Aviv Gaon: Okay, so just I will wait one more minute to everyone to find their places. And we'll start with the next panel, diving deeper into the data discussion. We had a very interesting talks, and a fascinating open to the conference. I'm very excited to start with this next session about data barriers. While building on the fascinating IP panel and subsequent discussions, we are moving onto the next item on our list, data barriers. My name, as my supervisor already presented, is Aviv Gaon, and I'm PhD candidate, and almost done with that title, I hope.

I think we already established the importance of data to AI research. We are developing tools and software which relies on gaining access to data. And not just data, AI and machine learning system need good data. But what is good data? And how we can make sure that we are not using copyrighted data or infringe privacy rights? And maybe we should follow Star Trek Vulcan saying, "The need of the many outweigh the need or privacy of the few." This panel will dive deeper into those questions, starting with explaining what is data, why data is essential for AI development, and is there really good or bad data?

Next, we will explore ways to gain access to data, such as licensing, and learn how Canada and other countries have addressed AI in their legislation. I had the pleasure to work with the Professor D'Agostino recently on IPO's submission to the standing committee on industry, science and technology for the statutory review for the Copyright Act here in Canada, urging for enacting a special exemption for AI data mining. And we are very excited to see what is going to turn out for that discussions.

The order of business for this panel is as follows:

First, I will introduce our distinguished speakers. They will forgive me for keeping their long and accomplished CVs short, giving that we can all access our speakers' bios on our website. And I will use this chance to urge you to do so. Our website provides much more than just a list of bios. You can watch last year videos and we intend to upload today's talks in the coming weeks as well. So stay tuned.

After the presentation, I will open the floor for more questions, and I hope we'll be able to give some interesting comment as well.

The first speaker is Dr. Momin Malik. Momin is a data science post-doctoral fellow at the Berkman Klein Center at Harvard. He is a PhD in computing science. And among other things, he is working on how to understand statistics, machine learning and data science from critical and constructivist perspectives.

Following Momin's presentation, Paul Gagnon will discuss data licensing. Paul is the legal counsel of Element AI, and his practice focusing on data, IP and partnership.

Our last speaker, Dave Greene, you already been introduced, is the Assistant General Counsel Microsoft, which is our leading partner and supporter of this conference. And I wish to join Professor D'Agostino in extending my gratitude for Microsoft's support for this conference. Dave will provide us with a broader view

on AI legislation, initiatives in Canada and other jurisdictions.

So, I'll start with Momin, so please...

Momin Malik: May I have the slides, please? Maybe I can start. This is the first all male panel, so we were doing good until now. We did have another person who was scheduled, but she dropped out. So, unfortunately... Huh? That is back, okay.

Today I'm going to talk to you about what I think AI can do with copyrighted data. The session is Resolving Data Barriers, and the question is what are the end goals that we might want to strive towards.

To tell you a little bit about myself, my background was history of science. And that's still very much where I come from. Thinking critically about the claims and content of scientific knowledge. And I think this gives me a very unique perspective. I worked at the Ben Berkman Center with Dr. Urs Gasser, whose background is information law from Switzerland. And we were working on issues about information equality. And so I also have some knowledge of that literature.

It was very unsatisfying, though, to be reacting to, or what I felt to be reacting to, a lot of the technological developments. I did a Master's at the Oxford Internet Institute, where I also met John. And after that, I went on to do a PhD from the School of Computer Science, along with a Master's in machine learning.

My mission now is to use everything I've learned. And my time in my PhD, I really went into the weeds learning statistics and machine learning from the ground up, trying to build some of these systems myself, understand the theory. But again, with a critical lens, not taking the claims at face value, but trying to go deeper and saying what are the underlying assumptions? What are the implications? What does all this rely on? And now I am back at the Berkman Klein Center now to bring back what I've learned to the communities that I came from. And that's very much what I try to do, is provide a critical perspective that people who are critics, might not know the technology enough to be able to say, and people who know the technology enough may not have that training that I have from a discipline like history of science.

I'll talk about five topics. I'll review some of the things that John brought up and give my own take on it about why data. Also, look at what AI can and can't do. I think drawing these boundaries is very useful. Then I'll look at what data is useful for the things that AI is good at. Conversely, the things that are copyrighted or behind copyrights, what can they be useful for? And lastly, can we use copyrighted data for AI uses without necessarily giving full access to it?

First topic is why data? This is a metaphor I've been playing around with. This is an illustration of the Pepper's Ghost illusion. It goes back to the 1800s. Versions of it have been described even earlier than that. Where you have some sort of transparent film, low light setting, if you shine a light brightly onto something hidden below the stage, below the audience's view, to the audience it looks like

there's a ghostly apparition on stage. And this as been used as in the specter in stagecraft. The recent Tupac and Michael Jackson holograms were high tech versions of fundamentally the same illusion. This is, I think, not a bad way to think about, I'll put it in scare quotes, "AI." Filling out some of the rest of the actors, the shocked man theatrically acting on stage is industry, nobody in this room mind you. I think we don't have anybody from IBM, so I think IBM Watson I could blame as an example of this.

While we in the bewildered public look on in amazement as this ghost appears on stage, what's going on behind the scenes is statistical machinery and data. What this is to say, is that "AI," again in scare quotes, "AI" as an academic discipline has a lot in it. But what's deployed that we see that has big effects in our lives, machine translation, self-driving cars, image recognition. All of this is based fundamentally on applying statistical machinery to data. Whether or not in the limit one's things are seamless, that ghost is actually effectively a real ghost. Or that the limit of all this statistical approaches is actually intelligence. I don't think it is, but there are people who will say that. And so, those are open philosophical questions. But there is no actual intelligence as we might think of it in a colloquial sense. It's all this carefully constructed illusion. And if you just shifted your perspective a little bit, look from the side of the stage, it would all collapse.

I think Dr. [Yanisky-Ravid] could tell you, if you looked at all of the prototypes before that final piece that we got to here, a lot of it is junk. There's a tremendous amount of work and effort that goes into making something that holds its illusion of the machine reproducing human knowledge, human actions, human creativity. And a lot of the effort that goes in is similar to this man underneath the stage, and the whole set-up it makes it seem seamless to us when we're looking as members of the public.

Going beyond this, to what AI can and can't do. I'll break up four tasks broadly. I think if we want to understand a system and what's going on, what exists now under the scare quotes of "AI" is actually terrible. And I'll use the example that John gave about medical systems. There's this great example from Rich Caruana, who is a Microsoft researcher who worked on a medical project in the early 2000s who said, "We strongly found that people with asthma less often get pneumonia." This is a very strong finding. But, of course, that's people because people who have asthma get more consistent medical care, and so are better protected from pneumonia. It's not that asthma has any causal connection to preventing pneumonia. And so, you can build systems that do things very narrowly, but if you're trying to understand what's going on beneath the surface, AI is pretty terrible in terms of the tools that exist today. People are trying. There's probabilistic reasoning. There's causal inference. There's a whole bunch of tools, but that's not what we see deployed in really fantastic working systems and applications.

Conversely, if you're trying to build a system, machine learning and AI are fantastic. I'm putting "predicting behavior" in scare quotes, because prediction means something very specific in statistics and machine learning. It means the outputs of a

model. And what that means in a colloquial sense are post hoc correlations. Sometimes, post hoc correlations do fantastic at making predictions about the future. Circularly, you have these weird papers talking about predicting the future, which is self redundant in a dictionary sense, but makes perfect sense in a technical sense. And there are ways in which predictions can fail if the underlying causal structure changes. Google flu trends is an example of this. H1N1 was something new. It was off the winter seasonal trend for flu, and so the system that Google built of trying to predict flu from search results turned out to work terribly when it was deployed in the real world, because correlations weren't good enough.

Lastly, planning interventions. This is another area where I think you really need to understand the underlying causal structures for which a lot of these correlations that AI can use so fantastically is really not good enough. Econometricians, building on earlier work from statisticians, are talking about how you can predict well without understanding the causal structure. Spurious correlations do work really well, and a whole bunch of other weirder things that, just because you have an AI system that can work, doesn't mean it'll tell you how to intervene and actually change the world.

Next thing is that what data is useful for which tasks? Looking back at these same four tasks, the things that AI is the best at doing, that data is locked behind commercial databases. By which I mean things like Google, things like Facebook. And historically, a lot of things that came out of AI were data mining. They were opportunistic. I have all of this amazing data, what can I actually do with it? And developing approaches about how to extract value from that, extract insights from that. But this is not necessarily locked behind copyrights. My behavioral data from my cell phone records, from my call logs, from my emails, this is not behind copyright. This is just owned by the companies that manage the systems. And so that, I think, is the biggest barrier to somebody like me being able to do interesting things with the data. It's these commercial databases.

That said, on the topic of this session... Oh, as another point, there's this article I came across at the Oxford Internet Institute, Savage and Burrows, 2007, The Coming Crisis of Empirical Sociology. This tremendous fear that was coming up in sociology, that the data to do large scale sociological research was all being held and gathered by private companies. My own dissertation research was about how these data are not as great as we might think because we don't know who's captured in the data. This is also gets back to some of the issues of bias. Geotag tweets are what I did one paper on. They're fantastic. You can get the contours of city blocks, and rivers and roads, and so it seems deceptively powerful. But if you compare that to the census, certain areas are vastly over represented, so you can't really use that to study the population in general. You can only use that to study a narrow segment of it. And if you don't understand how narrow that segment is, you'll probably be wrong about the generalizations you make.

Copyright, on the other hand, as was mentioned before, things are copyrighted because they are valuable, or somebody thinks they're valuable. And so they do form a natural frame. And I think copyright can be really useful for building systems

and for understanding. And I think where it intersects with useful AI applications are on systems, but I also think that there's some understandings that may not come from AI but that still can be from large bodies of copyrighted works. Taking these two themes of building systems and understanding, I'll give a few examples that I see. One is books, another is image recognition and image search, and music recommender systems. Under understanding, I'll talk about news media and I'll talk about case law. I'm not a lawyer. I have been around lawyers for a long time by being at Berkman Klein at two points in my career, but I'll give somewhat of an outsider's perspective on some things that surprised me.

Book search. It's very powerful that, let's say, I came across a book I vaguely remember, The Pawn Copyfight. And I want to use that, and so I enter owners versus users copyfight. In Google Books, I can find Professor D'Agostino's book, and find the exact passage that that comes from. This is incredibly powerful. This is not necessarily an AI modeling type task. This is more information retrieval. Can you find an exact match in a huge database, but there's certainly fuzzy matching that you can do that would fall under AI. This is really powerful, and only Google has the ability to do this. Google Books early on had lots of controversies with copyright. They worked out agreements with publishers, but they're the only ones with those agreements. They're the only ones with the access to the data that they can use to build systems that help me, as an academic, find citations, find exact phrasings that I remember somewhere. If I know something's quoted, I can find where it came from. That's really powerful, and with that access to the actual copyright text, only Google can do that. Other people can't.

Image recognition. There are huge databases of stock photos that have content in them. And if you want to know what's in an image, you need to have access to these photos. And there's really no way around that. If you've ever tried to get stock photos for presentations, the high quality images either you have to pay a lot of money for them, or you get iStock photo watermark that ruins the photo. So a lot of this isn't accessible. And I don't know the vast majority of photographs on the planet, if most of them are stock photos or if most of them are privately held, but there are large bodies of copyrighted photos that would be very useful for a lot of tasks that we don't have access to unless we have a licensing agreement.

Lastly, music recommender systems, things like Spotify, things like Pandora. There's a whole area of music recommender systems, music information retrieval, where you need the actual content of the songs to know what songs are similar to each other by some auditory features, to know people who like certain things in common, what else they might like. Again, you need access to the actual song. And if you don't, then you're not going to be able to develop these systems or do this research. And so, the early movers that can get licensing agreements from large publishing houses, they have a definite advantage that the rest of us may not have access to that, can not have. This particular graph is from a paper on the million song dataset, but things that are released like this depend on the goodwill of companies to release the data that they have. This is copyrighted, so they could file copyright claims for a researcher who collects this.

The other thing that I want to talk about is understanding. My own sense is that while a lot of AI today is a carefully constructed illusion, the underlying statistics is very powerful and does work. There are a lot of ways in which it doesn't work, but the data is very valuable, and the modeling we can do from that data is very valuable. And that will persist whether the labels of AI come and go. Some of the techniques that are underlying AI will stay. One example is a project that I'm on now, Media Cloud, which is a joint effort of the Berkman Klein Center and the Center for Civic Media at the MIT Media Lab. Media Cloud, for the past 10 years, has been scraping online use data and collecting it in a database. With that data that has been collected, they can do things like studying the similarity of words between different media sources.

What's shown on the plot, and this is from the recently published Network Propaganda: Manipulation, Disinformation and Radicalization in American Politics. They're able to use measures of word similarity, and use certain clustering and dimension reduction techniques to show that while you have Fox News grouped in the far right, you actually have two parts of Fox News. You have Fox News itself and then Fox News Insider and Fox Nation, which are far closer to what we think of as the far right media. Then, as a totally different cluster at the bottom of that plot, there's the Daily Stormer and White Nationalist pieces. At the back, you have the National Review and other kind of conservative media that think of themselves as more intellectual and not the sensationalist right wing media.

Of course, similarity and word frequency, as word occurrences, is not a perfect proxy for the underlying meaning, but it is a useful thing to study in concert with what we think of as the underlying meaning. And this is the type of work that we couldn't do had Media Cloud not been scraping online news, because I don't have access to LexisNexis or these other large commercial databases of copyrighted news data. Similarly, we're struggling that we don't have access to video data and to transcripts, which again, are behind copyright or just not accessible to us, which probably is a big part of the US media ecosystem in terms of how people are influenced or what information they're getting.

Studying case law. I am not a lawyer. I was very surprised to find that law schools have to pay a lot of money to these private entities like Westlaw to get access to case law, to judicial decisions, that these are not publicly accessible. The Harvard Law School's Library Innovation Lab has spent the past five years in a massive effort scanning these books of case law, digitizing them, working out agreements with Westlaw who can claim the copyright to a lot of this material, to make this digitally available online in a database. Currently, only two states you can have bulk access to, but we have the data. It's just difficult for us to give access to anybody without getting Westlaw upset.

But with this data, I can do amazing things. And I'm only starting to look into this. I can see are there certain judicial theories that we can test empirically? If they come out false empirically, maybe that just means that the metadata isn't picking up the right signal, but there are a lot of interesting things I can do. Do people in the south of the United States cite large states more? Do they cite neighboring states more?

Questions like this. What sort of influence do we see based on corporate defendants or parties in these law cases? And that I can do at scale, which I couldn't do without access to things that are behind Westlaw's copyright.

The last thing I want to talk about is, can I do all of these things without access to the data itself? Recalling this picture, well, can't I just build and set up the system and have the guy in the sheet walk in at the very end? Not quite. Like any model, this model onscreen is a simplification. A slightly better one might be that the data is the light that we're shining, and that the actual person in the sheet is the labels.

My colleague, Mary Gray, at Microsoft Research, has a forthcoming book called Ghost Work that I'd highly recommend. It's coming out in I think a month or two, with Siddharth Suri, where they talk about automation's last mile. A lot of work that goes into making machine learning work, and AI work, is tedious human labeling and annotation. For things like prose recognition, image recognition, I've done some of this. You have graduate students at Carnegie Mellon who are spending long hours clicking little boxes on pictures to say hey, machine, this is the face, this is the knee, this is the elbow, drawing a line, these are the hands, these are the joints.

Once you have tons and tons of labeled data, then you can start to find correlations and patterns of pixels that do generalize, that can recognize things in pictures that you haven't specifically tagged. But this kind of dirty secret of all of the human effort, really tedious effort, it takes to make these systems that work so magically, so beautifully, is a major part of what makes AI work. And if I don't have access to the full copyrighted data for a lot of the tagging I would want to do, hire researchers to do, hire research assistants to do, I wouldn't be able to do. That's where a lot of the value of machine learning and AI comes from this manual labor, and without the full access to the copyright data, I could not do that.

As a summary, I think of AI, a useful heuristic is the versions that we see commercially is an illusion of statistics. Again, this is not all of what AI is, but this is a useful heuristic. The main body of data valuable for AI, I think is restricted by access, not necessarily copyright. That said, there are some very powerful uses of the data that is behind copyright for AI or for statistical analysis. And it's not possible to fully extract this use, this value, just from building the systems without the data in it. I also didn't mention, but debugging, even just sanity checks as we call them, are really important. You can build a system, but you really need to know how the data move through the system in order to do a good job of building that system. It's not as simple as a total division of labor. Thank you.

Audience: [crosstalk].

Momin Malik: Thank you.

Paul Gagnon: Thanks everyone. Thank you to the organizers for having me on this panel. That was off with a really amazing start. Thanks for that perspective. It's one that's in the industry when you're knee or neck deep in it, depending on the week, you get to

kind of be familiarized with this. From an outside perspective, you're very much seeing the ghost onstage. And I thought that first image a really apt analogy.

My name is Paul Gagnon. I'm one of Element AI's legal counsels. I work out of Montreal and Toronto every so often to support our Toronto office. Our other offices are set up in London, largely working on AI for good projects, and Seoul and Singapore as well. Our company is two and a half years old now. I joined about a year and a half ago, and it's been a whole lot of fun.

The theme is data barriers. Do they exist? Yes, they do. And how to dissect them and how to overcome these barriers is a bit what I'm aiming for in the time that's been allocated.

First off, the initial panels were quite excellent at exposing how IP law applies, or can apply, to different phases of the work behind AI. Is sui generis legislation a good way forward? I think the previous panel made a good job of saying that no, we have a lot of tools to work with.

My legal background is one of a Quebec civilist, so as a civilist the urge to legislate is obviously, at times, the first reflex. But being in Toronto, the old common law tradition of letting things happen is, I don't need to convince the attorneys in the room, that it might be a better way forward.

Copyright, as such, as contributing to these barriers, there's two sides to this coin. The first is that copyright is a barrier preventing that access or limiting it, or limiting how works can be used, has pretty drastic consequences. There's a great article that came out in 2017 from Amanda Levendowski, and similar works, showing that ultimately better data obviously with the old adage that garbage in, garbage out, or AI systems will only be as good as the data that feed them. This work showed that ultimately high barriers and uncertainty around fair use would exacerbate these biases and make for lesser quality AI.

On the other hand, with respect to copyright, is one should not necessarily always assume that works that are treated and dealt with with AI are actually copyright protected. The examples that were drawn up were actually quite useful. When it comes to the music study, all of this metadata, is it inherently part of the work? Is it a body of work from which we can draw as a culture? And obviously, in getting there, music labels and publishers have a vested interest, of course, in publishing music. How much of that extends to the analytics and the data inherent and behind all of that music? I think is one way, if we do assume copyright theory about incentives and monetization, then obviously I think we should challenge this view. If we don't accept the view that copyright exists for monetization and incentivization, which I think is also a defensible view, we get to the same point of challenging whether data can be used or not.

One key point is the dichotomy that's built in, in Canadian copyright law at least, between data and the work. And where do we get this dichotomy from? The definition of compilation under the Copyright Act states that, "The compilation is

either a collection of works or a collection of data." Obviously, data compilation has to itself be original, but let's not overshoot it and say that dichotomy is fundamental. And whether you use the work as a work in and of itself, or if you use a work as data, could be an interesting way forward.

This discussion about creating a separate fair dealing exemption under Canadian law for data mining, we've had the chance to testify last October at the parliamentary committee, arguing that one exemption would be the way forward. But, recent developments in Europe show also that when you do kind of open this Pandora's box of reform, you can end up in a worse place. For example, in Europe, to a large extent there are still concerns that this data mining exemption would only be accessible to research institutions, which would be catastrophic.

So, one of these themes, research or research institution, largely misses the point on how AI research is actually being driven. It's being driven in companies. Obviously, in publicly funded research institutions, but at this beautiful intersection between the two fields.

Availability of data is the barrier, right? Data can be available or not, and definitely we live in the age of big data, but we for sure live in the age of big barriers. The statistics that we can compile about data generation in terms of volume largely miss the point. Bad data that's generated is behind different ecosystems and paywalls that one can not readily work with. And I think Momin did a great job of illustrating that.

Then we have another way of getting data is through data brokers. And data brokers are a largely under regulated industry. It's been around for a long, long time. We know these data brokers, for example, like Nielsen through the TV ratings. And we don't yet see how active this industry is. There is legislative change showing and requiring increased transparency, for example, in Vermont. The data brokers, at least some of them, fell under investigation. And there was an extensive report in 2014 written by the FTC, responsible for applying competition law in the US. And I moonlight as a competition lawyer through my work at the Max Planck Institute, I worked on the interactions between competition law and IP, a fascinating field. But definitely, this report from 2014 from a competition law standpoint, shows that actually there is a market advantage in holding this data. And because there is a market advantage, and because these things are profoundly entrenched, that means that contracts need to be analyzed extra careful and under the realm of competition law. The fact that data's such a key input to AI, only makes that concern stronger.

In the day to day, when you analyze these data licenses from data brokers, I often joke with my teams that the only way to comply with them once they're done, is that we would either lobotomize the team or invent one kind of special device like in Men In Black, to wipe the brains of the people that have been exposed to that data. It's an exaggeration, but these contracts basically sell information as if they were nuclear secrets, and have very strict restrictions about what can and can not be done. A lot of these use restrictions, they're not so much there for any

fundamental reason. They're there to better monetize ulterior use cases. So this is definitely one of those barriers, and how to resolve that is an open question.

I hope that we're at a point of inflection in that information's abundant, our system's created, and yet it's not accessible. So when you have historically these markets that have profound commodification, and yet very few gatekeepers, those are the markets that are ripe for disruption. And so, the example of iStock photos and Getty Images is a good one. What do you do when want to have access to all of these pictures, all of these labels? Well, you can also go to Unsplash, a Montreal based company that's actually open sources and has a very easy licensing model that's free and open and without restriction.

So, open source in the truest sense of the word, with all the necessary ramifications, is a way forward, is a way that can break these data barriers. Break these barriers that kind of prevent, or at least limit, the evolution of AI itself. And prevent or limit is a key thing. If you can say well, no, it's still taking place and it's not a real limitation, well, it's still a riskier proposition because you are still making use of data in ways that either contradicts licensing terms or that is on shaky grounds.

From a copyright standpoint, from a privacy standpoint, as well, the example that came about last week was IBM making public, well, the intention of doing this was couple months in, but the headlines came out last week. IBM built a million picture dataset from Flickr. The aim of that publication was to say well, look, we published this dataset with labels in order to make facial recognition less biased. The price to pay was a perceived violation of privacy. It also goes to show that when you do have openly accessible images, it doesn't necessarily mean that all of the inherent rights to exploit those images are there.

For example, a number of these licensed images, even the Getty ones, will contain trademarks. All of the licensing terms of any commercial provider of these images will tell you that you're not cleared for trademark rights. And you'll have specific images that are cleared from privacy rights and that you can use a wide range of use cases. But that's not a given assumption. So, beyond all of that, what do we do about it?

What we've found is that the real crux of the data roadblock, is that this notion of use is profoundly ambiguous. So, you have to understand that open data or accessible data are two different things. Data can be readily accessible yet not usable, or at least there's a huge gray zone around what use means. This happens all the time for data that's used for AI applications. A trend in the industry is having these competitions. A dataset is released, develop a new algorithm, develop a new model, and test performance on this data. And usually this data's licensed by saying, "This data's licensed only for the purpose of the competition." What happens if the competition is passed, like two years ago, and the data's still available? How do reconcile the original intent, the licensing terms is but a one-line sentence that says, "Only to be used for the competition." And yet it's still available two years later. How do you reconcile that?

You often have a lot of data that's licensed that says, "Academic use only." The notion of academia, does this mean that it's only university research institutions? What about non-university research institutions that are entirely dedicated to research? Is that academic? One would think so. Then there's research use only. Does this mean it can be used within private context within companies? Again, unclear. Hard to reconcile the intent with this notion of use, because use is too broad, at least for AI.

And so what we've done, and the papers, and in true I guess on brand for both Element AI as a company and AI more generally, instead of going the peer review route, we worked on a paper and just made it publicly accessible. With my co-authors, Misha Benjamin, who is my colleague at the legal department, Chris Pal and Negar Rostamzadeh, two researchers associated with Mila, the AI research lab in Montreal. And Yoshua Bengio as well, who co-authorized the article. We wanted to expand this notion of use, create more granularity so that use actually means something. The way we set it up is that there's use of the data itself. What can you do with the data? And then there's use in conjunction with the models. And in doing that, we identified the different use cases and introduced higher granularity on what can and can not be conferred as rights. And in doing that, so the article kind of explains the issues that we found with data licensing.

And we also suggest a new license to work with a new family of licenses. We called it the Montreal Data License, or MDL. And the goal of doing this is much the same as the early work in open source software. Open source software works so well because we know what the licensing terms mean. The standardization of legal terms drove the adoption of open source software. And if it didn't necessarily drive it, it for sure de-risked it because it was much clearer what can and can not be done. We wanted to do the same with data. So, for those that are connected, if you turn to montrealdatalicense.com, beta version of the site. So, if you see any improvements that can be made, let us know. The paper's available there, and there's a license generator there.

So we built this kind of Q&A in which, when you want to release data, you can go through there, pick and choose use cases through the questions that are asked, and generate the license text. And again, license text, we wanted to be consensus based. So if our language is off, if there's improvements to be made, we more than welcome those comments and suggestions. So in doing that, we want to reduce that gap. We want to standardize the use of data licensing because we think that it's an issue, and because we think that it's a barrier that's present. The Berkman Klein Institute tried to do the same recently with licensing language around AI generated artwork, which I found was very great work. Definitely useful. I think there's a ton of these discussions.

So when we do this, we also drive understanding on a technical level of what AI is and isn't. So when you go through our definitions and our use case, that's what we worked collaboratively with the researchers that we worked, to understand what can and can't be done. So, for example, with data we say can I create a

representation? This is a technical term. A representation of a data is a technical term. Whether you want to grant that right or not is based on your understanding of it, but to educate and to bring forward the discourse on that we saw as a natural first step.

So these barriers exist, and it's pretty much up to us to resolve them and to kind of challenge the notion that this age of big data means that there's an abundance of which to work with. And ultimately, the risks that are tackled are those that our researchers faced, that our nascent AI industry has faced. And those are risks that actually favor more unscrupulous actors, or actors that have deeper pockets to litigate. When you're a burgeoning company, risking these, building fundamental products with data that you weren't certain you could use, it's a risky bet and it's one that we think that increased standardization of data licensing language can help resolve.

A quick historical note at the end, the evolution of AI is decades in the making. But the progress of AI research faced two AI winters. AI winters that we qualify as an absence of funding and very, very slow momentum. We're not in an AI winter yet, but we may be. And one of the reasons of, at least partially, the two first winters was that the hype was so big, that fundamental disappointment led funding to disappear. And so, an exercise like today I think is hugely important in making sure that the hype is connected to what's going on.

There's a lot of really promising applications being developed, lots of promising products that are available, but we have to challenge those notions and make sure that that cynicism, that expectation, that we've actually truly challenged what's limiting progress, like data barriers, and what can really facilitate more democratic use of AI. And largely, that those capabilities that are now present are used by many. So, hopefully, this third AI winter will never come today. And if we brace for impact, we should also brace for the impact of over hype and of having legal systems and how they interact with AI be a cause of that AI winter.

So, thanks again to the organizers.

Dave Green: You might remember me from such panels as IP at a Crossroads, apologies for the re-use of clip art. By the way, Dr. Malik, what a wonderful and then just clear and clarifying use of PowerPoint, in addition to the terrific information that's there as a PowerPoint user.

I love this data topic. I've spent the last two years traveling around the world, primarily in Europe, educating regulators, educating lawmakers, educating trade groups about the importance of data use. And really, thinking about the democratization of data, and I think where that's culminated and why I titled the talk sort of the Right to Research, is understanding fundamentally what the challenge is. The challenge is really about how do we, as a society, and certainly how do we as researchers, how do we as private entities, how do we as governments access and utilize data in a scalable way, in a way that benefits and does not burden society? But also, in a balanced way, in a way that respects

traditional notions of intellectual properties, respects the right and freedom to contract, and respects obviously the impact of this wonderful technology on the users.

And I think the IBM example is a perfect example of a really wonderful intent, the notion of releasing a set of publicly distributed and publicly accessible data, to try to reduce the challenge of data bias. And doing so in a transparent way, and yet copyright and rights of privacy and other issues are sort of raised as potential friction. I think it's a perfect example of how we have to carefully walk through, break those issues down and figure out what is it that we're trying to solve for, and what is the regime and role of copyright, for example, or intellectual property, and resolving those issues.

I think you've heard really clearly from Paul and from Dr. Malik about the sort of the right, the need for machines to learn. And I want to use a couple examples. This is a photograph I could have easily taken with my cell phone. I didn't actually, if you'll see from the attribution there, that's an image that was obtained via creative comments license. Would anyone here argue that this image is not, just a show of hands, this image is not subject to copyright protection? Okay, we got one hand up, maybe two. I mean, again, as a photographer myself, and as a photographic lawyer for many years, I might look at this and say it's a factual image. If you're an ardent photographer, you'd say no, no, no, it's a contrast of new and old. You see an old bus there, there's sort of new technology. There's arguments. And if I took a copy of this without permission, without a license, and I used it for purposes of this presentation, I think fair dealing might give me an excuse.

If I used it in a class to educate photographers about how to compose a particular image, or how to juxtapose, fair dealing may or may not apply, or fair use may not apply. If I used that image and distributed it freely without really any protections or restrictions, whether I did so for profit or not, I think we'd now get into some grayer areas. And clearly, if I ran an ad campaign on come see the sights of London, I think none of us would argue that there's a real potential claim for infringement here. So I want you to park that thought for just a moment with this particular photograph. I'm going to come back to that.

If you're an engineer, we've used the example of self-driving cars, you have a really unique challenge. Your fundamental challenge as an autonomous vehicle developer is to ensure that your passengers get to their intended destination without causing harm to themselves or the world around them. And so, as a car travels autonomously down the street, it's got a number of decisions and choices and predictions that it needs to make. What is that object on the right? Can we make a prediction about what that object is likely to do as we get close to what appears to be an intersection? There's some other large objects in the way. When things go wrong, obviously it's not just simply catastrophic. People get injured, people die. But there's also a need for transparency and responsibility. Why did the machine make those predictions? Why did it behave that way? How do we correct that? And do we attach liability at what stage to those decisions. These are fundamental decisions.

So, to get to a point in autonomous vehicle safety that we trust ourselves, and we trust the system, to be able to enter into these vehicles. I used Uber to get here earlier today, and it was just sort of pondering as I was thinking about here. I'm using my phone to jump into a car with a complete perfect stranger, who I trust is going to get me here. And that's the task I effectively I'm trying to delegate in an autonomous world to a machine, to an algorithm. How do we do that? Well, we take images around the world. There was an example of images in Getty and Corbis. I would argue those are horrible images, and I'll tell you why. Because those images are an edited curated set. They're chosen not necessarily for their inherent value as data. They're chosen for their aesthetic capability, or for their factual reporting of a particular event. They may or may not be labeled, but they're traditionally not labeled in a way that's useful to machines.

And so, as I think both speakers, Paul and Dr. Malik, pointed out, there's a tremendous amount of work that would have to be done to this image to make it useful for machines. But fundamentally, what we're doing with this particular image is very different in our "use," to use Paul's words in finger quotes, than in the first example of that image in a traditional context, being used in a copyrighted sense. The challenge, of course, is to use that image for a machine a couple things need to occur. That image needs to be reproduced, needs to be modified and labeled. For transparency and for safety and for just simply going back and retraining machines when things go wrong, it needs to be stored and preserved. So these are the fundamental things that implicate copyright law.

I think we would argue, and if anyone disagrees I'm happy to see a show of hands, that understanding the information in a copyrighted work is not itself infringing, even if the activities associated with that understanding, that learning, implicate what are traditionally copyright rights of reproduction or distribution. Does anyone disagree with that concept? So, are we really talking about copyright restrictions and friction, or are we talking about other issues such as rights of access, right of contract. Well, that's the fundamental challenge, is viewing copyrighted works when they're used as a work, and we can apply traditional copyright norms, and we know very well with precedent and law what that outcome is. It's a little more challenging in a code based or an exception based regime like Canada has, as opposed to a regime like in the US, or the proposed regimes in Europe, and the actual regimes in Canada, where you don't have these codified concepts where you're looking and they're construed fairly narrowly.

In the US, obviously, the concept of fair use can be very broad. You're looking at, at least, four factors that as case law in the US has developed, it's very clear that the commercial factor, the impact of the use of that work on the market for the original is becoming less and less important, where it was significantly a dominant factor, versus the use's data. If you read that Harry Potter book yourself, I think copyright would not restrict you from learning and absorbing and applying the knowledge from your exposure of that work to your normal tasks and routines.

So, if you're a data scientist and you're going about your normal tasks, whether

you're a researcher or whether you're a commercial entity trying to develop next generation technology, you have some fundamental questions. Can I access this data? Do I need a license to use it? Does the license adequately permit my use? And what if my use produces a commercial benefit? Can I let others use or see or access this particular data? Can I share this data publicly? Do I have to attribute my sources, and what license should I use? And what about the ethical implications, as we understood with the most recent example?

Now imagine doing this at the scale that it takes for autonomous vehicles, in our example, to get to a level where we can trust them. My research teams over at Microsoft Research have told me it takes approximately 10,000 images to train an algorithm to just simply do some really basic recognition. That this is an object. That this object is this versus that. It takes about a million images for that algorithm to approach a level where it's actually coherent. For example, that it can identify a table or a setting. And it would probably take several tens of millions of images before, I think, all of us would trust that algorithm to make predictive decisions that could impact our literal safety and lives.

So to try to answer those questions at scale, really it's a huge challenge. And I think, from Microsoft's perspective, copyright should not be a barrier. It should not pose an obstacle. And the friction, if there is any friction, whether it's implicated by an ambiguity in law or expressed because the law as it exists does not permit these activities, or limits those activities to a defined group, we perceive that as friction. And this is, I think, where we spent a fair amount of time kind of overseas really trying to make those distinctions between the use of copyrighted works as works versus the use of them as data.

So, how have we done? What's the reaction across the world? What's interesting is, I think if you correlate these examples with where the investment and where the energy and where governments are contributing to the development of AI, I think you'll find a strong correlation. So, for example, the United States has a fair use regime. It's quite a bit of robust technology and activity. That fair use regime, thanks to cases like Google Books and Hafele Trust and TVIs, and a myriad of other cases, have really paved the way for an understanding of what non-commercial and commercial researchers can do with copyrighted works.

In Japan, Japan has made leaps and bounds in changing and broadening an already decent exception to ensure that not only can these works be utilized regardless of the nature of the entity using them, or the user, but that they can be aggregated and stored. Because I think the government regulators there understand the importance of maintaining transparency and accessibility. Australia is developing an exception. Singapore had gone through a review, and has just proposed an exception that's very similar to, in concept, to the Japanese exception.

And then there's Europe. Next week we find out. I think next Monday afternoon, the European Parliament is set to vote on the entire copyright package. And that copyright package is very controversial, because it addresses things other than text and data mining. But let's look at sort of what the European approach was.

Originally, the approach as Paul mentioned, was limited really to uses and users. It was public interest research performed by non-commercial research institutions. For those of you who have ever worked with academia, the concept of non-commercial research is an illusion. Because all research, particularly research that holds promise will have some commercial component to it. It will either be a joint sponsorship or joint research, or universities are charged typically by their tech transfer offices with transferring. That's the purpose of that research, to transfer it into the community where it may actually be implemented commercially.

So, with a lot of discussions with certainly a lot of looks to the centers and places around the country where AI is being not only incentivized through government support, through government contribution, but also through helpful legislation, Europe did an about face on 3A. They didn't for the other articles, but they did it on this particular article. And it was fascinating. They created a very broad exception. And I just want to talk about 3 and 3A. 3 applies to a smaller category of what's considered non-commercial or public interest research, and with the European council, and then now before the European Parliament, said was there should be no restrictions under copyright. And there should be no restrictions that are imposed as a consequence of contract when researchers access those particular works. That was a very broad pronouncement that for at least a subset of research, they didn't want contract intruding on a right to learn. And so, I would look at 3 as a not perfect, but simply arguably the Europeans' expression of a right to research.

Article 3A was a nod to, I think, the concerns that were raised. It's not perfect. It lets copyright owners withdraw their works. They have to do so in a machine readable way, and they have to express that in an unambiguous way, that their works are reserved and can't be utilized for TDM. So typically, that would be done via paywall or via some contract in which acceptance occurred and perhaps consideration was a portion of that. It's a good step. It's a step in the right direction. And I think as Chairman Ansip recognized, it's fundamental for Europe's AI goals.

Well, now that we've got sort of a momentum towards a right to learn, and the recognition that that right to learn needs to be broad. It needs to apply to all users and uses. Query, where does Canada go? And Canada's obviously spending a fair amount of time debating that particular question now, trying to figure out do they set the right balance? Do they approach it from a European perspective? Do they approach it from a Japanese or a Singaporean perspective? Do they apply it as a broader fair use standard, as the US would, or a limited exception? These are fundamental questions. And I think there'll be a healthy debate on that, and hopefully by this presentation you clearly know where we stand.

This is critical for Microsoft, this right to research. Because ultimately, and I think the speakers did a really good job, it's about not just access to this kind of material, but it's about access in a scalable way. Because if researchers still have to answer licensing questions on a one by one by one basis, that scalability it really does impact and restrict them. Very few companies have the resources and can perform either that level of review or take the legal chance, the legal risk, that they're going

to do it anyways, and then address the legalities of it in litigation. And more fundamentally, what we're talking about are not lawyers. We're talking about developers. We're talking about researchers. And so there's a groundswell around the world, and I think you heard Paul use some wonderful examples of entities that are trying to create a methodology, a mechanism, for making that data publicly accessible on platforms in a machine addressable way.

Here's some examples, I think, beyond the ones that you've heard, of attempts to resolve this data license friction. And this friction sort of stems from a couple of areas. Part of it is what standard do we use? Do we use an open data standard? Do we use a proprietary standard, and what are the appropriate standards that should apply? What is the license that we should use? And I think you've seen examples of the open database license, examples of open source type licenses. Fundamentally, my problem with those licenses is they're all founded on the concept of copyright. And I think from my discussion here, I reject the notion that the uses that we're talking about implicate copyright. And to the extent that they do, that friction and that limitation that's imposed by copyright should be relatively minimal.

But, copyright still is a barrier. I think a lot of the examples that Paul raised about non-commercial use or academic use only, I think when you probe the intent behind those as we do on a regular basis with our researchers, what you find is there's not an intent to control or restrict, particularly with the examples on Kaggle and others. What you find is that people actually don't know what they have. They're uncertain about their rights and their ability to distribute. So they believe if they apply a non-commercial or an academic only restriction to this data, because they don't own the underlying components of the data. These are just aggregations of typically publicly accessible data. The most that they would own would be any labeling or contributions that they added. And there's no intent to control or restrict those because they're not marketed or licensed in a commercial way. That really what they're doing is they're just trying to eliminate uncertainty by controlling use.

And look, from an open data and right to research perspective, that's a fear based approach, and that's really not an approach that's scalable or that works, particularly if you're trying to build platforms or centers where this data can be distributed. If you're a government entity, and most governments around the world are charged with making their data accessible. There's a number of open data provisions and open data licenses. How do you distribute this material? That's the fundamental decision. And I think one of the challenges that all of these approaches pose, is that they haven't thought about this in a schema or sort of a machine readable way.

So remember, a lot of this data mining, in order for it to really take place at scale, is not going to be done by humans. Humans don't read terms of service. Excuse me, bots don't read terms of service, humans do. And so I think as we look at this from both a copyright perspective, and as we think about distribution of aggregated publicly accessible material, making that available in a democratic way that any developer, that any government institution, that any private or public researcher

can access, we have to think fundamentally about how we achieve this at scale. And we're not there yet. We have a fair amount of work, I think, that's left to be done. I'm actually confident. Two years ago, I was very not particularly optimistic about the legal regimes. And in a very short period of time, I'm seeing those legal regimes change to recognize the friction that they need to eliminate.

Now, I think the challenge for Canada obviously is to address that copyright friction, but really to then for governments and entities that are distributing data, to do so in a way that truly makes that data accessible and allows us to achieve this promise of a right to research. And that's it. Thank you.

Aviv Gaon: Well, thank you speakers. I think it was very interesting panel. And I think we have time for some questions, so if any of you wish to ask our wonderful panelist a question, I think that's... yeah.

Dave Green: So, I'll pose a question, because it was alluded to by Paul.

Momin Malik: There was question in the back.

Dave Green: Oh, please...

Speaker 6: So, it's on? Okay, so I agree. It was really interesting, but I was wondering first, can we compare the AI machine system to people? Just absorb a lot of copyright works before they create some work of art and it's kind of like implemented in their mind, so is it the same use? That's on the one hand. But, on the other hand, if we speak about data and data should be accessible for all, so would firms, like the one you, yeah the last speaker, Dave, represent, would open all the data that is being used to everyone? Or is it kind of like two sides of the coin? It's we want to use the data, but we don't want to expose it to everyone for future use because that's our commercial trade secret?

Paul Gagnon: So, thank you for the question. To quote an analogy that was made by a civilist attorney that said, "Laws are like sausage. If you like laws, never see them made." I kind of apply the same to AI systems and products. It's messy, and it takes a lot of time, and a lot of iteration to get to something that has a relative capability to do things. To take new input and to adapt and give output that's not hard coded in. I don't think that's synonymous to automation. I don't think that's synonymous to intent or minds. And I think it actually shows that, especially in the field of art, I think that those are the best ways to engage with the discussion. Because art is a reflection of so many things in society. When Jackson Pollock started painting, half the planet wouldn't have said that was art.

Automated approaches to art are often disregarded. The example of this work of art that was sold for over $400,000, actually that's a good example of an underlying dynamic here. The people who sold it, didn't even develop the algorithm. The solution was made available under an open source license. And, if anything, the only legal issue there is that the people that sold it did not credit the use of algorithm in the right way.

When you look at different songs that are, so-called, AI generated, I think the best one for me is the new Beatles song called Daddy's Car. If you haven't heard it, it's worth a listen. It's the best teacher of all the cliches of the Beatles, but it actually what I'm curious about is how that song was made in the first place. And it illustrates issues around representivity of data and so forth. Like, for example, who your favorite Beatle is would probably skew you into selecting different songs to train from, right? You look at George Harrison's contributions, any self-respecting top 10 of Beatles songs will have at least four George Harrison songs. And we can talk about which ones later at cocktail.

But statistically, a very low impact of George Harrison's work made it through the huge corpus of the Beatles. And yet, our impression of it is immense. And we know that a lot of these songs are Beatles songs, despite the fact that statistically speaking George Harrison's contributions were not that significant statistically. So it actually shows you this question of bias in data, right? So this question, Daddy's Car, how was it made? These AI generating tools, especially in art, are actually a new canvas, a new re-mix, a new way to re-appropriate art that exists. So instead of remixing a song, you remix a whole body of work. And in that, there is creativity. And in that, there's no automated approach that there is automation possible. But that doesn't exclude the fact that there's ways that humans can use these as tools.

Dave Green: If I understood a portion of your question properly, and apologies, I think you might fundamentally be asking about use. And does it matter whether the use in one context is permissive versus the non-permissive, and how do you distinguish the implications of copyright? I'm changing my thinking on this. My thinking now is that you don't actually focus on the use. If you think about machine learning, and you think about the techniques that are applied, and the reproductions and uses that occur kind of along that linear timeline, really copyright issues might come into play at the tail end of that process. And so, you might have a regime that doesn't punish the inputs. It doesn't punish the intake of content and the utilization of that content to do the things that we have humans been doing for centuries. Reading things and utilizing that information to perform tasks, and then some of those tasks and then create potential issues.

You really focus on the outputs. And you say if the output of a copyrighted work implicates or harms the kinds of traditional things that we think about from a copyright perspective, that's really where you're focused. And what's interesting is, I think you're seeing that in precedent. The recent TVI's case is a really good example of where TVI scanned a number of broadcast and radio news and traditional news. They created an analytical engine that allowed folks to track how a story was trending, or the sentiment around a particular keyword query, or what have you. And then, to make that content relevant or useful, they allowed snippets, if you will, or components of it, to be made available so that the user could place the information in the context of the article.

That's really where the court was very comfortable with the search, indexing, machine learning, etc., sentiment analysis. And where the court was uncomfortable

in those cases was at the other end where the distribution of 10 minutes of a news clip, I'd argue that's a documentary not a news clip, the way news is typically distributed, that was starting to intrude. And so that was a case that focused really on the outputs, and wasn't so concerned, and the parties ultimately dropped litigation, around the inputs. And maybe that's the way to think about that particular issue.

Aviv Gaon: And, yeah [inaudible].

Speaker 7: Okay. Let me ask the question that the panel might not have addressed. So I'm a scientist. In my lab, I collect data from our participants. And over the years, and I get consent to use the data to do our research, right? So now, as I understand it, EU rule right now is if you want to use that data for a new purpose, you actually have to track the participants down and re-consent. So I don't know, what do you think of this and how you believe this should be applied to the Canadian or North American context?

Paul Gagnon: So that has made... There's a layer more in the university context, which is the ethics board approval as well, that is also addressing how the data can and can't be used. So the regulation actually that you refer to is GDPR, the General Data Protection and Regulation, shows that there's three really realms with respect to data. There's privacy, as Carol pointed out in her intervention this morning. There's privacy. There's the copyright angle. And then, later downstream I alluded to it, the competition angle, which is when data as an economic asset becomes something else.

From the privacy standpoint, a lot of this actually largely already exists in terms of re-affirming consent or consent in new uses. I think that those standards are increasing. The question is how does one revoke it later. We see consent as like these punctual checkpoints, which is what GDPR has done as well. The area of research that's emerging is about control. So instead of these punctual consents, how do you make sure that you have meaningful knowledge of how it's currently being used, and can you revoke it, can you modulate that consent? So, no clear answers. It's an important component. And I think that the privacy protections are definitely trending in the same direction.

Aviv Gaon: [inaudible].

Momin Malik: It definitely makes things much more troublesome. And I think we tend to think that it's not a big deal. I mean, what are we going to do with the data that's so different from what people already gave consent for in the first place? I mean, they came into a lab. They were compensated for their time. I don't know where I stand on that because I would trust many people to not do anything objectionable, but I don't trust everybody to not do something objectionable. And I don't know what I could think of that would be objectionable to the people that originally consented to have their data collected, but I'm sure there could be things. And so, this is one solution. I mean, it's going to definitely make a lot of work harder. Maybe that's an acceptable trade off, but I don't have a good sense of that as somebody who's

based a lot of my research on the re-use of data, that I maybe couldn't have done any more, doing work that I think is in the public interest.

Dave Green: So, speaking from a platform perspective, I would say platforms have a tremendous social responsibility. The challenge, I think, is letting platforms perform that function of deciding what is good and what is not good. I'd say that shouldn't stop them from doing so, and they should do so in an aggregated collective way, because this is much more difficult than any one particular platform can solve for. And I think Microsoft and Google and others have recognized this early on. Our ethicists and our researchers clearly see the good, the bad and the ugly with respect to this technology. And I think they're very concerned that it be done in a thoughtful way so that the beneficial uses don't encounter friction from some of the fear around it.

There are times, and many times, when it's more appropriate to have governments regulate in the space because there's an important public debate that needs to occur and important interests that kind of need to be heard. It is not an easy answer, but I can tell you from a platform perspective it's not something we can ignore and it's not something we should ignore. And we're embracing a number of different approaches, including voluntary initiatives, principles, as well as calls for government regulation. Brad Smith, our General Counsel, recently addressed and made a call to have governments weigh in on the important concept of facial recognition. So, good question.

Aviv Gaon: Thank you. And Carol, yeah...

Carol: So thank you very much. Thank you to the panel for addressing some of the barriers, and the use of copyright as a potential barrier to data access. My question is specific. Paul, you said something, you were talking about the fact that in the Canadian context in particular, a sui generis exemption under the Copyright Act is probably not required. Rather, we already have the sort of legislative infrastructure in our existing laws to be able to data mine for the purposes of AI creations, essentially, or algorithm training. And my question there is, do you have concerns that the current interpretation under much case law, with the use of that data in a commercial context is often seen as sort of unfair. And are you concerned then that maybe the sui generis exemption not being in place, will in fact down the line prohibit that data mining?

Paul Gagnon: So, let me clarify. The sui generis statement I made was, I think, more for AI at large. It's such a complex field that there's so many different bits and pieces of regulation and bodies of law that I think we should draw from to work with. From a copyright standpoint, we did argue for the fair dealing exemption on text and data mining. Definitely, better a clear exemption that's not based on the identity of the entities doing it. If that was the case, then we'd argue for no exemption. But for sure, that clarification we actually much welcome it, alongside the positions that Dave, I think, explained that are pretty much reflective of where we are.

Carol: And if I can just follow up quickly on that, though? Then in the interpretation of the

fair dealing exemption, is there a concern that, again, in the commercial context or the commercial use of that data, or the ultimate creation from the use of that data, that it will be deemed unfair?

Paul Gagnon: Well, we do want the exemption to be within the confines of fair dealing as well, right? What's a little trickier is tying it to the use cases that are accepted. So, having a purpose limited fair dealing exemption is the issue. Not so much that the dealing is fair. We think the dealing is fair. Where it's uncertain is how it ties into the existing exemptions. And that's the clarification that we think is useful, not so much whether it's fair or not. Because, for example, the impact on the market for the work, though as Dave mentioned, I won't say fading but with variable interpretations under US courts, I think that in Canadian courts that would still be an important one to consider. So it you're scraping a database of images to resell and monetize the database of images, and you've done that using Getty images, well, Getty's main market is selling images, so that wouldn't necessarily qualify as fair dealing. Though the purpose of it would be text and data mining. I think fair dealing does bring that granularity that we can still ensure that the overall use is fair.

Dave Green: So, the very fact that we're having this conversation and that we're still debating this issue, can telegraph where our perspective is. It may be that fair dealing is, under Canadian jurisprudence, is appropriate and that courts can apply and look to other jurisdictions and apply a precedent that can incentivize. The world in AI is moving at an incredibly rapid pace. Investment is occurring at a rapid pace. Governments are infusing capital and pushing resources. I think everyone sees AI as table stakes for a next generation digital economy. And so, I think as governments look at this issue and decide how to approach it, there's sort of the legal approach which is, is the law appropriate and adequate to deal with these kinds of situations? And if not, how do we alter it but not alter it too much? And then there's a policy approach, which is what is the Canadian government and the Canadian public want to incentivize in terms of how they would approach a law?

And so, I think when you look at countries like Singapore and Japan, they're very much from the incentive perspective. They want to claim, they want to put a flagpole and say we are serious about artificial intelligence. We want investment to occur here, etc., and so they go for an exception based approach, and then draw the limits of that. Europe, I think, is probably more in the middle, which is they clearly want to compete. They want table stakes, but you can see in the text of that legislation just some concern and worry about, and then a general European perspective of really disfavoring exceptions, and wanting to construe them narrowly.

So, Canadian government and lawyers and lobbyists and folks looking at this issue, now have a pretty clear roadmap of how possible the different avenues to approach. And I think that's the way they need to think about it. Do they want to look at it as an incentive based activity and pave the way and clarify it, or are they more concerned or want to take a more measured or careful or cautious approach with the implications, positive and negative, that that carries?

Aviv Gaon: So I just want to add very short points here. First, I share optimism about Canadian policy. And I don't know if many of you read the recent budget report, but it seems that at least in the government they do understand the problems and difficulties we addressed here. At least for expecting something like soft regulation. And considering your question, Carol, I think that I'm more inclined to Dave approach. I think that's a problem that we have now, is that there are some attempts, in Canada at least, to take this AI initiatives and open up for more like a fair use exemption here in Canada, which I think is not very good and very useful.

And I think that, given the complexities of this issues, I think it would be better at least to try to come up with some specific limitation or exception for AI and data mining. And I think we definitely can do it, and should do that, and not go through this fair use. Because I think the problem with fair use that, or fair dealing for user, for trying to get fair dealing to be more like the version, the fair use version in the States. And I think the problem we have that, that we give too much power to the courts to decide, or else what is this AI? Is it something we can regard as an infringement. And I think this is something that we at least try to avoid. That's my take on that. And I think we have time for one more question, or... Yeah?

Paul Gagnon: Sure, great.

Aviv Gaon: So...

Paul Gagnon: Shall I go? All right. First, I wanted to say thanks. So, John, from the first panel, thank you so much for an excellent discussion. Really interesting and Momin, great to see you again. Always good to see a fellow Berkman alumni. Maybe my question is actually a very quick follow up, I think, from the prior comments and Carol's question. I'm really interested, David, in this notion of the right to research and learn. And because one of the themes that have come up in a lot of the panels and in this panel, has been this notion of inequality of access. And some of the friction created by copyright leads to biases and can lead to barriers to innovation.

And I completely agree with Aviv, this notion that if we just use limited exceptions. If you want to talk about fair dealing or fair use, there's a lot of uncertainty in that. And that leads to courts making a lot of decisions, but for a sort of smaller market player doing this kind of AI work at scale can be taking on a lot of risk. And so I'm wondering, maybe asking David, but anyone can comment on it, what would you envision like an ideal version of the right to research, or if it were to be implemented in a jurisdiction like Canada? Would it be an exception to copyright, like fair dealing or some kind of limited exception? Or, what would an ideal version look like?

Dave Green: So, rather than talk about the implementation of it, whether you would approach it from an exception perspective or whether you would expand on a fair dealing and carve out a real clear safe space for that. I mean, there's various approaches. I'd rather approach it as what should the right of research include? What are the contributions that should be, however the implementation occurs? Because look,

asking and American on how to implement Canadian regulation, is probably a non-starter. I kind of look to Canadians to determine what the right course of action is. I'm certainly not a scholar on Canadian jurisprudence or, for that matter, Canadian copyright law, whether I've read and know about it.

But I think from a fundamental right of research, which to me is more of a global concept. I mean, you could take that right of research, establish a set of principles, and have those principles, however they're implemented, applied globally. Because research in the 21st century is not an isolated instance. It occurs on a global basis. I think there's certainly a responsibility component of that. And we've talked a little bit about that. I think fundamentally there certainly is a right of access, and I would say at a bare minimum that publicly, lawfully acquired or lawfully accessible public material should be available without restriction, contractual restriction included.

I wonder, and I think from a platform perspective, I'm conscious we're a platform, we're a research institution. We have one of the largest R&D components in the world for a tech company on R&D, and so it's important to have this right of research. But we also have important databases and important materials. I want to disclose our biases. And so we've thought and talked a lot about those folks. Actually, one of our business units is involved in litigation around access from a data scraper, LinkedIn, and that poses important privacy considerations as well as IP considerations. But I do think at a bare minimum, publicly accessible material that is unconstrained, is not behind a paywall, does not subject to measures that evidence a clear intent to make it unavailable, should not be burdened by copyright law.

I would almost go further and say for a fundamental right of research to exist, I think I agree with my colleagues on the panel, that we need to be really careful about what those contractual prohibitions can permit or not permit. And I would certainly be comfortable with a prohibition on contractual restrictions for material that would otherwise be lawfully accessible. I think there is a real concern, and I've heard the concerns about Lexis and Westlaw, etc. They do provide value, and they do provide some insights. And the notion that folks can go and mine that material without burden, I think is concerning. And maybe the way to tackle that, obviously, could be some governments in this to make those laws and that case material publicly accessible.

But, I also think that, from a competition perspective, you do have to be worried about imposing contractual terms, certainly on a category of research that would not compete with their business models. I can probably outline others, and that's probably worth another topic. I think as you think of these principles about what does a real right to research mean? How do we make that data accessible? The only other thing I would add is, having gotten that data and information, that information should be able to be freely shared and freely accessed on platforms if the aggregator so chooses to do so, under the right contractual provisions, to make sure that it doesn't interfere with a legitimate interest of those copyright owners. But the right of research is useless without access by a number of entities, on a number of platforms, to that kind of material.

Paul Gagnon: A courageous proposition to offer a follow up when it's the only thing standing between us and lunch. So, politics is the art of the possible. So the fair dealing exemption is what we saw to be what was possible. If you ask me what I prefer, I prefer non-purpose driven US style fair use. It's still a challenge, decades now into trips and WTO framework, as to whether that US style fair use exemption even is legal under WTO law. So obviously, you don't necessarily want to poke the bear on these unresolved questions and aim for a super blue sky exemption. So that's kind of where this more targeted approach comes from.

And then to the point about monetizing public content, which is what it is. When you put court decisions behind paywalls, and in Canada we've had the benefit of CCH decision and whatnot, that actually we're discussing this. But the underlying question is how much of that is actually public information in the first place, and what can we do with it? Crown Copyright comes into consideration in Canada, creates some weird issues. But fundamentally, this notion of public information is a bit even broader than just court decisions or government documentation. That's pretty clear slam dunk.

In the US, this is heavily monetized. The SEC API is deliberately throttled, so that you're brought on to other pay for play service providers. To which, I spent some time at the SEC and revolved back into business, and then back again. One could argue, and I think hopefully we might touch upon it this afternoon in the smart city debate, what about Uber's data? The Uber API says you can't build a competitive service. Well, why not? We pay for the roads. Shouldn't we be able to pay for public transit data, subject to privacy limitations? Why not? This is all publicly used infrastructure. Is that not public data as well?

In any event, accessibility is one thing. You have to compete on computing power. You have to compete on hardware. You have to compete on scale. And that is where ultimately open data can also be a false flag. Oh, here's open data, but knowing somewhat cynically that no one can compete in the first place. So the real question is how do you build a competitive economy that's data driven? And copyright is definitely a part of it, but there's a lot more to touch on, for sure. Thanks.

Aviv Gaon: Okay, thank you. Please join me in applauding this fascinating panel.