AI ETHICS

AI's Interference with the Judicial Process and Governance Approaches to Regulate AI

By Jane Loo, Research Associate and Adjunct Faculty, Singapore Management University and Willow Wong, Research Associate, Centre for AI and Data Governance, Singapore Management University

This module contains two parts. The first half of the lecture aims to examine the impact of AI on the judicial process and on an accused's due process rights. Specifically, it will look at how predictive algorithms are used in the criminal justice system and debate whether the use of such algorithms improve or worsen the fairness of the criminal process. In the second half of the lecture, a cross-cultural examination is conducted to see how AI governance initiatives in Europe and China seek to regulate the complex challenges posed by AI technologies.

Lecture Transcript

Loo 0:00 Hi everyone. Welcome to this module on AI's interference with the Judicial Process and Current Regulatory Approaches. My name is Jane. I'm an adjunct faculty and research associate at Singapore Management University. My background is in law and human rights, and my research interests include examining how technology influences the criminal process, the rule of law, and the respectful fundamental rights and liberties. Together with my colleague, Willow, we will give a short lecture today on how predictive algorithms could potentially impact an accused's due process rights. And then we will consider the current governance and regulatory approaches in place. I will just quickly pass it on to Willow, who will also introduce herself.

Wong 0:46 Thanks, Jane. Hi, everyone. I'm Willow Wong, adjunct faculty and research associate at the Center for AI and Data Governance, Singapore Management University. My academic background is in philosophy and one of my primary research interests is in AI ethics. In our work as researchers, Jane and I often think about the challenges in AI regulation, and we explore questions of how ethics and law can be used to steer the overall path of AI innovation. And today, I look forward to sharing my perspectives on cross cultural AI governance approaches that are used to regulate AI technologies.

Loo 1:31 Thanks Willow. As you can see, this lecture will be divided into two main sections. In the first part, we will discuss the potential impact of AI on an accused person's due process rights and we will address this question by first defining what due process is. We need to also consider why due process is so important, and what it protects in the criminal law context. We will then move on to determine the type of due process rights that we should be concerned with in this digitized age. This includes the right to a fair trial, the presumption of innocence, the right to a timely trial, equality of arms and so on. Turning to how AI interferes with the judicial process specifically, we will consider the use of predictive algorithms by the courts, and determine how it interferes with an accused's due process rights. I will point out three reasons here. That is, how predictive algorithms interfere with the nature of judges discretion, the uncontrollable influence and power of private actors, and finally, how

Transcribed by https://otter.ai

light cases are treated too alike when AI is used in judicial determinations. After which we will look at some practical suggestions on how to safeguard an accused due process rights. Principles of accountability, transparency and fairness are important here. I will then pass the floor over to Willow who will walk you through the second half of the class. She will go over some AI governance and regulatory approaches, which would be critical in ensuring that an accused due process rights are safeguarded in the digital age. Let me quickly pass you over to Willow.

Wong 3:16 That's right, Jane. So in the second half of this short lecture, we will look at two major players in the AI ecosystem, namely Europe and China, to compare the AI governance initiatives seen across the two jurisdictions. This includes the GDPR and the draft AI Act from the EU, alongside the Cybersecurity Law in China. We will also identify their strengths and drawbacks in addressing technical, legal, ethical, and social issues posed by artificial intelligence. Lastly, we will consider what these distinct governance approaches might reveal about the cultural values unique to the European and Chinese regulatory environments.

Loo 4:10 So let's look at the first question that is what is due process? The concept of due process is a multifaceted, and perhaps even contested concept that varies from jurisdiction to jurisdiction. At its simplest, due process refers to the constitutional guarantee that a state commits to respect all the legal rights it owes to a person. That means that all legal matters must be resolved in accordance with established rules, principles and procedures in the country. It is possible to speak of due process in both procedural and substantive terms. Substantive due process determines whether the state has violated constitutional protections and it asks the question of whether any such deprivation of constitutional protection can be justified. On the other hand, procedural due process seeks to question the process, and method of government action used to convict defendants accused of a crime. Essentially, it asks the question of whether the state has followed proper procedure when it infringes on a constitutional right. Due process is significant in both the pre trial and trial stage, meaning to say that if evidence used to convict a defendant had been obtained improperly during the early investigation stage, such as when a defendant is coerced or manipulated to say certain things during his interview process, a defendant's due process rights would be infringed. That is to say that an accused person has due process rights, even prior to coming to court.

Loo 5:53 We should also talk about why due process rights are important. Essentially, this boils down to how due process rights function as a check on the exercise of arbitrary state powers. Institutions of the criminal law such as policing and imprisonment grant immense coercive power and discretion to the state that could potentially threaten and infringe on individual security and liberty. For instance, police powers to exercise their powers of arrest and detention could ordinarily constitute crimes of assault and false imprisonment, if not for their authority under the law. Due process acts to limit and constrain these powers so that these enforcements and sanctions are legitimate and in accordance with the law. In short, the protections afforded by principles of due process are designed to safeguard citizens against any untrammeled powers by state authorities by limiting and guiding state discretion.

Loo 6:56 While many rights and immunities can be associated with due process, it is perhaps possible to identify a core conception of it that is inalienable in most criminal disputes. These core tenets of the process apply at some level or some form or another to all West liberal legal systems. Some of these terms will already be familiar to

Transcribed by https://otter.ai

you. I think they are quite self explanatory, so I will not go through each of them. But I will just quickly go through a few that will be relevant later.

Loo 7:28 So firstly, the right to a fair trial is one of the most fundamental due process right. It simply means that the defendant has a right to a fair and public hearing within a reasonable time before an independent, impartial and competent court. The right to a fair trial is an umbrella term that also encompasses many of the rights reflected here. The presumption of innocence, as we know, is a presumption that a defendant is to be assumed innocent until they have been proven guilty by the prosecution beyond a reasonable doubt. A timely trial is an important due process right, because justice delayed is justice denied. Equality of arms simply means that there should be a fair balance between the opportunities afforded to the defendant and prosecution in a criminal proceeding. The right to a public hearing recognizes a defendant's right to be heard openly so that judicial determinations and reasoning can be scrutinized. Consistency and certainty of decisions is another important aspect of due process, simply because like cases should be treated alike in line with the principle of fairness. Any preferential treatment and discrimination in decision making reflects a denial of justice. The independence and impartiality of the judiciary must also be secured to ensure that decisions taken are fair and free from external influences. With this understanding, we can now explore how AI could potentially infringe on due process.

Loo 9:05 For the purpose of this lecture, it will be useful for us to narrow in on a particularly controversial technology, and that is the use of predictive algorithm tools by the courts. Most of us will already be familiar with the technology since it is in use in many legal systems worldwide to make sentencing, probation and bail decisions in the criminal process. Essentially, what the softwares are able to do is to make predictions about a defendant's risk of reoffending or fleeing based on variables such as a defendant’s criminal record, their failure to attend court, crime severity and other factors deemed relevant or important by the risk prediction tool. Based on the absence or prevalence of these variables, the offender is then given a risk score from low to high. The court would then rely on these assessments to determine whether a defendant should be incarcerated for a shorter or longer duration. A high risk score notes that an offender is more likely to reoffend. Therefore, to protect the public from further crimes of the defendant, the courts may be inclined to incarcerate the offender for a longer period.

Loo 10:18 Traditionally, these risk assessment decisions are performed by law enforcement, correctional staff or clinical professionals working in prisons. These personnels will have knowledge of a number of variables about the defendant, either connected to their criminal history or sociodemographic characteristics. There are many of these tools available in the market. And one tool that is particularly notable is the COMPAS software. You can see it on the right side of the screen. But as some of you may be aware, risk prediction tools like COMPAS have been widely criticized as being inaccurate and biased. A research by ProPublica found that only 20% of the people predicted by COMPAS to commit violent crimes actually went on to do so. And more troublingly, it was captured by the researchers that the algorithm was twice as likely to falsely flag black defendants as future criminals, as it would falsely flag white defendants. Of course, ProPublica’s papers are not without its own criticisms. In other publications, other researchers have found that the COMPAS score translates into roughly the same likelihood of reoffending regardless of the defendant's race.

Loo 11:39 Yet, despite the controversy surrounding these tools, let's talk about why there is still strong support for its utilization in judicial decision making. Advocates for these tools highlight that predictive algorithms help rationalize the decision making process for a judge by summarizing information in a more efficient way than a

Transcribed by https://otter.ai

human judge. Hence, it is arguably more accurate at predicting risk, and a better balance can be struck between reducing incarceration and promoting public safety. Additionally, there is the expectation that these tools can eliminate racism and discrimination embedded in the criminal justice system by reducing individual judges discretion and any personal bias that they might have. It is hoped that the introduction of algorithms in courts can have a positive impact on the overall fairness of the judicial system. One final reason relates to the utilitarian benefits that these prediction tools might bring to the courts. It arguably helps to cut short the time taken to conduct individualized judgments on an offender's risk. In other words, important and limited resources can be redirected elsewhere in the criminal justice system.

Loo 13:00 For these reasons, predictive algorithm tools are increasingly gaining popularity worldwide among both practitioners and government institutions. Despite the advantages we just examined, there is some consensus that data used to build these systems may often be based on biased or flawed data that tend to produce discriminatory outcomes. Discrimination exists when the predicted outcome for a particular group is systematically different from other groups, and therefore one group is consistently treated differently to others. For instance, probation bills or sentencing predictions can be made based on factors that are racially biased and highly correlated with race, like the fact that one lives in government subsidized housing or had expressed negative attitudes towards the police, for example. Bias can also occur in the AI system's programming because an algorithm's design and functionality will reflect the values of its designer or creator. Any probation bills or sentencing outcomes made based on these factors would no doubt be unfair when relied on and can produce discriminatory outcomes.

Loo 14:16 Building on this discrimination issue, we will now turn to the threat that such predictive algorithms pose on an accused due process rights. We want to look at three main issues here, when it comes to how discriminatory predictive algorithms cancel or impact on an accused due process rights. The first theme that we will look at concerns how predictive algorithms interfere with a judge's discretion, and where any residual discretion is transferred to. Second, and related to judge's discretion, is the influence of private actors in the criminal process, and how judicial scrutiny is moved away from the public eye towards non-transparent, proprietary methods owned by private corporations. Finally, we want to consider the principle and law that like cases should be treated alike and question whether it is taken too far by predictive algorithms. We will examine the extent these predictive algorithms commit to this principle and at what extent, and at whose expense. All these factors combined will interfere with an accused due process rights and their overall experience with the criminal justice system. Here the absence of proper due process can mean a difference between an accused's freedom and their incarceration.

Loo 15:34 First, let us look at how predictive algorithms change the nature of judge's discretion. What is judge's discretion and why is it so important? In the common law world at least, the ability of judges to exercise their discretion is important, because there will be occasions where the law is ambiguous. Judges need to apply their discretion in order to clarify and interpret unclear laws, and sometimes even to develop new principles for these cases in order to recognize the unique factual particularities of a case. Judicial discretion needs to be applied if justice demands such a subjective, personalized application. Individualized justice here is important to secure the defendant's right to a fair trial. And in any event, judge's discretion in these special cases are not unlimited. Judges are still constrained by a defendant's due process rights and significant legal protections are still available to the accused, such as the right to judicial review.

Transcribed by https://otter.ai

Loo 16:35 So how do predictive algorithms impact on a judge's discretion? Simply, they interfere and sometimes even eliminate judge's discretion altogether from the cases that demand it. Two things are worth mentioning in this context when predictive algorithms are used in courts. First, algorithms cannot be a substitute for judicial discretion. Judicial decisions are not simply taken rationally or is always black or white, but include a consideration and balancing of emotions, intuition and a sense of justice. As we know, AI systems are not able to provide individualized justice because they work by following a set of rules. If a defendant's case has a particularity or a mitigating factor that is not covered or unknown to the algorithm, it will not be taken into account by the system. Judging involves various social considerations and human values that are at present still beyond the capacities of AI. An example of such a mitigating factor could be a defendant's young and naive age. If this particularity is a determining factor for a judge, such that they might consider the factor relevant enough to lower a defendant sentence, one can see the injustice that reliance on a predictive algorithm might cause if the defendant is denied that grace and discretion. In this scenario, maintaining judicial discretion is important to preserve the defendant's right to a fair trial. Meaning to say judicial discretion operates to ensure that a defendant sentence is justified in accordance with the justifications of punishment, and not just the motivation of incapacitation as often embedded in these machines.

Loo 18:24 Second, while acknowledging that predictive algorithms used in sentencing, bail and parole decisions remain advisory and voluntary in nature for now, as in they are only used to assist and guide judges in their decision making process and need not be followed, there is still the possibility that they can strongly influence the outcome of the case. Research in behavioral economics and cognitive sociology review that it is psychologically difficult and rare to overwrite the recommendations provided by an algorithm. In fact, judges are likely to follow the predictions provided by risk assessment tools, if the risk prediction of the algorithm is higher than the one a judge has in mind. Thus they might increase the sentence of an accused without realizing that they are following the algorithm's recommendation. In this case, judicial discretion is not only lost, but the use of these predictive algorithms may also interfere with the independence and impartiality of the judiciary. Judge's disposition to favor and abide by decisions made by these machines would contribute to the emergence of de facto AI judges endangering the independence of the court. The fact that judicial discretion is interfered with or even lost, but is still claimed to be exercised is particularly problematic because the role and influence of AI in judicial determinations is not straightforward and transparent. In summary, by interfering and limiting a judge's discretion, predictive algorithms, especially when it produces discriminatory results, will impact on the defendant's right to a fair trial, as well as compromise the independence and impartiality of the judiciary.

Loo 20:16 Let us move on now to where this discretion flows to. The use of predictive algorithms do not simply cause a judge's discretion to dissipate into thin air. Instead, what happens is that this discretion is moved out of the judiciary to some type of marketplace. So instead of this discretion falling on a judge with authority and limitations on their power, it is redirected to private corporations and technology developers who possess the power and resources to build these machines. A wide range of predictive algorithms currently exists in the market. These algorithms are constructed by different actors and stakeholders, including governmental organizations, nonprofit organizations, and private bodies. In building these algorithms, these organizations would have to make considered choices and decisions regarding the type of datasets and testing methods they wish to preference in their model. Note that each of these actors have different resources, different priorities, and different motives for building and contributing to this analysis task. Different choices and preferences will naturally be made.

Transcribed by https://otter.ai

Unsurprisingly, the variables used to build these predictive instruments will vary widely and this means that the results provided by risk assessment tools will also differ depending on which model is eventually adopted for use.

Loo 41:43 In this particular instance, you can see for yourself how a judge's power to exercise their discretion to consider factual peculiarities in a given case is being transferred to a non state body who is at liberty to pick and choose what variables are considered important in the building of their risk prediction model. Depending on the model adopted for use by the judiciary, discretion will automatically and inevitably shift to the company in question. To give you an example, if a corporation considers that historical variables outside of the defendant's control are important, such as whether or not an offender's parents had previously been incarcerated, the discretion lies with the company to highlight that variable in the model. Whether or not a judge will take that same factor into consideration is no longer material. In short, we need to be cautious of any unintended shifts of discretion that might stem from the use of predictive algorithms.

Loo 22:43 Moving on, what happens when actors are motivated by profit to incarcerate as many individuals as possible, or when their data set and model suffer from some type of bias problem? Alternatively, even if there is no bias involved in the data set, what if the model took into account irrelevant variables, for example, whether the defendant was wearing a black or white t-shirt when they committed the offense? Courts relying on these faulty risk prediction tools would as a matter of course, also hand down faulty decisions if not scrutinized by human judgment and oversight. One can definitely see how this will interfere with a defendant's right to a fair trial, meaning also the right not to be subjected to discriminatory or irrelevant decisions taken by an AI in the determination of one's guilt. But what is more problematic here is that these algorithms are usually kept secret and proprietary. Most actors are reluctant to share the algorithm or the data used to train their model. Even when shared, there is also no guarantee that persons on the outside, lacking technical knowledge, are able to appreciate or understand the use of such data.

Loo 23:56 Most judges likely lack the technical knowledge also to understand algorithmic risk assessment tools. In the case of State and Loomis, Justice Abrahamson in her concurrence pointed out that the lack of understanding of COMPAS was a significant problem and the court needed all the help it could get. Therefore, there is the problem of even spotting the use of biased or irrelevant variables in the first place. Unsuspecting individuals would be prosecuted or denied bail and parole without even knowing or understanding the reasons why. This lack of transparency relates to the difficulty in challenging any discriminatory or incorrect decisions taken by these machines. Again, a defendant's right of access to justice, and a defendant's right to a fair trial is effectively denied.

Loo 24:49 Also, because these models are shrouded in secrecy, there is also interference with a defendant's right to a public hearing. The right to a public hearing seeks to protect litigants against the administration of justice in secret with no public scrutiny or review. This right helps to promote and maintain confidence in the courts by rendering the administration of justice visible and transparent, so that justice is not only done but is also seen to be done. The inherent lack of transparency in these predictive models when used by state actors would render the administration of justice a secret. Once again, this compromises due process guarantees.

Loo 25:32 One final point relates to the commonly invoked principle that like cases should be treated alike. This is a well established principle in common law jurisdictions and the effects of consistency, uniformity and certainty underlie the concept. Justice is said to be at stake when a person is treated less favorably or differently from

Transcribed by https://otter.ai

another person in a like position. As we know, predictive algorithms are often marketed as being able to reduce sentencing disparities, since it removes individual judges discretion or bias in risk determination. Judges are said to be nudged into following the predictions of algorithms that are based on factors that are already fixed and already deemed as relevant to an offender's risk scoring. Advocates of predictive algorithm tools argue that such quantification tools make judges more accountable for their decisions, and thereby, sentencing becomes more consistent and certain, which arguably also means that like cases are in fact treated alike.

Loo 26:34 However, two important considerations are ignored in this thinking, and we will explore this in the next slide. The first problem relates to our previous point that there are many stakeholders involved in the construction and implementation of algorithmic approaches to criminal justice. The proliferation and abundance of preferred choice in the marketplace raises questions about the fairness and consistency of the judicial system as a whole. Depending on what predictive model a court chooses to adopt in its system, prediction outcomes will indeed differ. For example, if a Birmingham court decides to adopt the XYZ model and a London court adopts the ABC prediction model, it might possibly lead to an increase in sentencing disparities within and between the two cities, in the sense, because the algorithm is inconsistent throughout different states, like cases will definitely not be treated like therefore impacting on an accused due process rights.

Loo 27:33 Secondly, we want to also address the problem of statistical generalization. Ordinarily, the use of group tendencies as a proxy for individual characteristics is in non permissible and is constitutionally forbidden discrimination. To give examples, this includes one's race or age or gender. Accused persons have the right to receive personalized justice, and not be sentenced on account of risky characteristics of a group to which they belong. However, the lack of transparency in these predictive systems means that there is a lack of information on how most algorithms operate and what type of data is being used to build these systems. Challenges have been raised in the courts on the constitutionality of such risk assessment tools, as infringing on fundamental principles of equality and fairness. Risk assessment tools based on such statistical generalization would no doubt be discriminatory in nature. Additionally, in these examples, like cases are definitely not treated alike by the nature of the facts of the case. Instead, like cases are treated alike based on discriminatory variables included in these prediction models. Again, you can see how AI can infringe on a defendant's right to a fair trial and a defendant’s right presumption of innocence.

Loo 28:57 To summarize, I have explained how predictive algorithms could potentially infringe on a defendant's due process rights in the following ways. Firstly, because it interferes with a judge's discretion. Secondly, because of how it shifts such discretionary powers to private actors. And finally, because it infringes on the principle that like cases should be treated alike. Let us move on to consider practical solutions on how to protect an accused’s due process rights before I pass you over to Willow, who will go over some AI governance approaches.

Loo 29:28 So when it comes to finding practical solutions on what can be done to protect an accused’s due process rights, a few suggestions have been raised. On the part of the state, given the very real risk of discrimination from these tools, one suggestion would obviously be to eliminate the use of AI in sentencing decisions completely. But to avoid such drastic measures being taken, the judiciary can develop some sort of guidelines that permit the use of predictive algorithms only in situations where it is absolutely necessary and can be employed with the greatest precaution, including granting the defendant an opportunity to contest and inspect the algorithm. Other researchers have suggested that the lack of understanding surrounding predictive algorithms

Transcribed by https://otter.ai

is especially problematic because it could lead to situations where judges and other legal professionals are inclined to trust the technology and will just blindly follow the outcome. On that note, there needs to be more awareness building and education about these new technologies, so individuals relying on these tools can have a better understanding of the capacities and limitations of these systems. This is important to ensure that inappropriate weight is not placed on the perceived objectivity and accuracy of AI to the detriment of an accused's due process rights.

Loo 30:52 On the part of private corporations, both moral and legal responsibility needs to be imposed, and there is a need for more transparency, explainability and accountability. Some mutually beneficial way needs to be considered for private corporations to disclose their predictive algorithms' inner workings, so that it is both possible to safeguard a defendant's due process rights, while ensuring their corporate interests are also safeguarded. In relation to how to read these prediction scores to ensure that a defendant's due process rights are respected, one solution is to ensure a collaboration between human and machine by making sure that a human judge gets to exclusively decide in uncertain or unclear cases. Human oversight is maintained in this way. Again, this is of course, easier said than done, because as we discussed earlier, a judge can easily be nudged into following the results of a predictive algorithm. In essence, if we translate these practical solutions into higher level principles, we can see that what is needed is more fairness, transparency, explainability, accountability and the requirement of the human in the loop. On the back of these principles, I will now pass you on to Willow who will discuss current AI governance approaches.

Wong 32:10 As Jane described, there are many legal, ethical and technical challenges posed by AI technologies. So how do we go about solving them? The difficult task of negotiating where to allocate limited resources, time and expertise to positively influence AI development is, in essence, the primary concern of AI governance. For some background context, the global landscape of AI governance has seen active participation from the private sector, the public sector, and NGOs or other third sector entities that seek to amplify civic voices within AI initiatives. An example would be the UN, WHO, or think tanks that generate their own research insights to influence activities in the private and public sectors. With much of the cutting edge AI research and development coming from prominent tech companies, we have also seen industry efforts to attempt to address fiercely debated issues, such as algorithmic harm posed by AI. For example, Google DeepMind's Ethics and Society Research Unit and IBM's Cloud Tools are examples of private sector initiatives aiming to monitor algorithmic bias and associated impacts on people and societies. Over in the public sector, we also see many efforts to address AI's ethical concerns through national strategies. Canada was one of the first governments to release its National AI Strategy in 2017. Since then, more AI players have established their national and regional strategies. Today, we will look at EU and China.

Wong 34:20 In Europe, existing legislation already shapes how AI is developed and deployed, such as the General Data Protection Regulation, otherwise known as the GDPR. While the GDPR does not explicitly mention AI, it establishes important boundaries on the use of data and the risks of automated decision making. Article 22 of the GDPR refers to the right not to be subjected to automated individual decision making. The actual wording is more complicated, but essentially, it implies a right for explanation for automated decision making, including profiling. The data controller will have to give meaningful information about the logic involved in those decisions and any consequences that the data subject might face due to such automated processing. As a result of the GDPR, controllers will need to design, develop and apply their algorithms in a transparent, predictable and verifiable

Transcribed by https://otter.ai

manner. However, the GDPR has limited effectiveness when it comes to addressing situations where decisions are only partially rather than fully based on automated processing of personal data. A second problematic aspect is that the criterion of having legal effects or having similarly significant effects on him/on her is open to interpretation. How is this effect measured, and by whom? Finally, the enforcement method for the GDPR is usually through fines for data breaches due to lack of compliance. There is no evidence to indicate how this type of measure encourages responsible AI innovation in a way that proactively stops harm from occurring in the first place.

Wong 36:37 Moving on, there's also the draft AI Act proposed by the European Commission, which represents a coordinated European approach to lay down legislation to regulate the human and ethical implications of AI. The draft AI Act is a world first attempt at horizontal regulation of AI systems based on risk classifications. It goes from unacceptable to high to limited and minimal risks. The draft AI Act issues hard prohibitions on unacceptable risks, while using voluntary codes of conduct to manage the minimal ones. In the draft AI Act, three categories of AI practices are banned in their entirety. This includes two on manipulation, and one on social scoring. The final category of prohibited AI practice is the use of real time and remote biometric identification systems, which is not allowed except for specific law enforcement purposes, and only if they are accompanied by an independent authorization regime for accountability purposes.

Wong 37:55 If we look at the AI Act for the specific wording of this prohibition to regulate against manipulation, it becomes quite confusing. A - the placing on the market, putting into service, or use of an AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behavior in a manner that it causes, or is likely to cause, that person, or another person, physical or psychological harm. B - the placing on the market, putting into service or use of an AI system that exploits any of the vulnerabilities of a specific group of persons due to their age, physical or mental disability, in order to materially distort the behavior of a person pertaining to that group in a manner that causes, or is likely to cause, that person, or another person, physical or psychological harm. Now these emphases are added; they are not found in the original documents but it does show the multiple components that make up these statements. Do we think the above adequately characterize manipulation? If we think about how we understand manipulation in everyday contexts, we know that tricking someone into behaving a certain way can be done in an intentional and obvious way, not just covertly or against someone's knowledge. This seems to contradict the emphasis on subliminal techniques. Furthermore, the manipulator can also trick people by exploiting vulnerabilities that go beyond age, physical or mental disability. This misses out on social and economic vulnerabilities, gender, race, or religion based characteristics that expose an individual to specific risks that lie beyond the scope of age or disability. Since the draft AI Act only refers to whether the activity is likely to cause, or actually causes, harm in the physical or psychological sense, does that also mean that manipulation by AI systems can be allowed insofar as they won't cause harm? Overall, this wording is not precise enough, which really compromises how effectively this provision of the AI Act can be implemented in practice.

Wong 40:47 As a concluding thought, before we shift over to the AI governance landscape in China, it is worth reflecting on whether the GDPR and the draft AI Act will inhibit European AI development due to its hard restrictions. Now, scholarly debate over the relationship between innovation and regulation is ongoing, and it's a very fierce one. The tension between those two goals also reveal how different nations respond to the rapidly changing landscape of technology, and how they respond to that change using governance approaches. The

Transcribed by https://otter.ai

European approach is more hardline than what we'd see otherwise in the US, which favors a freer market for businesses to operate without many government mandated restrictions, as that may scare away business activities or research developments due to red tapes. To the Europeans, though, the trade offs seem to be worth it because they value the protection of individual privacy and dignity enough to make a pact that hard regulation will actually contribute to an overall healthier market in the long run.

Wong 42:21 Over in China, AI governance is a prominent area that is fast developing. The first white paper on trustworthy AI released by China was in 2021. A clear sign of how the country is paying close attention to the global conversation on AI governance is its advancing definition of trustworthy AI, which largely conforms with international definitions. In this figure, we see that China sees trustworthiness as meaning an AI system that is reliable and controllable in the sense of the system's security and robustness against attacks. That it is transparent and controllable in terms of explaining the intentions of algorithmic decisions. That there is the presence of data protection by using risk evaluation, security monitoring, and governance approaches to protect privacy. And that there is a clear delineation of responsibility, referring to the design for system responsibility and tracing whether it is the human actors fault or the system fault, and the use of documentation plays a role here. And finally, what is featured is also diversity and inclusiveness through design for system fairness, and using comprehensive data samples. On face values, these principles are very similar to other frameworks around trustworthy AI found in other parts of the world, including the OECD definition of trustworthy AI, as having respect for human rights and privacy, for AI system that operates in a fair, transparent, explainable way that is secure and safe, and that the actors involved in AI development and AI use should remain accountable.

Wong 44:38 So, it seems on paper that there are some convergences. However in the paper by Hannah Rose Kirk et al. called "The Nuances of Confucianism in Technology Policy", they essentially argue that digital ethics in China is guided by its own distinct ideology, even if we're seeing similar language around trustworthy AI. The authors argue that the digital policy in China is shaped by the interaction between existing cultural value system of Confucianism, and the presiding political system of authoritarian rule by the Chinese Communist Party, or the CCP. So, underpinning AI governance initiatives are the Confucian values of hierarchy, stability and unity, which serve to justify and legitimize the deepening of authoritarian control and the expansion of digital surveillance. In China, when corporate actors behave unethically against public interest, the responsible management of technology is seen as within the government's responsibility to control data collection for the public good. Confuscian thinkers feared division of power believing it will lead to chaos. So China's unique legislative free environment governing digital privacy exemplifies this notion, because if governments allowed technology companies to collect and freely share personal data, this is seen as dangerous to the citizen, both on a personal safety and on national security levels. Instead, through the enforcement of government open access to any data sets, under the guise of national safety, data of citizens become centralized and unified. And by presenting centralization as necessary to protect citizens from corporate data misuse, a Confucian hierarchy is established where the government knows best.

Wong 47:16 As the Cybersecurity Law exemplifies, most privacy legislation introduced in the past decade governs the relationship between citizens and corporations with very little regulation concerning citizens and government agencies. There is not only an absence of a law protecting citizens against government internet surveillance, but laws instructing the necessity and utility of government surveillance. Well where does citizens' voices sit in this picture? We often make the assumption or sweeping statement that Chinese citizens place less

Transcribed by https://otter.ai

value on their own privacy because the collective good weighs more, or that Asian cultures don't recognize privacy as something intrinsically good, like, quote unquote, the West. But this is an overly simplistic narrative. The value of individual privacy in China is not measured in absolute, but in relational terms. So in Confucian thinking and in the Chinese tradition, the state authority adopts a paternalistic duty of care towards the population. They have the responsibility to make the right decision for the family, community and nation. So putting yourself in the shoes of the Chinese citizens, we see that it's not for the lack of care or knowledge that citizens adopt the understanding of their privacy need not be guarded from their caretakers, especially if the government's access to their personal data fits within existing hierarchies of duties, where the state exercises surveillance, or censorship, in the name of security, harmony and order for the society.

Wong 49:20 This contrasts with American perspectives, where the individual is the gatekeeper of their own rights and interests, often against the perceived over intervention by government entities whether that's on a federal or state level. Comparing these two perspectives then, they are very generalized, but we can consider the possibility that the Chinese and American citizens have very different understandings of which are the main actors endangering them versus protecting them from threat. As such, China's data protection regime is uniquely fitted to the Chinese context. Given the ideological motivation underpinning the Cybersecurity Law, AI technologies and big data solutions have become convenient tools that facilitate the rapid centralization and digitization of government-society relationships. In the case of the Chinese social credit system, the quantification and numericising of private life takes a regulatory effect that conditions people to produce a curated range of pro social behaviors judged as desirable by the state. As emerging technologies are added to the government's resources, including financial data on people's credit scores and mass surveillance video footage using facial recognition software, the social credit system allows the government to centralize the quote unquote righteousness score of an individual across the financial, social, behavioral and health domains. By collecting digital records of citizen information, such big data insights enhance the state's ability to then shape Chinese society in a way that mirrors the Confucian ideals promoting harmony and allegiance with the ruling party. In this way, we begin to understand how specific legislation, ethical frameworks or governance toolkits are actually embodying the cultural values unique to that specific regulatory environment.

Wong 51:58 To conclude our thinking, we have explored AI governance initiatives such as the GDPR and the draft AI Act from the EU and China's Cybersecurity Law. We have considered some of their strengths and weaknesses in connection with the deeper values and cultural contexts that informed why the Chinese or European or indeed American approach might take different emphases to work towards the same goal of AI safety. My sense is that AI governance now and where it will go next is hard to predict. It is an exciting space to watch. And if you'd like to learn more about the research by the Center for AI and Data Governance, please visit our website. If you have any questions, comments or feedback, please feel free to contact us via email. Thank You.

Transcribed by https://otter.ai