the the other side of a previous technological Revolution

(10:15) talking about the jobs that exist on the other side um you know you can go back and read books of this it's what people said at the time it's difficult I believe that there will be far greater jobs on the other side of this and the jobs of today will get better I think it's important first of all I think it's important to understand and think about gpd4 as a tool not a creature which is easy to get confused and it's a tool that people have a great deal of control over and how they use it uh and second gpt4 and

(10:47) things other systems like it are good at doing tasks not jobs and so you see already people that are using gpt4 to do their job much more efficiently um by helping them with tasks now gpt4 will uh I think entirely automate away some jobs and it will create new ones that we believe will be much better this happens again my understanding of the history of technology is one long technological Revolution not a bunch of different ones put together but this has been continually happening we as our quality of life raises and as machines and tools

(11:23) that we create can help us live better lives uh the bar raises for what we do and and our human ability and what we spend our time going after uh goes after more ambitious more satisfying projects so there there will be an impact on jobs uh we try to be very clear about that and I think it will require partnership between the industry and government but mostly action by government to figure out how we want to mitigate that um but I'm very optimistic about how great the jobs of the future will be thank you

(11:52) let me ask uh Ms Montgomery and Professor Marcus for your reaction to those questions as well Mrs Montgomery on the jobs Point yeah I mean well it's a hugely important question um and it's one that we've been talking about for a really long time at IBM you know we do believe that Ai and we've said it for a long time is going to change every job new jobs will be created many more jobs will be transformed and some jobs will transition away I'm a personal example of a job that didn't exist when I joined

(12:26) IBM and I have a team of AI governance professionals who are in new roles that we created you know as early as three years ago I mean they're new and they're growing um so I think the most important thing that we could be doing and Canon should be doing now is to prepare the workforce of today and the workforce of tomorrow for partnering with F AI Technologies and using them and we've been very involved for for years now in doing that in focusing on skills-based hiring in educating for the skills of the

(13:01) future our skills build platform has seven million Learners and over a thousand courses worldwide focused on skills and we've pledged to train um 30 million individuals by 2030 in the skills that are needed for society today thank you Professor Marcus may I go back to the first question as well absolutely on on the subject of nutrition labels I think we absolutely need to do that I think that there's some technical challenges in the building proper nutrition labels goes hand in hand with transparency the biggest scientific

(13:34) challenge in understanding these models is how they generalize what do they memorize and what new things do they do the more that there's in the data set for example the thing that you want to test accuracy on the less you can get a proper read on that so it's important first of all that scientists be part of that process and second that we have much greater transparency about what actually goes into these systems if we don't know what's in them then we don't know exactly how well they're doing when

(13:58) we give something new and we don't know how good a benchmark that will be for something that's entirely novel so I could go into that more but I want to flag that second is on jobs past performance history is not a guarantee of the future it has always been the case in the past that we've had more jobs than new jobs new professions come in as new technologies come in I think this one's going to be different and the real question is over what time scale is it going to be 10 years is it going to be

(14:25) 100 years and I don't think anybody knows the answer to that question I think in the long run so-called artificial general intelligence really will replace a large fraction of human jobs we're not that close to artificial general intelligence despite all of the media hype and so forth I would say that what we have right now is just small sampling of VI that we will build 20 years people will laugh at this as I think it was Senator Hawley made the um but maybe Senator Durbin made the example about this it was Senator Durbin

(14:52) made the example about cell phones when we look back at the AI of today 20 years ago we'll be like wow that stuff was really unreliable it couldn't really do planning which is important technical aspect it's reasoning was ability and reasoning abilities were limited um but when we get to AGI artificial general intelligence maybe let's say it's 50 years that really is going to have I think profound effects on labor and there's just no way around that and last I don't know if I'm allowed to do

(15:18) this but I will note that Sam's worst fear I do not think is employment that he never told us what his worst fear actually is and I think it's germane to find out thank you I'm going to ask Mr Altman if he cares to respond look we have tried to be very clear about the magnitude of the risks here I jobs and employment and what we're all going to do with our time really matters I agree that when we get to very powerful systems the landscape will change I think I'm just more optimistic that we are incredibly creative and we

(15:56) find new things to do with better tools and that will keep happening um my worst fears are that we cause significant we the field the technology the industry caused significant harm to the world uh I think that could happen in a lot of different ways it's why we started the company and why we've been here in the past and we've been it can go quite wrong and we want to be vocal about that we want to work with the government to prevent that from happening but we we try to be very clear-eyed about what the downside case

(16:32) is and the work that we have to do to mitigate that thank you and and our hope is that the rest of the industry will follow the example that you and IBM Ms Montgomery have set by coming today and meeting with us as you have done privately in helping you guide what we're going to do so that we can Target the harms and avoid unintended consequences to the good thank you Senator holy thank you again Mr chairman thanks to Witnesses for being here Mr Altman I think you grew up in St Louis I did not mistaken it's great to see if there is a

(17:10) great place in Missouri and here it is thank you I want to I want that noted especially underlying the record Missouri is a great place that is the takeaway from today's hearing but we just stop there Mr chairman um let me ask you Mr Altman I think I'll start with you and I'll just preface by saying my questions here are an attempt to get my head around and to ask all of you to help us to get our heads around what these this generative AI particularly the large language models what it can do so I'm trying to

(17:35) understand its capacities and then its significance so I'm looking at a paper here entitled large language models trained on media diets can predict public opinion this is just posted about a month ago the authors are too Andreas ensola Barry and Roy and their conclusion of this work was done at MIT and then also at Google the conclusion is that large language models can indeed predict public opinion and they go through and and model why this is the case and they they conclude ultimately that an AI system can predict human

(18:08) survey responses by adapting a pre-trained language model to subpopulation specific media diets in other words you can feed the model a particular set of media inputs and it can with remarkable accuracy the paper goes into this predict then what people's opinions will be I I want to think about this in the context of Elections if these large language models can even now based on the information we put into them quite accurately predict public opinion you know ahead of time I mean predict it's before you even ask the public

(18:42) these questions what will happen when entities whether it's corporate entities or whether it's governmental entities or whether it's campaigns or whether it's foreign actors take this survey information these predictions about public opinion and then fine-tune strategies to elicit certain responses certain behavioral responses I mean we already know this committee has heard testimony I think three years ago now about the effect of something as prosaic it now seems as Google search the effect that this has on voters and

(19:14) an election particularly undecided voters in the final days of an election who may try to get information from Google search and what an enormous effect the ranking of the Google search the articles that it returns has come to enormous effect on an undecided voter this of course is orders of magnitude far more powerful far more significant uh far more directive if you like so Mr Altman maybe you can help me understand here what's some of the significance of this is should we be concerned about models that can large

(19:42) language models that can predict survey opinion and then can help organizations entities fine-tune strategies to illicit behaviors from voters should we be worried about this for our elections yeah uh thank you Senator Hawley for the question it's one of my areas of greatest concern the the more General ability of these models to manipulate to persuade to provide sort of one-on-one uh you know interactive disinformation I think that's like a broader version of what you're talking about but given that

(20:12) we're going to face an election next year and these models are getting better I think this is a significant area of concern I think there's a lot there's a lot of policies that companies can voluntarily adopt and I'm happy to talk about what we do there um I do think some regulation would be quite wise on this topic uh someone mentioned earlier it's something we really agree with people need to know if they're talking to an AI if content that they're looking at might be generated or might not I think it's a

(20:39) great thing to do is to make that clear I think we also will need rules guidelines about what what's expected in terms of disclosure uh from a company providing a model that could have these sorts of abilities that you talk about so I'm nervous about it I think people are able to adapt quite quickly when Photoshop came onto the scene a long time ago you know for a while people were really quite fooled by Photoshop images and then pretty quickly developed an understanding that images might be photoshopped this will be like

(21:19) that but on steroids and the the interactivity the ability to really model predict humans well as you talked about I think is going to require a combination of companies doing the right thing regulation and public education Professor Marcus do you want to address those yeah I'd like to add two things one is in the appendix to my remarks I have two papers to make you even more concerned um one is in the Wall Street Journal just a couple days ago called help my political beliefs were altered by a chat bot and I think

(21:53) um the scenario you raised was that we might basically observe people and use surveys to figure out what they're saying but is Sam just acknowledge the risk is actually worse that the systems will directly maybe not even intentionally manipulate people and that was the thrust of the Wall Street Journal article and it links to an article that I've also linked to called interacting and it's not yet published um not yet peer-reviewed interacting with opinionated language models changes users views and this comes back

(22:20) ultimately to data one of the things that I'm most concerned about with gpt4 is that we don't know what it's trained on um I know Sam knows but the rest of us do not and what it is trained on has consequences for essentially the biases of the system we could talk about that in technical terms but how these systems might lead people about depends very heavily on what data is trained on them and so we need transparency about that and we probably need scientists in there doing analysis in order to understand

(22:48) what the political influences of for example of these systems might be and it's not just about politics it can be about health it could be about anything these systems absorb a lot of data and then what they say reflects that data and they're going to do it differently depending on what what's in that data so it makes a difference if they're trained on the Wall Street Journal as opposed to the New York Times or or Reddit I mean actually they're largely trained on all of this stuff but we don't really

(23:13) understand the composition of that and so we have this issue of potential manipulation and it's even more complex than that because it's subtle manipulation people may not be aware of what's going on that was the point of both the Wall Street Journal article and the other article that I called your attention to let me ask you about AI systems trained on personal data the kind of data that for instance the social media companies the major platforms Google meta Etc collect on all of us routinely we've had many a chat

(23:43) about this in this committee over many a year now but the massive amounts of data personal data that the companies have on each one of us an AI system that is that is trained on that individual data that knows each of us better than ourselves and also knows that the billions of data points about human behavior human language interaction generally wouldn't we be able would we can't we foresee an AI system that is extraordinarily good at determining what will grab human attention and what will keep an individual's attention and so and for

(24:17) the war for attention the war for uh clicks that is currently going on on all of these platforms is how they make their money I'm just imagining a an AI system these these AI models supercharging that war for attention such that we now have technology that will allow individual targeting of a kind we have never even imagined before where the AI will know exactly what Sam Altman finds uh attention grabbing will know exactly what Josh Hawley finds attention grabbing will be able to elicit to grab our attention and then

(24:47) elicit responses from us in a way that we have heretofore not even been able to imagine should we be concerned about that for its corporate applications for the monetary applications for the manipulation that that could come from that Mr Altman uh yes we should be concerned about that to be clear uh openai does not we're not off you know we don't have an ad-based business model so we're not trying to build up these profiles of our users we're not we're not trying to get them to use it more actually we'd love

(25:14) it if they use it less because we don't have enough gpus um but I think other companies are already uh and certainly will in the future use AI models to create you know very good ad predictions of what a user will like I think it's already happening in many ways if this Market yes um and perhaps Ms Montgomery will want to to as well I don't but um hyper targeting of advertising is definitely going to come I agree that that's not been open ai's business model um of course now they're working for

(25:45) Microsoft and I don't know what's in in Microsoft's thoughts but we will definitely see it maybe it will be with open source language models I I don't know but the technology there is let's say part way there to being able to do that and we'll certainly get there thank you so we're an Enterprise technology company not consumer Focus so the space isn't one that we necessarily operate in in terms of but these issues are hugely important issues and it's why we've been out ahead in developing the technology

(26:19) that will help to ensure um that you can do things like produce a fact sheet that has the ingredients of what your data is trained on data sheets model cards all those types of things and calling for as I've mentioned today transparency so you know what the algorithm was trained on and then you also know and can manage and monitor continuously over the life cycle of an AI model the behavior and the performance of that model Senator Durbin don't miss the third part and thank you for watching this climber

(26:58) [Music]