**An Interview with Mark Chinen**

**Marcus Schwarting**

Department of Computer Science

University of Chicago

meschw04@uchicago.edu

**Introduction**

I recently had the opportunity to interview Professor Mark Chinen from the Seattle University School of Law. He was educated at Pomona College and Yale Divinity School before receiving his law degree from Harvard Law School. Before he began law teaching, Professor Chinen practiced in the areas of international trade, banking and corporate and securities law in Washington D.C. with the firm Covington & Burling. Professor Chinen teaches contracts and courses in international law and writes on various aspects of international law, particularly international governance, theology and international law, and the relationship between domestic and international law. His first book is *Law and Autonomous Machines: The Co-evolution of Legal Responsibility and Technology* (Edward Elgar Publishing, 2019), and is the subject of some questions from our interview. His most recent book, *The International Governance of Artificial Intelligence*, is projected to be published in mid-2023.

Over the course of our hour-long interview, we discussed several interesting and prescient issues. These include the legislation and division of responsibility surrounding autonomous vehicles (AVs), how the legal system should address baked-in biases in artificial intelligence (AI) systems, the continuum of autonomy and its implication for AI legal personhood, and the ability of blockchain smart contracts to be legally binding. We end with a brief preview of the topics included in Professor Chinen’s newest book.

**On AV Responsibilities and Legislation**

*Marcus S.*: How do you think about the way that developers, manufacturers, and legal councils ought to weigh their responsibility for passenger safety versus pedestrian safety in the case of an AV? And can discerning those differences be codified in a way that is explainable to an AV?

*Mark C.*: What I would start off by saying is that you will not find it surprising that I tend to look at this from a legal perspective. What we think of as the balance between, the interests between a passenger or the operator of the vehicle versus the pedestrian, you could make the argument that a rough balance has been achieved by different sources of the law. For example, there are regulations regarding passenger safety and the safety for vehicles, and the standard that those vehicles need to meet before they can operate on the roads. Then there is the whole panoply of traffic safety regulations. Then there are basic rules of the road regarding pedestrians and rights of way. In a sense you can argue that the law has already achieved the balance that you are asking about. So then when you talk about the responsibility of developers as they are trying to design these devices, I tend to think we ought to keep those laws and those regulations in mind. Because at a minimum that should have achieved that rough balance that you are describing. And then we can work out where you start to bump into edges or gray areas where maybe the law cannot reach some of the new phenomena that are created, because now we are talking about an AV as opposed to a human-driven vehicle. Does that make sense?

*Marcus S.*: Yes, that does make sense. And is the hope that we can really take those laws of the road and start to break them down into the first- and second-order logic that is explainable to an AV?

*Mark C.*: I would put it this way. Right now, every attempt is being made for that to happen. It has been a few years since I wrote my book, but at that state researchers were breaking down every sort of driving scenario into various parts. In a sense, they already had a head start. When you think about the regulations that already exist and some of those safety requirements, those are already framed in terms of specific operations like rates of speed, rates of changing lanes, and those kinds of things. You can make an argument that a lot of it is already operationalized. Then it is about writing the code that matches it. It is often not as cut and dry as I am describing it because there are often times when, particularly when you are talking about regulations or traffic codes, there is room for interpretation. Then if you are a programmer of a developer, in a sense you are guided by the general council of your firm. Or if you have some legal expertise, then in a sense you are interpreting the law as you develop the code.

*Marcus S.*: Should developers really be put in the position to be interpreting the law in that way?

*Mark C.*: Personally, I think the answer is no. As an institutional matter, we leave it up to courts and those who are trained in the law to do that. Of course, the reality is we are all interpreters of the law in that we, in our everyday activities, have vaguely in mind a sense of the law even though most of us are not trained lawyers. Even if we do not talk about autonomous systems, if you are an engineer designing a conventional vehicle, in a sense you are also interpreting, but obviously not to the same level as if you are designing something for autonomous operation.

*Marcus S.*: That really places a weight of responsibility on the developers of that code. I think that is a fantastic thing for people who work on those types of systems to keep in mind.

*Mark C.*: Indeed. It raises these broad questions about who gets to decide those kinds of issues. I do not know if the legal community, or the public more generally, is aware that by merely designing, developers are applying and interpreting the law.

*Marcus S.*: Sure. Another question along the same vein is, especially with AVs, you see this incremental progress of an AI where it starts to supplant more and more of the human in the loop.[[1]](#footnote-1) But in the meantime, you have what I have heard people call a “chimera” [centaur][[2]](#footnote-2), where you have this computer plus a human and you get even better performance than just a computer or just a human. In the situation of AVs, or even more broadly, if a [centaur] were to commit tort, what heuristics do you see being available to separate the blame between the human and the machine?

*Mark C.*: I think we would have to flesh out a particular scenario. When you think about a [centaur], I am assuming this is where the AI system or component has a relatively high degree of autonomy. Because if it is fair to construe the AI device or system as simply a tool of a human being who is using it, then I think the lines of responsibility are going to be straight forward in drawing. A human making use of a tool then committing a tort are standard legal constructs that I think are going to hold. But then as the AI system gets more and more autonomous, or more intelligent if you want to call it that, I think that is where the constructs are not going to work as well. In that scenario, you can make an argument that the AI system is just as responsible for the harm as the human with which it is working. Now what might happen in that situation is, assuming the human whom the AI system is working was not a manufacturer or designer of the system, then we can start talking about the concept of products liability. The responsibility would then turn to whoever designed the system. There is a debate in the literature about how effective products liability is going to be in addressing the designers or manufacturers of those kinds of devices or systems.

*Marcus S.*: Sure. That is a very tight line to walk, on either side it seems one could err. Building on that idea, in the future are you at all concerned that the legal system could potentially scapegoat a human bystander for errors that may be caused by an AV that are completely outside the human’s control?

*Mark C.*: That is interesting that you would use the term “scapegoat”. I might use a different term, but it is closely related to what you are describing, Marcus. As AVs become better and better, part of the impetus for these vehicles is to take away the prospect or the potential for human error, which is the leading cause of accidents.[[3]](#footnote-3) So as AVs become more and more common, it is almost as if they will set the standard for care in terms of what constitutes careful driving. Or in the medical practice, what constitutes due care in providing medical treatment. It is almost as if when a human steps in who is prone to error, almost by definition if something goes wrong they will be the first identified as being probably responsible for it as opposed to the AI systems and devices, especially as they get better. But of course, this depends on our ability to make progress with these devices. My guess is that they will get better and better.

*Marcus S.*: And therefore less likely that the car is at fault and more likely that the human is at fault?

*Mark C.*: Yes, right. It is almost as if we will all take risks if we decide we want to go driving by ourselves. Because if 90% of the vehicles are autonomous, you probably would be identified as the first thing to look at when we are trying to assign responsibility.

*Marcus S.*: Right. These AI systems are working around you at that point as the unknown factor.

*Mark C.*: That is exactly right. They will have identified us and will try to avoid us because they know humans are a risk. I think that is a little far in the future. I do not know how long this transition period will be, but there will be a transition when we will see gradual adoption of these devices. Maybe as they get better and better, we will see them taking over more and more tasks that humans tend to do.

*Marcus S.*: In the case of an AV getting to a point where a collision is practically inevitable, earlier this year the NHTSA released a report on Tesla AVs where, if [the AV] saw it was about to be in an eminent collision, it would effectively abort the autopilot.[[4]](#footnote-4) I think that raises a lot of questions about the responsibility of AV manufacturers to really “see it through” for the whole autopilot process. What role does an AV manufacturer take when a collision is essentially inevitable?

*Mark C.*: I think this is a very complex question, and there are many facets to it. It appears that the designers at Tesla have designed their autonomous system, and rightly so, to essentially cede to human control in those kinds of situations. That reflects the kind of common approach that is being taken now. In any of these kinds of systems we are enabling a human to take over, having the human in the loop, or having the human be able to make meaningful decision-making power when a vehicle or another device is autonomous. Now that is not a fool-proof response. Maybe what we really want to do is not allow humans to intervene, because humans are the source of error. Particularly when it comes to these trolley car-like problems or dilemmas[[5]](#footnote-5), it is almost as if we want the human to make that decision and not the autonomous system, and by extension the designers of the AV. It is almost like we do not want them to make that call. We want to allow a default so that we continue to do what we do now, which is we let humans make that call. Part of the reason for that is, and I do not want to minimize the trolley car dilemma, it is a real one, but most of us will probably go through our entire driving career without ever encountering a trolley car dilemma. Hopefully. So far that has been true for me. And I think that is the case mostly. Then the question becomes how much time do we want to put into the design for that dilemma? It is an empirical question for how many times a driver or an AV might be faced with a trolley car dilemma. If it happens a lot, then maybe we should have the designers make that call. Maybe we should have some other institution deciding how to do that. Some people have tried to argue that what we ought to do is have a menu, where when you step into an autonomous taxi service, for example, and you select the kind of defaults that you want. I do not know how workable that is. But you see that [such a menu] is trying to to maintain this decision-making process that we are already doing, which is that we leave it up to each individual to decide for ourselves how we will “solve” that dilemma. Again, I do not know if that is workable, because dilemmas, by definition, are not capable of resolution.

*Marcus S.*: Right. And there are so many different permutations and context dependencies that it seems like something a human really does have to negotiate. Would you agree?

*Mark C.*: I think so. Although I do not know if we are great solvers of the trolley car problem. We just want [humans] to be able to do that. Whether we evolved that way, or depending on whether you have a faith tradition, however we are, and whatever makes us ourselves and gives us the abilities to make these choices. And there are philosophical debates about determinism versus free will and all these kinds of things. We will leave it up to us at this point. And I think that makes sense to me.

*Marcus S.*: I suppose, unfortunately or fortunately, we do not have the empirical data to back up human decision-making on trolley problems.[[6]](#footnote-6)

*Mark C.*: Yes, I think that is a fair way to put it. And as you said, the other approach is to try to imagine all scenarios and then decide in every scenario. Of course, that raises two questions. One is, can you imagine all possible scenarios? There are some in the literature who argue that it is [possible], and they can show formally that you can identify all scenarios. Of course, the effort it would take to decide the solution in each scenario would be massive. Then someone must decide, and then we must decide who gets to decide the right response in every given scenario.[[7]](#footnote-7)

**How to Legally Address AI Errors and Biases**

*Marcus S.*: In your book *Law and Autonomous Machines*, you discuss the concept of cultural bricolage, which was a new term for me. My understanding is that this means incorporating known legal concepts when dealing with new scenarios, going back to precedent. One illustration you provide is the idea of considering AI like a misbehaving animal. If an animal owned by an individual commits tort, then the individual is oftentimes liable. If an AI possessed by a corporation commits tort, is the company then liable? Is this a suitable analogy? What are the merits of using this case and, more generally, trying to use these past legal precedents to inform our legislation around autonomous systems?

*Mark C.*: So I would start by saying that the idea of bricolage comes from Jack Balkin[[8]](#footnote-8), who is a law professor, and who is also drawing from the anthropologist Claude Lévi-Strauss[[9]](#footnote-9). So that idea is not unique to law, but rather that all of our cultural tools, whether it be the law or any physical tool, is subject to this phenomenon. Bricolage is that, when faced with new phenomena, we simply use the tools at hand to both conceptualize and respond to this new thing. One thing about these kinds of tools is that they tend to have multiple uses and sometimes unintended uses. Then they fold back on each other so that they can alter themselves because of the way bricolage works. To go back to your question of the analogy of animals, the idea is if a corporation or any organization were to use an AI system and it caused harm, again going back to our earlier conversation, one of the first things we would ask would be about the degree of autonomy and sophistication of the device. Again, as long as there is a valid argument that the device were under the control of that corporation, then the lines of responsibility will be fairly easy to draw. Ultimately it will be the corporation's responsibility. I think that would even be true if the autonomous system were fully autonomous because in corporations, and this is true of other organizations under principles of agency law, you would say that we would draw yet another analogy: the device is an agent of the principal, who is in this case the corporation, and is acting under the supervision of the principal. If that device were operating within the scope of authority it was given, if it was engaged in the work of the corporation, then any “tort” that it committed would be attributed to the corporation, which is its principal. There are exceptions to that in agency law. If an agent were to do something that was completely outside the scope of the agency, then the principal would not be liable. But for the most part, I think agency laws should be able to take care of those issues. Notice that we are no longer within the wild animal analogy, we are using a completely different legal tool to help us understand what is going on there. I think [the analogy of] wild animals attempts to capture the idea that eventually AI systems will reach the levels of sophistication where they might not be within the control of the “owner” or “master”. Then we start to wonder if maybe we need to talk about the strict liability that, even if we cannot draw straight lines between the “animal” and the “owner”, we are still going to hold the owner liable because of that relationship of ownership.

*Marcus S.*: One example that comes to mind of an autonomous system that may have committed tort is the “flash crash”…[[10]](#footnote-10). It was sort of this algorithmic recursive loop that just spun out of control. And there were a handful of powerful market makers that were pretty much to blame for this algorithm running out of control, and I do not think anyone was ever held liable.[[11]](#footnote-11) How would you square that case with this situation of an “agent” and the liability that may have been incurred?

*Mark C.*: In that situation, as long as an argument can be made that human beings were using this tool, and they were responsible for deciding to use it and had some control over it, then I think what we are talking about in terms of agency law applies. This comes into play when we think of the autonomous system as being so sophisticated that it really can act on its own and make decisions. I think we would have arguments back and forth about whether those particular trades were made by a system that was completely beyond human control. And, by the way, if it were, then we can talk about agency relationships and we can even talk about using strict liability as a way of assigning responsibility. I will say though that some people have argued, and I think I might agree with this, that the law is not good at handling major disasters like the one that you are describing. Do not get me wrong, we do try to use the law to address major catastrophes like that, but I think we can agree that oftentimes the law does not do a very good job of it. Usually the results are very binary in some ways. For example, we could try to hold those traders, the human beings that were responsible for that crash, liable. But then what would that mean? Maybe we could argue that yes, there could be some fines, maybe even imprisonment, and yet I think we would have a sense that this would not quite be a satisfactory response to the massive damage that was caused, right? And this just shows some of the limitations of the law to provide remedies with major phenomena like that.

*Marcus S.*: I suppose if you imprisoned every trader that trained a bot and lost a lot of money, there would be a lot of traders in jail.

*Mark C.*: Particularly when we talk about criminal law, we must talk about the particular mens rea of each human trader and all those kinds of things. I would agree that we have a strong sense of personal culpability that must be met before you impose criminal sanctions. This would be hard, and by the way, this is one of the reasons why these questions about responsibility for AI systems is so pressing, particularly in the criminal area. Especially when you talk about very sophisticated systems, there will be issues about whether you could hold the human users criminally liable for those harms because of the state of mind that would be required before you could hold that human individual criminally liable.

*Marcus S.*: So moving from a situation of finance to more social media, internet content, and the like, AI are often constructed from imperfect data. It is hard to get high-quality data. Whether we are talking about criminal justice data or healthcare data, there are these racial, ethnic, gender, etc. biases that are baked into that dataset. So when you train an AI on that data, it becomes “garbage in, garbage out”, and you have an AI that is also replicating and distributing these types of biases. In a situation like that, especially when a company knows the kind of imperfect data they are feeding into an AI, what is their responsibility for making sure that an AI does not replicate these biases?

*Mark C.*: I think that if a company or any user is aware of the limitations of the systems that they are using, particularly when it comes to momentous decisions like law enforcement decisions or the decision to give credit or to employ someone, and if you are aware that the devices you are using are flawed, then I think you do have a responsibility to, at a minimum, make sure there become other ways in which you reach these kinds of decisions. You might consider the “recommendation” or prediction made by an AI system, but you would automatically know that you have got to double-check that, because of the limitations of the data for which that model was trained. The state of the art is still grappling with how we really govern these systems. I think most would agree that you want a system of checks and balances so that you have internally within an organization human beings who are looking at and auditing, monitoring the "decisions" made based on the AI system, and allowing other folks to assess those results for fairness. If you are using AI applications to, say, sift through resumes, it is just incumbent upon that company to check those results and question whether they are noticing biases for who is being invited for second interviews, and things like that.[[12]](#footnote-12)

*Marcus S.*: Building on that, in medical applications especially it seems like there is the potential for these biases to go far awry. I totally agree that it is beholden upon the company to ensure that they are not replicating biases when they are aware of their imperfect data. However, are you familiar with federated learning?

*Mark C.*: To some extent, yes.

*Marcus S.*: Okay, the concept being, let us say with GDPR compliance[[13]](#footnote-13) I have a healthcare system that covers a lot of EU countries. But medical data from Switzerland cannot interact with medical data from Germany, which cannot interact with medical data from Poland, etc. So the federated learning process is where every healthcare system takes their data, trains a model, and then those models all get pooled. A decision-making agent can be represented by an ensemble by a lot different healthcare companies that cannot share their data, but can share their models. From a security standpoint, I have a lot of qualms with this type of setup.[[14]](#footnote-14) But from a legal standpoint, it also seems troubling where I might not know about the kind of biases that occur from Spain, but suddenly I am incurring those biases on people in Poland. What sort of oversight would you consider in a situation like that, where individual entities do not even have access to the data, and they only have access to this [ensemble] model?

*Mark C.*: You can imagine why folks are moving to that model, and some of the advantages of that kind of a model. Because in a sense you are arguing that you can pool data from many sources, and that adds to the diversity of the pool. But as you pointed out, then you are also importing these biases that might exist in that country, of which you may be unaware. I do think that is where the law stands as a kind of backstop. Say, for example, a company decides to use a model based on a federated training process. And unbeknownst to them is importing inherited biases, if you want to call it that, from another country. You still must meet the requirements in the sense of the law, if it leads to discriminatory results, for example. You would still have to meet certain kinds of standards within that country, with regards not only to fairness but to safety and efficacy. Also, this goes to education, the idea is that as a company that is using this, as a chief technical officer you ought to know that the system you are using or you have purchased has been trained in this way, and hence has this potential flaw. Again, it speaks to the need for several pairs of eyes monitoring the outcomes of these decisions once these applications are used. Does that make sense?

*Marcus S.*: That does make sense. Although in, say, a medical context, oftentimes these AI systems are imported to distance a human being from the decision-making process, and to remove the subjectivity they might have for a particular situation. It sounds like your recommendation for avoiding these biases is not to just have one human, but to have three or four humans that are investigating the decisions an AI makes. Is it maybe more of an aggregate type of analysis that must be done, versus a case-by-case assessment?

*Mark C.*: Yes, that's right. I guess I would start off with the aggregate because if you start to see trends that are systematic, then that is probably the only way you are going to be able to detect some of these things. If you start to see systematic biases or systematic decisions which lead to unwanted health outcomes or treatment decisions that turn out to be harmful, then you know something is wrong. I will be honest though, ideally you would want oversight to be fine-grained to the point where you could have that [model prediction] evaluated at a decision-by-decision basis. But then you are quite right, I think you would start to run into these pragmatic problems. It might not be workable or possible to do that.

*Marcus S.*: Right. It seems a tough balance to strike there too.

*Mark C.*: That's right. I mean, you could spit out hypotheticals when you are using a system that is assisting with diagnoses, and, not unbeknownst to you, the system has been trained on a dataset which will leave out a particular kind of group or does not consider a particular characteristic of a group that would point to another diagnosis, etc. You are right that on an individual level this would be hard to detect. But I would think that if you are looking at the data in the aggregate, you will start to see trends.

*Marcus S.*: That's true. You know, with federated learning, I think there are still a lot of practical issues we are still trying to keep up with. For example, if you had a medical facility in Africa where you are getting diagnoses of sickle cell anemia, which is fairly unique to people of African descent, you might start diagnosing people in Norway with sickle cell anemia because of this class imbalance; that is, the classes being different types of individuals in a particular healthcare system.

*Mark C.*: Indeed. But again, I would go back to education. You and I are having this conversation and are identifying that potential problem, and would you agree that most developers are aware of that problem? Then it is about educating users and getting users to also know that this is not magic.

*Marcus S.*: Yes, agreed. I think there has been so much hype and so many people coming into the space that I would be more on the side of making sure people understand the limitations of these systems. And you are right, it's not magic, it's statistics.

*Mark C.*: That's right. I don't know if this is the time for it, but a developer mentioned the idea for whether there is a difference between the probabilistic world in which AI operates versus the world in which you and I operate in. And then that goes to how we make our decisions. I don't even know if we can describe how we reach our decisions. Is it Bayesian? I don't know.

*Marcus S.*: Actually, Bertrand Russell wrote a good piece called “Why I Am Not A Bayesian” and it goes to this issue.[[15]](#footnote-15)

*Mark C.*: I would love to read that.

**The Continuum of Granting AI Legal Personhood**

*Marcus S.*: Looking at the kind of autonomous systems that we have discussed so far, one argument you make in your book is that autonomy exists on a continuum. And the degree of legal oversight and legal liability should change depending on where you are on the continuum. You state, and I thought this was very clever, that a mouse trap is an autonomous system, and you do not really have any legal ramifications around how it behaves. But do you think that there is a point at which AI qualifies even for personhood and therefore incurs its own legal liability? And if so, what criteria would you use to determine if this point has been crossed?

*Mark C.*: I would start by saying that the law has granted legal personhood to “things”. Ships, for example, have legal personhood as well as corporations. Those are done for almost purely pragmatic reasons, just as a way of enabling creditors and those who are harmed to be able to bring [legal] actions. In the case of ships, [brought against] the ship itself and its cargo, that sort of thing. And then corporations, like these entities, can engage in contracts on its own. Then we have found it helpful to give it legal personhood. In the case of corporations, to give limited liability to its owners because we think it is useful for society to have these kinds of legal persons. Several commentators in the law have proposed that for AI. Because of these problems in liability, as a matter of pragmatics or convenience, why not give legal personhood to some of these entities, just so it is easier for people harmed to simply make claims against these systems? And then once you do that, you could have the autonomous system take out its own insurance policy and create assets for itself so that people could make claims against it. The idea would be that the human or the organization that originally puts that AI system into operation can give initial funding to the new AI "person" for those claims. But then of course you would want to be able to sever the link between the humans who funded this device, or else it does not make sense. But as soon as you do that, you almost give the system a kind of quasi-property, right? So not only does an entity not only get these sorts of responsibilities, but maybe some legal rights as well. Then we are taking baby steps in the direction of personhood. There are lots of debates in the literature about whether we will ever reach the point where AI should be considered its own person as a moral entity worthy of moral concern. That does raise the philosophical questions that you and I were talking about earlier as to why we give ourselves moral concern. What makes us worthy of moral patiency[[16]](#footnote-16)? Then we have arguments about whether AI has reached the level that is common to what we possess or have in terms of our abilities. Some people have argued that any device which can collect some concentration of information is worthy of some degree of moral patiency or thought. The idea being that most of us would have some qualms with tearing up a book, the argument being that the reason why we have that feeling is [based on] its stored information. But it is not as if those [feelings] would override our duties to human beings, etc. I was reading some research where people are trying to create rough analogues to emotion, and the machine equivalents of pain.[[17]](#footnote-17) And as soon as you start going in that direction, you can imagine there are going to be arguments that you are making the AI like people. Then we start to want to treat them as people. But that seems very far in the future to me.

*Marcus S.*: Right. I have been skeptical in the past that AI could ever reach that point. I learned about Searle's “Chinese Room”[[18]](#footnote-18) not long ago, which really made my step back and reconsider some of those positions. I am not quite sure where I fall right now. One point you make in the book is that our legal system is just implicitly anthropocentric: centered on humans, for humans. Do you think that in the future we will need to take steps to include autonomous machines and AI when we go forward codifying such laws to make it more inclusive?

*Mark C.*: I think that goes to whether we are ever going to achieve general intelligence[[19]](#footnote-19). I am now completely speculating. I wonder if we would want to give personhood in the full sense to something that did not have that general intelligence. But as soon as I say that, many of us would argue that we have ethical duties towards, say, animals, which supposedly do not share our level of intelligence. So maybe I would start off by saying if we ever reach the level of AI that truly is a general intelligence or “just like us”, you could argue that the law does not need to be changed much at all. We could simply recognize them as being humans, and then toggle back and forth about intermediate steps, and what changes are necessary. Then we would make decisions like granting personhood to AIs just like we do to corporations, but they do not have the same panoply of rights that human individuals have. And then you would imagine that ethicists would start to ask the question “Is that ethically sound when we do that? And are there fairness issues?”, particularly (and now we are being truly speculative) you have reached a level where the AI arguably starts to have sentience. Some people argue that therefore we should not develop AI to get to that level, so that we don't have to answer these very difficult questions.[[20]](#footnote-20)

*Marcus S.*: I would imagine that if science is any indication, that this concern will not register with a lot of people building these agents.

*Mark C.*: I feel like that goes to a much broader conversation we could have later about the governance of technology. You have sense that technology is inevitable, that it is deterministic in some ways. On the other hand, we do have laws that, for example, that ban human cloning. Granted, there might be folks somewhere trying to do that, but it feels like a strong prohibition that most folks heed.So maybe you could make that kind of prohibition for developing a general intelligence. But you are right, Marcus, I guess it does not feel scary enough, or as august as, say, human cloning.

*Marcus S.*: I think that most of my fellow practitioners do not have strong considerations for that possibility.

*Mark C.*: I think they want to see if it can be done, right?

*Marcus S.*: Right!

**Can Smart Contracts Be Legally Binding?**

*Marcus S.*: I have one more question for you…[can] smart contracts[[21]](#footnote-21) issued through a blockchain be legally binding? And is the absolute permanence of a smart contract more of an asset or a liability?

*Mark C.*: I would start by saying that the law already permits electronic contracting. In the United States this is the concept of an “electronic agent”. You might be aware that “automatic contracting” already takes place in cases of inventory and those kinds of things. But they tend to be in very circumscript situations where this fits in.[[22]](#footnote-22) This is because we are talking about a technology that has been developed long before we ended the AI winter. The concept already exists. Now you translate it to blockchain, I have read where folks are experimenting with creating contracts in that way. I do not personally see any inherent difficulties with forming [smart contracts], so long as if there is a breach of that contract, that there are some kinds of remedies. You can make an argument that it maybe goes to automatic execution of certain kinds of contractual obligations. You program the system to make these kinds of contract decisions to purchase or to sell, then it becomes about [what happens if] the other side does not meet their obligation? I think that sometimes the blockchain would help in determining that. Then it is a question of whether we want contracts to automatically police themselves. Maybe sometimes the human parties would want that to happen, sometimes they don't.

*Marcus S.*: Sure. For example, if I put my will on the blockchain, how does the blockchain know if I am dead or alive?

*Mark C.*: Right. I guess that goes to a deeper question of how autonomous we want these devices to be. Because what I hear from you is that you need inputs from the real world for some of this to work.

*Marcus S.*: Right. So if I put my will on the blockchain as this immutable thing at a point in time, then I have to trust it is legally binding at the point of execution. Are we at the point where a smart contract can be legally binding?

*Mark C.*: Again, it would key on whether human beings are using that modality to create a contract, is my view. I have not looked to see if there is specific legislation which allows for blockchain contracts, so I want to do some more research on that. But I do not see anything that would inherently prevent you from doing that. But it is key to the human being, and it still does not solve all problems. To go back to your illustration of the will (by the way, trusts and wills is not my area), as you know, trusts and wills can be contested anyway no matter how they are memorialized. I would imagine the law would continue to make it possible to sometimes essentially change the wishes of the person who created the will even though via a blockchain it is immutable. You could trace everything that you put into that will, every change made. It would never disappear. But that does not really matter, because you could have some piece of parchment that contains that will, and the law could still decide that we are not going to honor the wishes of the attesting, or that we are going to honor the wishes.

*Marcus S.*: One thing that people really talk about as a feature of smart contracts is that they are effectively immutable. I put my contract on a block, I keep moving down the chain and so long as I get far enough away from that block it gets effectively permanent, unless you do some serious undoing, and it costs a lot of people a lot of money. Is that immutability really a bug or a feature to the legal contract process?

*Mark C.*: It would really depend on what we mean by “immutable”, because my understanding is that yes, if you put something on the chain, everything you put on the chain is immutable. But that doesn't mean that you can't add blocks. Sometimes those kinds of contracts that cannot be amended are arguably helpful, but I will be honest, most contracts parties probably want to reserve the option to amend that contract. So I am not seeing examples of where you want a contract that you could not change. I certainly can understand a contract that addresses anti-forgery or anti-fraud devices, where you want to make the “document”, or whatever is represented on the blockchain, immutable. But I cannot imagine why you would want to make it so you cannot change your agreement with your human counterparts.

*Marcus S.*: Right. You know, it is a little ironic to be thinking of smart contracts and cryptocurrencies as offering an anti-fraud device. Especially with automatic execution, there is really no way to walk back that smart contract once it is in a block.

*Mark C.*: It is true. And don’t get me wrong, there is now a whole area of the law that is focused on this. There are parts of automatic execution that are enormously helpful, but that goes again to these kinds of standardized transactions where there is not going to be a lot of argument back and forth as to the decision to purchase or the decision to sell. There is not a lot of anticipation that there will be times where you really want to stop the execution. But sometimes that might happen. I do not know what happened to some of those contracts, for example, when the pandemic hit and supply chains were disrupted. Imagine you had a supply contract that was on a blockchain and could not account for the pandemic and all the disruptions, and so were just automatically executing. I bet that if that happened, some parties must have gotten hurt if they were in those kinds of contracts.

*Marcus S.*: I think some parties did.[[23]](#footnote-23)

**On *The International Governance of Artificial Intelligence***

*Marcus S.*: I wanted to give you some time to talk a bit about your new book.

*Mark C.*: Yes, thank you. I will just say a little bit about it. I just turned in the manuscript last week, it is called *The International Governance of Artificial Intelligence*. The study is about arguments in the emerging level of AI and is a combination of hard and soft norms for its governance. What I do is look at the stakeholders in the development of those norms and the sources of loss. So [I look to] private firms, large technology companies in particular, programmers and academics, nation-states, and international organizations, and state legislation as a source of international law, and the private softer norms that companies adopt like codes of ethics. The basic thrust of the book is showing the interactions of those different sources and how those sources of law are given rise to, what might we call its international governance. I think it is nascent since there are no strong overarching set of rules at the international level that govern AI, but is emerging. The other thing I would say is that I conclude by saying that the overarching set of norms for AI should come from international human rights. I am certainly not alone in making that argument. I address some of the objections that are raised for using that, but still come out in favor of using those as a way of understanding, at a minimum, among nations of a common language, where we can conceptualize and debate the impacts of some of these on AI applications that are beginning to have international effects. We expect to have the book out by late spring to early summer of 2023.

**Conclusion**

I learned quite a lot from my discussion with Professor Chinen, especially with regards to the legal ramifications of autonomous systems and AI in the world of commerce, finance, healthcare, and beyond. I am looking forward to his new book, and hopefully we can interview him again around the time of its publication.

**Acknowledgments**

A big thanks to Professor Mark Chinen for taking the time to speak with me. Thanks also to Emily Wenger for checking interview questions and proofreading this document.

**References**

[1] Monarch, Robert Munro. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster, 2021.

[2] Case, Nicky. "How to become a centaur." *Journal of Design and Science* (2018).

[3] Stanton, Neville A., and Paul M. Salmon. "Human error taxonomies applied to driving: A generic driver error taxonomy and its implications for intelligent transport systems." *Safety Science* 47, no. 2 (2009): 227-237.

[4] Neal E. Boudette, “Federal safety agency expands its investigation of Tesla’s Autopilot system”, *The New York Times*, 2022.

[5] Nyholm, Sven, and Jilles Smids. "The ethics of accident-algorithms for self-driving cars: An applied trolley problem?." *Ethical theory and moral practice* 19, no. 5 (2016): 1275-1289.

[6] Navarrete, C. David, Melissa M. McDonald, Michael L. Mott, and Benjamin Asher. "Virtual morality: Emotion and action in a simulated three-dimensional “trolley problem”." *Emotion* 12, no. 2 (2012): 364.

[7] Rai, Tage S., and Keith J. Holyoak. "Moral principles or consumer preferences? Alternative framings of the trolley problem." *Cognitive Science* 34, no. 2 (2010): 311-321.

[8] Balkin, Jack M. *Cultural software: A theory of ideology*. Yale University Press, 1998.

[9] Lévi-Strauss, Claude. *The savage mind*. University of Chicago Press, 1966.

[10] Kirilenko, Andrei, Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun. "The flash crash: High‐frequency trading in an electronic market." *The Journal of Finance* 72, no. 3 (2017): 967-998.

[11] Michael Tarm, “Autistic futures trader who triggered crash spared prison”, *The Associated Press*, 2020.

[12] Deshpande, Ketki V., Shimei Pan, and James R. Foulds. "Mitigating demographic Bias in AI-based resume filtering." In *Adjunct publication of the 28th ACM conference on user modeling, adaptation and personalization*, pp. 268-275. 2020.

[13] Zekos, Georgios I. "Digital Politics, GDPR, and AI." In *Political, Economic and Legal Effects of Artificial Intelligence*, pp. 473-511. Springer, Cham, 2022.

[14] Boenisch, Franziska, Adam Dziedzic, Roei Schuster, Ali Shahin Shamsabadi, Ilia Shumailov, and Nicolas Papernot. "When the curious abandon honesty: Federated learning is not private." *arXiv preprint arXiv:2112.02918* (2021).

[15] Russell, Bertrand. *The scientific outlook*. Routledge, 2009.

[16] Glymour, Clark. "Why I am not a Bayesian." In *Readings in Formal Epistemology*, pp. 131-151. Springer, Cham, 2016.

[17] Bowers, Jeffrey S., and Colin J. Davis. "Bayesian just-so stories in psychology and neuroscience." *Psychological bulletin* 138, no. 3 (2012): 389.

[18] Pouget, Alexandre, Jeffrey M. Beck, Wei Ji Ma, and Peter E. Latham. "Probabilistic brains: knowns and unknowns." *Nature neuroscience* 16, no. 9 (2013): 1170-1178.

[19] Kuehn, Johannes, and Sami Haddadin. "An artificial robot nervous system to teach robots how to feel pain and reflexively react to potentially damaging contacts." *IEEE Robotics and Automation Letters* 2, no. 1 (2016): 72-79.

[20] Searle, John R. "Minds, brains, and programs." *Behavioral and brain sciences* 3, no. 3 (1980): 417-424.

[21] Bostrom, Nick. *Superintelligence*. Dunod, 2017.

[22] Buterin, Vitalik. "A next-generation smart contract and decentralized application platform." *white paper* 3, no. 37 (2014): 2-1.

[23] Schmidt-Kessen, Maria José, Helen Eenmaa, and Maya Mitre. "Machines that make and keep promises-Lessons for contract automation from algorithmic trading on financial markets." *Computer Law & Security Review* 46 (2022): 105717.

[24] Queiroz, Maciel M., Renato Telles, and Silvia H. Bonilla. "Blockchain and supply chain management integration: a systematic review of the literature." *Supply Chain Management: An International Journal* (2019).

1. For more information on this topic, the de-facto resource on this type of machine learning construction is Robert Monarch’s book *Human-In-The-Loop Machine Learning: Active learning and annotation for human-centered AI* [1]. [↑](#footnote-ref-1)
2. These systems are often called “centaurs”, not “chimeras” [2]. I believe I conflated my Greek mythological creatures during our conversation. [↑](#footnote-ref-2)
3. Empirical evidence suggests that around 3 in 4 automobile accidents are caused by human error [3]. [↑](#footnote-ref-3)
4. This report was filed in June 2022 and, as of this writing, the investigation is still ongoing [4]. [↑](#footnote-ref-4)
5. The trolley problem and its variants are a set of ethical dilemmas dealing with whether someone should act to sacrifice the good of one for the good of many. In the classic setup posed in the mid-1970s, a runaway trolley is on course to collide with and kill five people standing on the tracks. A bystander can pull a lever and divert the trolley to another line, where a single person is standing on the tracks and would instead be killed. These sorts of dilemmas and variants have been posed in many domains, including the case of AVs [5]. [↑](#footnote-ref-5)
6. While it is clearly unethical to test true life-and-death trolley problem dilemmas, much work has been done by cognitive scientists and psychologists to determine our revealed preferences when faced with moral dilemmas of this sort [6,7]. [↑](#footnote-ref-6)
7. Would you like to be the person to decide the outcomes of trolley problems? If you would like some practice honing your ethical intuition and have some fun along the way, I would recommend [this witty browser game](https://neal.fun/absurd-trolley-problems/) by Neal.fun. [↑](#footnote-ref-7)
8. Balkin describes these concepts in his books, but prominently in his 1998 publication *Cultural software: A theory of ideology* [8]. [↑](#footnote-ref-8)
9. Lèvi-Strauss introduced the concept of bricolage in his 1966 publication *The Savage Mind* [9]. [↑](#footnote-ref-9)
10. During the interview, I failed on my exact recall of the incidents surrounding the 2010 flash crash, so I will describe it briefly. On May 6, 2010, for around 36 minutes, the securities across United States exchanges dropped sharply by roughly a trillion dollars before rebounding. While the exact causes are contested, one agreed-upon point of failure was an error in many high-frequency trading algorithms, which compounded and over-magnified existing market sentiment. For further explanation, see Kirilenko et. al. [10]. [↑](#footnote-ref-10)
11. It should be noted that the London-based trader Navinder Singh Sarao faced criminal sanctions for his role in market manipulations that partially brought about the 2010 flash crash. He was sentenced in 2020 to one year of home confinement with no jail time [11]. [↑](#footnote-ref-11)
12. Researchers are already finding ways to identify biases in the recruitment process of AI, and finding clever ways to mitigate them [12]. [↑](#footnote-ref-12)
13. Since the passage of the EU General Data Protection Regulation (GDPR), effective in 2018, many questions remain to be answered about the role AI can play within this legal framework [13]. [↑](#footnote-ref-13)
14. My favorite paper regarding the security concerns arising from federated learning models is the clever work of Boenisch et. al. [14]. [↑](#footnote-ref-14)
15. I believe I was conflating a few sources here. Bertrand Russell was a proponent of Bayesian epistemology [15], whereas Clark Glymour, an opponent of Bayesian epistemology, wrote the article “Why I Am Not A Bayesian” [16]. Neither really addresses Professor Chinen’s question about whether our decision-making strategies are Bayesian. There does not seem to be a clear consensus on this topic [17,18]. [↑](#footnote-ref-15)
16. Moral patiency refers to the concept that we have some moral obligation towards some other being. If we owe a moral obligation to a being, then that being is known as moral patient. [↑](#footnote-ref-16)
17. Work on constructing AI to feel human emotions and pain is still very much in its infancy. One recent step has been a robotic arm that is programmed similarly to a human nervous system, such that it can “reflexively respond to potentially damaging contacts” [19]. [↑](#footnote-ref-17)
18. John Searle’s “Chinese Room” is a thought experiment where a human who does not speak Chinese is placed in a room with a book of “if-then” types of lookup statements in correct Chinese. A Chinese writer then communicates with the human in the room by passing notes. The human in the box takes statements from the Chinese writer, matches characters, and uses their given lookup statements to return a response. While the human in the room has no understanding of Chinese, the external communicator has no way of knowing this. Now instead of a human in the room, consider a machine translator. It may be carrying out similar operations, but does it have any real “understanding” of the Chinese language? Searle argues that there is effectively no discernable difference between the human in the room versus the machine in the room [20]. [↑](#footnote-ref-18)
19. Artificial general intelligence, or AGI, are considered very different from traditional AI. Today’s AI systems are trained on specific tasks and are generally baffled by anything outside their very narrow scope of expertise. An AGI is a machine that, in theory, has no domain-specific limitations. [↑](#footnote-ref-19)
20. For an in-depth discussion of arguments for and against attempting to develop AGI, see Bostrom’s *Superintelligence* [21]. [↑](#footnote-ref-20)
21. For those unfamiliar, a smart contract is a computer program or a transaction protocol that is intended to automatically execute, control or document legally relevant events and actions according to the terms of a contract or an agreement. Smart contracts, as described by Vitalik Buterin in his famous white paper [22], can be placed on a blockchain for automatic execution. [↑](#footnote-ref-21)
22. As Professor Chinen states, the extent of automatic contracting has been limited to cases such as inventory management and algorithmic trading [23]. It seems unlikely that automatic contracting is desirable outside such specific situations. [↑](#footnote-ref-22)
23. In the past few years, there have been several well-cited papers on the intersection of supply chains and blockchains [24]. [↑](#footnote-ref-23)