Consumer Finance Monitor Podcast (Season 9, Episode 25): AI Liability Comes Into Focus: A Conversation with Mark Geistfeld on the ALI's Civil Liability Principles Project

Speakers: Alan Kaplinsky and Mark Geistfeld

Alan Kaplinsky:

Welcome to our award-winning Consumer Finance Monitor Podcast, where we explore important new developments in the world of consumer financial services and what they mean for your business, your customers, and the industry. This is a weekly podcast show brought to you by the Consumer Financial Services Group at the Ballard Spahr Law Firm. I'm your host, Alan Kaplinsky, the founder and for 25 years the practice group leader of the Consumer Financial Services Group and now senior counsel of that group. I'm pleased as I am always pleased to be hosting and moderating today's program.

For those of you who want more information either about the subject that we're going to cover today or for that matter, anything else in the world of consumer finance, don't forget about our blog, Consumer Finance Monitor. We've hosted our blog since 2011, and so there is a lot of relevant industry content there.

We've also regularly hosted webinars on subjects of interest to those in the consumer finance industry. So to subscribe to our blog or to get on the list for our webinars, please visit us at ballardspahr.com. And if you like our podcast, please let us know about it. Our podcast is available on all the major platforms. And please let us know if you have other ideas for topics that we ought to cover, or the speakers that we should consider inviting to our program.

So let me now tell you a little bit about what we're going to be talking about today. As regular listeners to our podcast show know, we've been very closely following artificial intelligence, or we'll call it AI from now on, on all our digital media, our blog, our podcast show, and the webinars. And that's because of not only the enormous impact that they're going to have on the world, but because of the important impact they're going to have on the consumer finance industry.

And indeed, I think we really began getting, really following AI in depth beginning in August 2023 when we released a podcast that was entitled A Look at the Growing Use of AI, Generative Artificial Intelligence and Chatbots in the Consumer Finance Industry. And then we've gone on from there. And as those of you who are following AI know, the White House came out with a White House action plan. We did a webinar on that and we were very fortunate to have as our guest, Dean Ball, who was one of the architects of that plan.

And then much more recently, we released on April 30 a podcast show entitled The White House AI Framework: Ambition, Preemption and Uncertainty Ahead. And we've done many other programs as well. Just if you're interested in AI-related programs that we've done, I invite you to go on our blog and just do a search for that and you'll find everything we've been doing.

A little bit later this year, I'm going to be doing a podcast show that's somewhat related to the show that we're doing today, but I think it's really complimentary to what we're doing. I'm going to be interviewing Professor Dave Hoffman from University of Pennsylvania Law School about an article that he recently co-authored with the CEO of the American Arbitration Association entitled Agentic Commerce Needs a Legal Infrastructure.

And though this is, what I'm adding isn't really in the title, it should be. And if we're not careful, the courts are going to be coming whether there's a legal infrastructure or not. So that's going to focus on contractual frameworks and how they might be applied to agentic AI or agent-driven transactions, and why courts will play a central role in shaping the outcomes.

Today we're going to focus on the most consequential legal developments at the intersection of tort law and emerging technology that is civil liability for artificial intelligence. AI is embedded in financial services and many, many other industries. And as these systems become more autonomous and complex, they raise fundamental questions about accountability. When

AI causes harm, who's responsible for that? The developer, the deployer, the end user, or some combination of actors? The legal system is only beginning to grapple with these questions and existing doctrines clearly weren't designed with AI in mind.

That's why the American Law Institute, or we'll call it the ALI, has launched an important new project, developing principles of the law for civil liability for AI. Unlike a restatement which synthesizes existing case law, a principles project, in my view, can take a more forward-looking approach and provide guidance in areas where the law is still developing.

I'm joined today by Professor Mark Geistfeld, who is the reporter for this ALI project. Professor Geistfeld is the Sheila Lubetsky Birnbaum Professor of Civil Litigation at NYU Law School, and one of the nation's leading scholars in tort law and liability theory. As a reporter, he is responsible--principally responsible, I should say--for drafting the principles and guiding the project through the ALI process.

And having been involved myself in a process that was also driven by two reporters from NYU Law School, Oren Bar-Gill and Florencia Marotta-Wurgler, I can tell you it sometimes can be a very, very long process. In that case, I think it took a total of nine years to develop a restatement of consumer law.

So without further ado, I want to first of all provide a very warm welcome to Professor Geistfeld, who I'm going to refer to as Mark from now on. So welcome to our podcast show.

Mark Geistfeld:

Thank you for having me. I really appreciate it. Looking forward to our discussion.

Alan Kaplinsky:

Yep, me too, Mark. So let's start right at the beginning and build a little foundation before we get into a few of the more complicated areas. And let's talk about the procedure involved here. What prompted ALI to undertake a project on civil liability for AI? Was it that some people went to them and said, "Hey, you really ought to do that"? Tell us about that if you will.

Mark Geistfeld:

Yeah. The genesis of the project was right that fall after ChatGPT came out and everybody started to become aware of these models for the first time. And Biden was the president then and they were talking about everybody's P(doom), the probability that we're doomed because AI is going to take us over and become our overlords.

At that time, there was concern by both the Carnegie Foundation and the RAND Corporation that the industry was not taking its safety responsibility seriously enough, and that they had been overly presumptive that you need some kind of enabling legislation from Washington in order to get a regulatory framework. And so they hosted a conference in Palo Alto with a bunch of industry insiders and then it had a few tort folks like myself there to essentially let them know, "By the way, you're subject to tort liability right now and if you're not careful, you're going to pay for it."

And out of that discussion, there grew recognition internal to the industry. It's like we would welcome closer specification of the framework and so on so that we know what we're supposed to be doing in this space. And so then I developed a proposal for the ALI along with some people in the industry and Carnegie to put this project together. The ALI approved it and we've been up and running for a little over a year now.

Alan Kaplinsky:

When it originally got formulated, was it pretty much a foregone conclusion that if ALI took this on, it would be a principles type of project rather than a restatement or anything else that they might do?

Mark Geistfeld:

Yeah. No, it's a restatement as the term implies requires one to rely on existing case law and to restate what the cases have been holding. In the AI space, there's just very little law at this point to restate. And so it would be overly skeletal to try to build out a restatement framework.

A principles project is actually a kind of perfect fit in this space. A way to conceptualize the project is that what we're doing here is what lawyers tend to do all the time. We know we have a technology that we can understand. We have fact patterns that we can then think about, and we have existing doctrine and principles underlying that doctrine that we then apply to the fact patterns to try to figure out what the appropriate legal response is.

So that's essentially kind of the framework that we're doing. The project itself ends up being written exactly like a restatement. There's a black letter and there's comments and a reporter's note and so on. But the effort is really just to show how existing doctrine and law as modified to fit this new technology would apply to a range of different kind of harms.

Alan Kaplinsky:

Right. So at the time when this thing got launched, which as I take it was a couple of years ago, I think, or maybe a little over a year-

Mark Geistfeld:

Yeah, a year and a half ago, probably. Yeah.

Alan Kaplinsky:

As a torts professor, had you already begun to see courts that were grappling with trying to apply tort principles, common law principles to AI liability situations?

Mark Geistfeld:

Oh, yeah. There had already been a few, not necessarily even AI difficulties early on figuring out how software fits into tort law. Is it a product subject to products liability, or is it a service subject to ordinary negligence liability? So there are cases of that sort.

I had gotten into the space initially with autonomous vehicles, which had been on the radar screen. They're kind of old school now, but they were back in 2015, everybody was talking about AVs taking over the highways within the next couple of years. And it was pretty clear in looking at the liability issues in that particular space once I delved into it that yeah, doctrine works here. But I felt like I was taking a final exam and everything I knew about products liability in that instance to figure out how it would apply to the crash of an autonomous vehicle.

Alan Kaplinsky:

Right. So get back to how you got involved with the project, you got invited to this initial meeting or conclave where experts in the area of tort law and I assume that some experts in the area of the technology itself got together and decided there was a need to bring some order into this area and to do it through a principles project. How did you get elected? Did you draw the short straw or...

Mark Geistfeld:

It's a little overstatement of the earlier. The project kind of grew out of that discussion. The purpose of that particular conference at Palo Alto was not to generate tort liability. It was just one of the many panels. It was all about making sure that AI is as safe as it ought to be.

But for me, once the opportunity arose, I definitely was enthusiastic about taking this on because once I learned about the issues, it was quite obvious to me that this is what I want to work on moving forward. And this project ends up just being a great platform for me, both a privilege and a fantastic, interesting opportunity combined. So of course, I'm going to go for that.

Alan Kaplinsky:

Yeah. Have you done or had you done writing in the area other than the proposal that you wrote that you put together for ALI?

Mark Geistfeld:

Oh, yeah.

Alan Kaplinsky:

Yeah. So you've done law review?

Mark Geistfeld:

Yeah. So I have a law review article on Autonomous Vehicles in the California Law Review, I think 2016. And based on that, I did work for the European Commission when they were developing the EU AI Act. They looked to me on guidance for where the US was at with AI regulation in general, which at that time was very little, still in many respects. It's at the state level much more robust, but federally not a whole big difference from back then. So I just moved into the space.

Alan Kaplinsky:

I'm curious, just the work that you did for the EU, are they further ahead of the US in this area?

Mark Geistfeld:

Oh, yeah. The EU AI Act is very comprehensive and demanding. They're having trouble implementing it, trying to figure out how to satisfy all the various disclosure obligations. So yes, they were distinctively first movers and kind of proudly so and rightly so in the space.

Question whether they moved a little too quickly. They kind of built the framework originally around autonomous vehicles and associated forms of AI. And once the generative AI stuff came out, it was really a different kind of mode that they hadn't contemplated, and they had to kind of pull back a little bit and see if they could incorporate that into the overall framework. But without question, the EU is moving ahead on this.

Alan Kaplinsky:

Yeah. Okay. So I think you've already, I want to make sure our listeners, I mean, I already described the principles project a little bit and you've referred to it. But I'm sure our listeners, most of them are not all that familiar with what this end product is supposed to look like, and how you anticipate it's going to be utilized once it's actually published.

Mark Geistfeld:

Right. Yeah. Well, so it involves a series of sections and this is to principles for restatement project, they would be rules. But they take a black letter form like rule number, principle number one is common law principles of tort liability apply to AI models and systems. So then we have to give a working definition of an AI model and an AI system, kind of lay out the framework there.

And then next section will be like the obligation to exercise reasonable care applies to one who distributes, provides, or releases an AI model or system. And so the first portion of the project is really about distribution and ultimately about commercial distribution, because that's of course where the locus of primary concern is in this space.

And there'll just be a series of rules then kind of laying out to establish causation. These are the considerations that are applied.

Alan Kaplinsky:

Do you, like a restatement, give examples to? Do you come up with hypothetical examples and then explain how the principle is going to apply?

Mark Geistfeld:

Yeah. The whole thing reads just like a restatement until you get to the reporter's notes, and then you realize that we're actually just reasoning our way to these conclusions.

Alan Kaplinsky:

Yeah. I mean, is it fair to say that like a restatement project, which is supposed to basically restate the common law and not project what the law ought to be, the principles is more on the projection end of things rather than just restating the body of common law?

Mark Geistfeld:

Well, I think of it more as not so much what the law ought to be. It's what the existing law as adapted to the new technology would require. Because if we were to try to kind of make up on a whole cloth what we thought de novo the best way to do things, first of all, I don't think it would ever get through the body of the ALI because that'd be pretty contentious. And moreover, for good reason, because why would a court pay any deference to that way? There's nothing about that body that gives it expertise. Our expertise resides in the ability to take existing legal materials and to derive sound legal conclusions from those materials.

And so that's essentially what we're doing. So it has to be forward-looking because again, we don't have cases that we can use, but we do have the restatement third, which is a multi-volume undertaking. So there's a thousand of pages of black letter law and case support and so on. That's the corpus of materials that we use to figure out how that applies to AI-caused harms.

Alan Kaplinsky:

Yeah. So that's not final yet. Am I right?

Mark Geistfeld:

Just about final.

Alan Kaplinsky:

Oh, is it? Okay.

Mark Geistfeld:

Yeah.

Alan Kaplinsky:

So I take it... Well, I'll ask the question. Was one of the reasons why ALI decided there was a need for a third restatement, was that a result of technological developments like AI and things that were going on with software and things of that sort?

Mark Geistfeld:

No, it was just serendipitous in that respect. I think it probably, partly there's just the restatement second, probably I think finished in the 1970s. Its most successful provision, Section 402A, which spawned the regime of strict products liability. In the restatement second, that was about 14 pages, let's say. And by the '90s, the ALI decided to restate just that portion of tort law.

And the product restatement, it's like 300 something pages. So there's that much case law that developed over those three decades that warranted that kind of treatment.

So there's enough growth in tort law over the latter half of the 20th century that justified a third restatement, and then the AI stuff just drops in at this point in time serendipitously because from my perspective, the timing couldn't be better on the conclusion of the third restatement and the outset of this project.

Alan Kaplinsky:

Do they... In the third restatement, are there things where statements or black letter principles that are directly germane in a particular kind of way to the project that you're doing? I'm talking above and beyond the fact that general tort law principles are going to apply.

Mark Geistfeld:

No. So the products restatement, which was finished in the mid '90s takes a position at that point, says the case, we're not writing a black letter on software because there's just not enough case law right now that's relevant to this question. So we leave it to the developing case law to figure out whether software is a product subject to these rules or as a service subject to the rules of negligence.

So no, there's really very little in the black letter itself expressly applicable to AI stuff.

Alan Kaplinsky:

Got it. So tell us, the audience, if you would, Mark, on where things stand with the principles project right now. I know you said-

Mark Geistfeld:

Yeah. So we're drafting, finishing up the first set of materials that we've got. There's an associate reporter, Ketan Ramakrishnan at Yale, and myself are working on this right now, drafting. And we have a decent sized set of materials that we submit in a couple of weeks. There's going to be a meeting. We have an advisors group, which is about 60 people, let's say, comprised of general counsel from the leading labs, law professors like myself, practitioners like yourself, who are all in the AI space that then give input on the draft.

And then once we are kind of happy enough with the current state of that material, it goes to the ALI Council, which is a standing body within the American Law Institute that vets everything that goes through it. And then once it gets through that, it goes to the membership as a whole for a vote at the annual meeting. And once that's approved, then it gets out of the world. Now there may be distributions before that final endpoint, but that's where we are in the process right now.

Alan Kaplinsky:

Is there any way to project how long it will take before it gets to the point where the section council will approve the principles?

Mark Geistfeld:

Well, we're all keenly aware of the need for speed here because the technology is developing so rapidly. And so I mean the project as a whole, just as you alluded to earlier with the one on consumer law, the entire project before you end up with a bound volume of everything that's completed is going to be easily a decade. But because of the stature of the ALI, once we start to get material out into the public, it starts to shape discussion within the legal community. The courts will cite preliminary drafts and things like that.

So we could start to influence things well before the whole project is completed. And so our kind of decision making on what to do, the ordering, the sequencing is very much driven by that understanding and the need to do things as quickly as possible in a competent way.

Alan Kaplinsky:

Would you say in terms of the advisors that you have, 60 or so, and then I know there was called a members consultative group of people who are members of ALI, who want to just be involved in the meetings can do that, are all the various stakeholders do you think represented the AI, the Mag 7, who I would consider to be among the leaders and the chip makers like Nvidia and et cetera, et cetera?

Mark Geistfeld:

Yeah. Well, the ALI is very aware of the importance of having a group like this being represented across the board because the extent it's overly the project, the material ends up being overly one-sided. It just diminishes the effectiveness out there in the world. So is it as well-represented as it could be? That's never going to be the case, but it's a highly representative group across all the relevance.

Alan Kaplinsky:

Right. Let's turn now to what I'd call the core liability framework. And my question is this, at a high level, what are the core liability frameworks that the principles are developing for AI-related harms?

Mark Geistfeld:

Well, the threshold question that you get in, and you've already got cases litigated along these lines and it's going to happen in any complaint that's filed involving physical harm, certainly, is the plaintiff is going to allege both negligence and products liability. And the courts then going to have to decide whether or not the AI, the algorithm more generally, falls into one category rather than the other.

And so the social media cases, the addiction cases, the verdicts that came down a few weeks ago, the papers were talking about how these were dealt with as products liability suits, that the algorithms are defective in certain respects. And so underlying that is a determination of what is it that makes an algorithm a product.

Now the issue, interesting, so it turns out it kind of illustrates the difficulty of applying established doctrine to this technology because the products liability law developed with respect to tangible products. I mean, when you think about a product, you think about something that you pick up. And then you think about a service, the provision of legal advice, financial advice, whatever, that's an intangible.

And so the tangible-intangible divide has basically done all the work on the product service distinction, so much so that like the New York Court of Appeals just a few years ago in a case had a reason to define a product for purposes of tort liability and recognize we've been doing this for decades and we actually never got around to defining the product concept because it was obvious we didn't need to do that.

And so when you get into what is it that makes an intangible good like an AI model system platform algorithm, what is it that makes it a product? Turns out to be surprisingly difficult. In the first instance, you have to figure out, well, why is the tangible versus intangible divide? Why was that so important? What is it about tangibility that really matters for purposes of the tort inquiry?

Once you work through that and nobody's ever, again, it just hadn't been thought about before because it wasn't an inquiry that was ever necessary. And once you work that through, then you realize, okay, the tangibility matters because the manufacturer's safety decisions are baked into the product. So you can evaluate the product independently of the manufacturer's conduct.

And that's a pretty decent marker between products liability and negligence, which is conduct focused. So product liability, you focus just on the performance of the thing, the product. Negligence, you focus on the conduct of the provider.

Now the difficulty, so you can work that through, that turns out to be not that difficult once you go down that path. But then the problem that arise is that when tort law in the past is dealt with product claims involving the provision of information, you can instantly see there's a real awkward fit. In what respect is the provision of information defective? What's the appropriate design of information? Is it the truth? Well, sometimes, but usually not.

And so it gets into how do you apply this framework to the provision of information? The early cases, the social media cases adopted this position that said, well, if the plaintiff's allegation goes to the content of the information, then it is a service subject to negligence, not a defect. But if it goes to the method of delivery, kind of like the cover of a book, then that's more like a product subject to defect based rules of strict liability.

Now the problem with that approach is that every output from an AI model or system is information. That's all it gives you is information. So if you're deeming the provision of information to be a service full stop, that's the end of the inquiry. It's uninteresting. So you need a theory of information. What kind of information? Is it that counts as a product versus what kind of information counts as a service?

So you can work all that through and we've made good progress in that respect, but you notice how far you've had to move from just a simple thing where you didn't even need to define a product or a service. It was obvious to now you're in this space where you're having to come up with a theory of what kind of information is the right type for products liability as it compare to ordinary negligence liability.

Alan Kaplinsky:

Yeah. And so have you and the advisors reached a conclusion of where you draw that line?

Mark Geistfeld:

Well, our first cut at it was received lukewarmly. There were some problems with the way that we had initially specified it. And so we went back to the drawing board. It was very helpful, I think located the source of the ambiguity that was giving rise to some of the concerns.

So I think we've actually got a... I feel pretty comfortable with the way we've got it worked out right now. I know that the European Union folks over there are struggling with the same issue, so we're all kind of deeply interested in this right now. But we have to send this set of materials out in a couple of weeks and we'll have a meeting in late June when we go over it.

Alan Kaplinsky:

Right. Okay. So one of the biggest challenges I'm sure that you're facing is... Or let me ask the question. I assume it's how you allocate responsibility among all the players that are involved in an AI system, the developers, the deployers, the integrators, the users. How will the principles approach that issue?

Mark Geistfeld:

Well, as a first cut, we're setting off the users the liabilities that one can incur by using an AI model or system and that use causes harm to somebody else where the responsibility primarily resides with the user. We're leaving that as the round two. So now we're really focused on the providers, again, ultimately with the primary focus on the commercial providers.

Now within that particular set of actors, as you recognize with your question, it's incredibly complicated due to the way that AI models and systems are developed. It's not like a stream of commerce where you have the component part supplier gives a part to the manufacturer who assembles it and then distributes it through the warehouses into the retail store. It's much more layered. It's called a system stack and the base model runs everything. And then there's their guardrails and scaffolding put on top of that.

So you have to define the responsibilities of the commercial actors in this space in relation to the layer they occupy in the stack because each layer affords different kinds of opportunities and therefore gives rise to different kinds of precautionary obligations. But it's all tractable, but it takes a fair amount of time developing the right framework for liability because it's just, again, different from the kind of standard setup that we've had before.

So I've been living with that problem for the last few weeks. Again, I think there's a decent framework, but you have to first just make sure you're clearly locating everybody within the system stack so that you can then clearly identify what their obligations are in relation to their position.

Alan Kaplinsky:

And do you look at the concept of control of the AI system as a factor, or to what extent did one of the parties have the ability to prevent the harm?

Mark Geistfeld:

So each layer affords different opportunities. So the base layer, the developer, the commercial lab that develops the foundation model, it's called the base model, foundation model because of this stack idea. Everything's built on top of it. And so that's really the only player, at least with respect to the performance of the base model itself, who can affect its performance by the data that it selects to train the model on.

And so the obligation to make sure you're using representative data that's not biased, et cetera, et cetera, once the base model is trained with that, the parties higher up in the level can't really do anything in that respect. They don't have any control over the curation of the model itself. So the primary responsibility for that has to reside with the foundation model developer.

Now there can be parties up in the stack who then do so called fine-tuning of the model to get a specific kind of financial transaction or work in a law firm or something like that. So they can then curate the model further on specific data tailored to that particular application domain. So they can have curation obligations at that level. So you can pull it in a little bit, but the basic training of the model itself, responsibility is going to reside with the foundation model developer.

And you can kind of then run through all the different safety measures that are available like monitoring. Is the system being used in irresponsible way or not? It can be varied for a foundation model developer that just builds the model, but then other parties are the ones who end up commercially deploying it. The monitoring ability resides with the upper layers of the stack and not with the bottom layer of the stack.

So you just have to be very careful rather than talking about things in general like monitoring or curating data, et cetera. You have to be very careful about locating those obligations to the layers in the stack where the actors can actually control those kind of safety decisions.

Alan Kaplinsky:

Right. Let's turn to the concept of reasonable care, which is a fundamental concept in the area of tort law. And how does that translate in the context of AI systems? Does that include some of the things you mentioned already, testing, monitoring, human oversight, for example?

Mark Geistfeld:

Yeah. So there's a lot of respects in which the application of reasonable care to these models is very much like a product liability problem. When you're talking about the architectural design of a model, there's not a one-size-fits-all architecture. There are trade-offs involved in any kind of architecture, just like there's trade-offs involved in any kind of design of the frame of a motor vehicle.

They work, they create risks, they have functional benefits and so on. But that kind of issue is one that products liability laws dealt with for a long time. And that framework you can apply to lots of the kind of structural features of AI models and systems.

Then when you get to things like behavior, should they be monitoring? Again, there's an analog to human behavior there. If you're in a situation where risky things are going on and you're in a position to be aware of those problems and control for them, you're going to have a duty to exercise reasonable care with respect to monitoring.

So it's really, the task is you're looking for the right analogies and they pretty regularly exist. It just takes a lot more work to kind of structure the problem before you can then see the evident analogies.

Alan Kaplinsky:

In a lot of industries, and I'm thinking I'm in the world of consumer finance, a lot of industries have created best practices for a particular industry. And is there any role for industries to create best practices, or is that something quite apart from what you're creating?

Mark Geistfeld:

Yeah, no there is. And it's actually, this is a nice... I think this is maybe the most important, not the most, but it's certainly in a critically important way in which the tort system and the civil justice system more generally can interface with the regulatory system in a complementary way.

So the California SB 53, New York RAISE Act are state statutes that impose disclosure obligations on the large commercial labs with respect to their powerful models about exactly the kind of measures you're talking about, what are the safety measures you're utilizing and so on. And the underlying idea, which is very appealing and attractive is that this is a space where transparency is going to be really important because these models are very powerful. And so having knowledge of what's going on in the industry can help to facilitate good practices.

The world typically is one in which what goes on in companies is behind closed doors. You need litigation and discovery to air this stuff publicly. It'd be much better to be aware of different labs are trying different things, and then you could build that into the liability rules in a way that you'd like to create an incentive for a race to the top rather than a race to the bottom, which is the kind of market failure you always worry about.

Alan Kaplinsky:

Yeah. So you talked about this a little bit earlier, but I want to go back to it for just a moment, and that is how the principles treat AI systems when they're treated as products for product liability purposes, which could result in strict liability under some circumstances. And how do you deal with systems that evolve over time? I'm wondering if you could respond to that.

Mark Geistfeld:

Yeah. Well, part of it is to take it all the way in some respects, how do you deal with the fact that what you're unleashing in the world is in some respects autonomous. It's not like a product of the past where it's been designed and that design just moves through the physical world. This is not something that adapts to the physical world and in that respect is arguably autonomous. How do you deal with that?

At one level, the way you deal with it is identical. There's just a question of foreseeability. When you put this thing out into the world, what do you know or should you know about its general performance characteristics? You don't need to know the precise details, but there's foreseeable things, general kind of patterns that you can foresee. And then your safety obligations are kind of tailored to those general categories of risks, even if you don't know exactly how it's going to play itself out in a particular case.

Alan Kaplinsky:

Let's talk a bit about causation. That's another fundamental principle of tort law and it's already an area that's pretty complicated. It's not simple. AI complicates it even more. Am I right?

Mark Geistfeld:

Correct.

Alan Kaplinsky:

What's your reaction to that?

Mark Geistfeld:

Yeah, no, my intuition, although I've been proven wrong in the sense that every time I think something is going to be easy, it turns out to be really complicated. So it's really hard to know until you get deeply into it and we haven't moved deeply into the causation issues yet. But I had to think about them in the context of autonomous vehicles and of course I've been thinking about it. And you're right that there's complexities that are going to be raised here.

And probably the best analogy at least to think of in the first instance is these are probabilistic models, which is the difficulty of ascertaining what's going to happen in any given case as we were discussing earlier. And once you start to work with probabilities, the closest analog in tort laws is with drugs. There's a 3% chance the drug is going to cause a side effect. Toxic torts, exposure to a chemical elevates the risk of cancer by X amount. Medical malpractice, you don't give the chemotherapy within two months of diagnosing the condition, the chances of survival are reduced by 30%.

There's those spaces of tort law where you have to work with probabilities because that's the only information you have. We don't know why a particular drug causes cancer in some people, but not in others. So all we do is say, "Hey, 3% of the population who's exposed to the drug get cancer," something like that.

And so tort laws had a hard time with those probabilities. And so I expect to see the same thing play itself out in this context. I think those are going to be some of our hardest, most contentious issues of proving causation because the difficulty is that if the causation requirement ends up affecting barring recovery because you just can't prove it, what you've effectively done is you've just negated the associated tort obligation altogether.

If I have an obligation to exercise reasonable care, but I'm never held liable because nobody can ever prove causation under the ordinary rules, then the proper legal conclusion is that in fact, I owe no duty full stop. Because if you say I owe a duty and somebody has a right, but then we construct the liability rule so that anybody who's injured and claims a right violation can never recover because they can never, ever prove causation, we have a right without a remedy and that's problematic under multiple state constitutions.

So you got to walk this balance. On the other hand, you just don't want to say, "Hey, plaintiff, it's a hard time proving something, we'll give you a free pass here." So it's a very difficult balancing act in that space.

Alan Kaplinsky:

So how do the principles deal with what's referred to as the black box problem where even a developer may not fully understand the output?

Mark Geistfeld:

Yeah. No, for a lot of people think it's impossible and so let's just move to something like strict liability. We don't know. And so as long as the AI causes harm, that's something we can maybe observe, bracketed, qualified with our prior conversation there. How could we do negligence if we can't look in the black box?

It's true that the black box is difficult in some cases, but it's overstated that it is a problem that applies across the board. There are going to be times when, for example, you would want... There's good reason to think that there are going to be spaces and I think that the financial services sector is one where you're operating online and you're dealing with an agent, an AI agent on the other side of the transaction.

There's every reason to think that, and again, there may be qualifications here, this may be overstated, but it's plausible as a first cut that the baseline assumption in those kinds of shared environments is that simply that the AI agent acts like a reasonable person would under the circumstances so that you can coordinate behavior with humans in this shared space. Those kinds of AI models and systems aren't designed to make the world a safer place. They're really just designed to save on human labor, substitute for human labor.

And so if you think about that kind of problem, which is obviously a massively practically important problem, you could see a case being made that the agents should be held to the standard of human reasonable care. And then we don't need to look into the black box any more than we need to look at the black box of your head to figure out whether you exercise reasonable care. We just evaluate the behavior in relation to a baseline of what humans exercising reasonable care would do.

So as you work through the myriad kind of problems, you realize there's just times when really the black box is irrelevant. For sure there are times when it is relevant and those are going to be hard, and I suspect they're going to require targeted rules of strict liability to overcome some of these evidentiary hurdles. But it's almost always, you kind of end up with the lawyerly response of, "Well, it depends," because every case is different and you really need to work through the salient differences to figure out the right approach.

Alan Kaplinsky:

Right. What's the role that foreseeability plays in an AI system? Can it produce emergent or unexpected behavior?

Mark Geistfeld:

Oh, it can for sure. Yeah. But it's foreseeable if I am developing a model that's going to be used for medical diagnosis. It's foreseeable that if I train the model only on white males that it's going to be biased, going to be pretty good at detecting tumor whatever in white males. But if you start to apply it to any other kind of person, it's not going to perform well in that space. That's just a very easy, foreseeable consequence of the architecture of these models and the failure to adequately curate the training data.

So there are lots of problems of that sort where you can foresee if you don't put a guardrail on a filter on the model so that if somebody just asks a question, it's foreseeable that there's going to be a bad actor out there who's going to say, "Hey, how do I build a bioweapon?" So of course, we want filters like that on it.

So there's lots of things that are foreseeable that merit precautionary measures. You're still quite right, of course, that there's going to be things that due to emergent behaviors that we just can't foresee, but it's true also of humans.

Alan Kaplinsky:

How do you balance the risk of chilling innovation against the need for accountability? Your focus I guess is... Well, let me put it differently. Is innovation something that you have to worry about fostering or do you just do your job and if you do the best job you can and if it ends up chilling innovation in certain areas, that's not your problem to worry about? Or am I stating it incorrectly?

Mark Geistfeld:

Yeah, no, you're stating incorrectly. It is definitely a problem. It's something that courts have always worried about with liability. You don't want liability to chill socially valuable activities, and clearly the development of AI technology is a socially valuable activity. And to put an even finer point on the issue, AI has a tremendous promise to cure disease and to make the world a safer place.

So if in our effort to construct liability rules to make sure that AI behaves safely, we end up filling a safety innovation. We might have a self-defeating outcome where the world at the end of the day is a less safe place than it otherwise would be. So you most certainly worry, you're quite sensitive to making sure that you don't make liability rules that are going to have an undue chilling effect. What you want to do is you want to chill the unsafe practices like to have a targeted freezing device there is what you're looking for.

Alan Kaplinsky:

Yeah. Maybe this is completely off base and if it is, Mark, tell me. But the brouhaha that developed a few weeks ago when Anthropic came out, disclosed that a certain advanced model of its AI system was able to detect flaws in all kinds of other systems that people have such as systems that are protecting them against the data theft or hacking, the ability to hack into a system.

And they assembled, as I understand it, some of the people who were very knowledgeable in this area, maybe even some of their competitors, although I don't think OpenAI was there because they're direct competitors. Is that something that has an impact on what you're doing at all or is that in a different-

Mark Geistfeld:

Well, no, that's essentially what... I mean, the reason why this is such a moment is that it's clear that these models are developing quite rapidly and are getting more and more powerful in ways. And with the increase in power comes both the opportunity to do good and the opportunity to do bad.

So this particular example is really stark in that respect. The technology here is really dual use. The coding ability of the Anthropic model that would allow a malicious actor to hack into a utility and bring down the power grid, that same capability as you were describing, also allows the power grid and cybersecurity experts to use the same model to check their system to see whether there are any vulnerabilities in it that might get hacked.

So it's dual use. It helps. It has both offensive and defensive capabilities. And so it illustrates the difficulty of getting the right balance. So what Anthropic is doing in this instance is it's giving it to the cybersecurity community so they can go out there and try to fix all of the easy vulnerabilities before this model might become made more public where the bad actors might have access to it. So it illustrates the dual use nature and how to try to find the right balance between offense and defense.

But the other point, again, is just that this stuff is really rapidly developing and it's just that the time since I've been using it is extraordinary how much more capable the models are today than they were when I first started playing around with them. And it's just going to accelerate it. Certainly whether it accelerates or not, it's moving at a pretty rapid pace and this just illustrates that dynamic.

Alan Kaplinsky:

I'm just going to ask one more question because we're running out of time. To what extent will the principles consider regulatory compliance as being relevant to liability? There's not a lot, as I think you pointed out at the beginning, at least at the federal level, there's nothing really specific that I can think of related to regulation or a statute related to AI. I mean, Congress, the White House has come up with this framework and supposedly will lead to the enactment of legislation at some time down the road. I have no idea when.

But to the extent there are any regulations at all, I guess there are some at the state level. Colorado has a statute, and California does. There are bills in various states but there isn't a lot of regulation. Is that relevant to what you're doing?

Mark Geistfeld:

Oh, yeah. Regulation always matters. The way that tort system, rather than competing with the regulatory system, can complement it. And so you have to figure out when the two are working in tandem, when they're working at cross odds. When it's cross odds, then you're in the space of preemption. One or the other has to give way and the supremacy of the legislation means tort law gives way.

But if you're careful, you do it properly. The two can work together as they should because each has different capacities and comparative strengths compared to the other. And so an ideal regime is not just torts, it's not just regulation, it's a nice combination of the two. And to the extent we can help bring that about, that's most certainly an outcome we'd all be happy with.

Alan Kaplinsky:

Okay. Well, we've come to the end of our program today, Mark. And first of all, want to thank you again for taking the time to share with our listeners extremely important undertaking that you're involved in. And as this moves along, I'll greatly appreciate it when you reach the point where maybe you don't have a final product because I'm not sure I'll still be involved practically at that point. I'd love to hear about it because we'll want to have you back on the program to talk about various aspects of it.

Mark Geistfeld:

Great. Well, thanks for having me. It's interesting conversation.

Alan Kaplinsky:

So let me just leave our listeners with a few of my takeaways from our podcast show today. So first, AI is testing traditional tort concepts such as duty, causation, foreseeability and responsibility.

Second, ALI, the project is especially important because it's not merely summarizing settled law as applied to AI systems. There isn't much of that. It's helping courts and policymakers think through sound doctrine and what sound doctrine should look like and how it should be applied to AI systems in a very rapidly changing environment.

Third, one of the most difficult issues is how to allocate responsibility among the many actors that are involved in developing and using an AI system. And finally, striking that right balance between innovation and accountability may determine not only the future of liability law but also public trust in AI itself.

So Mark, thank you again for joining us and even more importantly, taking the lead role on these principles or project. And thanks to our listeners for tuning in and we'll continue to follow this project very closely as it makes its way through the ALI. I hope everybody enjoys the balance of their day.