# The Concept of Culpability in Criminal Law and AI Systems

Roman Dremliuga[1] & Natalia Prisekina[1]

[1] School of Law, Far Eastern Federal University, Vladivostok, Russia

Correspondence: Roman Dremliuga, School of Law, Far Eastern Federal University, Vladivostok, Russia. Tel: 7-914-707-1474. E-mail: dreamluck@yandex.ru

## Abstract

This article focuses on the problems of the application of AI as a tool of crime from the perspective of the norms and principles of Criminal law. The article discusses the question of how the legal framework in the area of culpability determination could be applied to offenses committed with the use of AI. The article presents an analysis of the current state in the sphere of criminal law for both intentional and negligent offenses as well as a comparative analysis of these two forms of culpability.

Part of the work is devoted to culpability in intentional crimes. Results of analysis in the paper demonstrate that the law-enforcer and the legislator should reconsider the approach to determining culpability in the case of the application of artificial intelligence systems for committing intentional crimes. As an artificial intelligence system, in some sense, has its own designed cognition and will, courts could not rely on the traditional concept of culpability in intentional crimes, where the intent is clearly determined in accordance with the actions of the criminal.

Criminal negligence is reviewed in the article from the perspective of a developer's criminal liability. The developer is considered as a person who may influence on and anticipate harm caused by AI system that he/she created. If product developers are free from any form of criminal liability for harm caused by their products, it would lead to highly negative social consequences. The situation when a person developing AI system has to take into consideration all potential harm caused by the product also has negative social consequences. The authors conclude that the balance between these two extremums should be found. The authors conclude that the current legal framework does not conform to the goal of a culpability determination for the crime where AI is a tool.

## 1. Introduction

Artificial intelligence was created to make daily life easier. For some people with a special attitude to society, a crime is their daily life. From ancient times criminals always try to apply technologies to increase the economic benefits of crime and its efficiency. For instance, pirates use speed boats to avoid navy ships and pursue a victim. An automobile is a favorite tool of bank robbers. Drones are actively adapted by drug dealers. The Internet is the main instrument for hackers. With minor exceptions, any new technology may be applied both for legal and illegal purposes. Some technologies become more popular for criminal applications than for legal ones.

For instance, almost 100% of child pornography which is illegal in almost all states is distributed in the Internet. Another example of a technology that was designed with good intentions, however, now actively used by criminals is Darknet. In the case of Darknet percent of illegal use often overestimates legal benefits. Currently, Darknet is the main forum for illegal drug trade and distribution of stolen credit cards' numbers (Perdolizzi, 2017; Duxbury, 2017). Hypothetically, some innovations are more natural to be used for illegal purposes, although developers of such technologies were led by noble intentions. The Internet was created as an instrument that facilitates science, makes people closer, destroys racial, religious, and cultural barriers, but in practice, the net development had a lot of negative consequences, and one of them is its use by criminals. The Darknet was developed as a tool for rebels to survive in tyranny states.

The Internet and the Darknet are very indicative examples because they demonstrate that an idealistic anticipation

of future application of technology during its development may cause significant problems in the future. The Internet was distributed in the world before clear legal rules appeared. It is obvious that most problems may be avoided at initial stages of development and introduction of technology. With minor exceptions, legal rules follow the development of technologies. Thus, the law follows innovations and it is not prepared for future challenges.

There are numerous research papers dedicated to ethical (sometimes legal) issues of AI development and application. One of the suggested proposals is to implement ethical (or legal) rules into AI behavior principles as a universal remedy from harm caused by such systems. It sounds logical but as Patrick Lin fairly commented on autonomous machines *"One natural way to think about minimizing risk of harm from robots is to program them to obey our laws or follow a code of ethics. Of course, this is much easier said than done, since laws can be vague and context-sensitive, which robots may not be sophisticated enough to understand, at least in the foreseeable future."* (Lin, 2012, 9). Thus, as the law and the ethics are so complicated systems of norms with high dependence on context of application, it is very difficult to design a machine which is impossible to apply as a tool of crime.

This paper studies the issue of how application of artificial intelligence influences on conceptions of culpability in criminal law. Is it significant that a tool has its own will and freedom of making decisions? Could an autonomous system that is a black box even for the developer be a factor that affects the guilt of the accused?

## 2. Culpability in Criminal Law

In the theory of the criminal law, *mens rea* is a culpable state of mind in the individual committing the crime. There may be criminal liability only with wrongdoing attributed to a particular person (Fletcher, 1998, 82). The term "mens rea" (or culpability) implies that if the person has the appropriate mental state, he/she may be held accountable for his/her criminal action. Usually, in criminal law, there are two, three, or four modes (kinds) of culpability that depend on the particular jurisdiction. This approach to attribution of crime to a person is common in many legal systems (Fletcher, 1998, 84). For example, the United States' legal tradition knows four modes of culpability: purpose, knowledge, recklessness, and negligence that are defined in The Model Penal Code (MPC) developed by the American Law Institute (ALI) (Model Penal Code; Alexander, 2009, 23). In the criminal law of England there are three modes: intention (divided into direct and indirect), recklessness and negligence (Mohamed, 2013). Russian Criminal Code mentions the following forms of guilt: intent, indirect intent, extreme recklessness, and negligence [1]. The Chinese Penal Code describes two modes of culpability: intent and negligence[2] but in criminal law three forms of intent are applied (Mohamed, 2013). Thus, intention, negligence and some other forms of culpability are ways of committing offenses that are typically described as the internal side of the offense. It is very significant to identify what type of culpability is appropriate to a crime because the character of liability, punishment, and legal consequences depend on it.

With minor exceptions, all mentioned jurisdictions "*deconstruct mens rea into the cognitive and volitional elements*" and "*the interplay and intensity of these constitutive elements determine the degree of person's culpability*" (Mohamed, 2013, 28). In turn, Artificial intelligence is defined as a machine or software that sometimes has wide cognitive abilities (to know, understand, and think) and autonomous will. In support of this thesis we cite some authors that describe autonomous machine as "*reactive (it responds in a timely fashion to changes in the environment), self-controlling (i.e. it exercises control over its own actions and is not directly controlled by any other agent), goal-oriented (it does not simply act in response to the environment), and temporally continuous (it is a continuously running process)*" (Brożek, 2017). Thus, in theory it implies that degree of a person's culpability when artificial intelligence system is a tool of a crime may be affected by cognitive and volitional characteristics of AI.

## 3. Intentional Crimes

It is generally assumed that committing a crime with intention is worse than committing it negligently (Fletcher, 1998, 82). Thus, from the mental element point of view intentional crimes are the most dangerous types of crimes. Criminal intent (purpose) means that a criminal foresees and wills the possible negative consequences of his/her action or, if his/her actions cause the attendant negative consequences, a criminal is aware of the existence of such consequences or he/she believes or hopes that they come. For example, if a person uses a gun to kill somebody, with minor exceptions, the most of the world's population are aware that a result of pushing a trigger of a pistol and directing it towards a victim is predictable. When a person denies that he/she predicts and wills death of a

---

[1]    Articles   24-26,   The   Criminal   Code   of   the   Russian   Federation   No.   63-FZ   of   June   13,   1996, http://www.wipo.int/edocs/lexdocs/laws/en/ru/ru080en.pdf

[2] Articles 14-21 Criminal Law of the People's Republic of China, https://www.cecc.gov/resources/legal-provisions/criminal-law-of-the-peoples-republic-of-china

victim, the court may apply a conception of *virtual certainty* (Wallerstein, 2014) (this term comes from English criminal law, other jurisdictions may apply different terms). It implies that probable consequences of some action are obvious for any sane person.

In case artificial intelligence is applied as a tool of crime the question of culpability becomes more difficult. The key challenge is the capability of the artificial intelligence system to make decisions and to learn in the environment in which it operates. In some sense these features turn artificial intelligence into independent actor of crime. Indeterminism of behavior, freedom of will and own complicated cognition create difficulties in evaluation of human degree of intentionality. For example, what would happen when somebody gives command to intellectual system that is already able to understand human speech or text as well as able to learn and control its own conduct. Imagine that during its mission this system kills somebody not because of intent of an operator but as a sidekick effect. Reasonable judgment is that an operator who commands is probably not liable because he/she is not able to anticipate, he/she does not factually push the trigger, he/she does not target.

Such position could exclude *any* commanders of military robots from any forms of criminal liability for war crimes and crimes against humanity. For instance, genocide, the most severe war crime, is a serious violent act "*committed with intent to destroy, in whole or in part, a national, ethnical, racial or religious group*" (Rome Statute of the International Criminal Court, 1998, Article 6). Some artificial intelligence systems already demonstrate inclinations to racial prejudice[3] and other biases (Caliskan, 2017). Imagine that an officer gives an order to eliminate all enemies to robot equipped with AI, and the machine inheriting human biases will target only people with a specific racial feature. In terms of objective consequences, it could be regarded as a serious crime but the officer or somebody else will not be liable because genocide is only an act "*committed with intent*" (Rome Statute of the International Criminal Court, 1998, Article 6). The genocide formula when a person is guilty of negligence does not exist. This imaginative example demonstrates an idea that to detect intent of a person could be more complicated due to autonomous cognition and will of intellectual system. It sometimes could lead to situation when there is no liable person for committed offence.

As direct intent implies that a criminal recognizes high probability of negative consequences of his/her actions and wishes for their commencement, it is necessary to pay a special attention to predictability of possible negative consequences of crime committed with purpose (direct intent) mode of culpability. AI system is programed or often learned in accordance with the developer/programmer's intent. Usually a user may rely only on some information from the developer to anticipate possible behavior of AI system. Thus, the user's conceptualization of the system is based on information provided by the developer and he/she could not fully understand how system reacts on his/her inputs.

It is very difficult for the offender to foresee consequences of his actions if he applies artificial intelligence. For instance, drug trade may be realized through a deployment of autonomous machines (King, 2020). These machines could "*send and reply to email or use instant messaging (IM) to spark one-on-one conversations with hundreds of thousand or even millions of people every day, offering pornography or drugs*" (Kerr, 2005). The variety of consequences could be from AI sold nothing to it involves in drug consumption billion of people. Thus, the offender may anticipate and wish less ambitious results of his action.

AI with ability to self-learning behavior is not only unpredictable from the user's point of view but also from the developer's side. Very demonstrative example of such unpredictability is Tay-bot developed by Microsoft. This chat-bot is equipped with AI and has capability to self-learning[4]. Tay may use statements of other users to study how to participate in conversation. He could acquire some phrases written by his correspondent. This feature of the chat-bot based on AI provoked a big scandal since Tay applied some pro-nazi phrases in conversations which it was taught by users. There are no doubts that Microsoft does not purposely develop AI that would be insulting people in the Internet. Probably those users who wrote Tay such insulted answers did not have an intent to provoke the chat-bot to distribute insulting phrases. As a result, nobody is potentially culpable.

Thus, the law-enforcer and the legislator should reconsider the approach to determining culpability in the case of using artificial intelligence systems for committing intentional crimes. As artificial intelligence has own designed cognition and will, courts could not rely on traditional concept of culpability in intentional crimes, where an intent is clear in accordance with actions of the criminal. Of concern is the use of automated lethal weapon, since international criminal law has no mechanism of liability for war crimes, crimes against humanity and genocide

---

[3] AI programs exhibit racial and gender biases, research reveals, The Guardian, https://www.theguardian.com/technology/2017/apr/13/ai-programs-exhibit-racist-and-sexist-biases-research-reveals

[4] https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html; https://www.telegraph.co.uk/technology/2016/03/24/microsofts-teen-girl-ai-turns-into-a-hitler-loving-sex-robot-wit/

without intent. The global community should negotiate new rules of criminal liability in case of AI application in which it may be possible to consider the degree of culpability of the commander, the developer, the machine learning specialist and other persons involved.

## 4. Criminal Negligence

There are no doubts that a person who may influence on harm caused by AI system may be the developer (the programmer). Is it reasonable to make him/her criminally liable? If somebody develops an autonomous program or a robot to intentionally commit some crimes or(and) cause valuable harm, it is an intentional crime. However, without oblivious intent presumably developer is guilty of negligence. It implies that a person "*should be aware of a substantial and unjustifiable risk that the material element exists or will result from his conduct. The risk must be of such a nature and degree that the actor's failure to perceive it, considering the nature and purpose of his conduct and the circumstances known to him, involves a gross deviation from the standard of care that a reasonable person would observe in the actor's situation*" (Model Penal Code). The general formula for negligence in the most of jurisdictions is that a person is not aware of substantial and unjustifiable risk, but he/she ought to be aware.

An issue that arises from definition of criminal negligence given in Model Penal Code is what would be standard of care in terms of AI development and design. This area develops very fast. Does it mean that nobody is liable or that any person who does not take into consideration all potential harm caused by its product is liable? Both alternatives are doubtful. If product developers are free from criminal liability, it would lead to highly negative social consequences. There would be a lot of dangerous products with AI on the market. Despite the developer still would be civilly liable for damages caused to victims by his product, if revenue is more than amount of compensation to victims, he will continue to produce dangerous product. The situation when a person developing AI system has to take into consideration all potential harm caused by the product also has negative social consequences. Such attitude on guiltiness of the developer will stop progress in AI sphere as less people will be involved in AI development legally due to unpredictable legal risks.

Due to lack of court practice in the sphere of AI developer's liability, it is necessary to review approaches of liability of the manufacturer in common to study this issue. Factually, when the judge decides whether the developer is guilty of negligence, he/she must establish what measures the developer has taken to reduce the risks of harm to the consumers. If somebody wants the developer to be held liable, "*it can be proven in court that the company was negligent, with regard to the defects, risks, and potential hazards arising from the use of their product, then the company could also be criminally, as well as civilly, liable for the damages caused to victims by their product*" (Asaro, 2012, 170).

There are certain limitations of the developer's liability that stems from legal practice in the past. The main strategy of the developer to defend against negligence charges is a reference to industry standard (Asaro, 2012, 172) or sectoral norms. For example, self-driving car developer must follow road traffic safety requirements. *If safety requirements in binding law are violated this can constitute a criminal offence, entitling the police to intervene* (Vellinga, 2017, 885). Thus, it makes task of the developer simpler, to avoid criminal liability just produce product that follows standards.

Policy when manufacturer blindly follows standards may provoke collision of social norms. Imagine a situation when a child jumps on the road in front of a car that is under control of AI. One of the options for the car is to turn to the oncoming lane to avoid bumping the child. However, this maneuver creates other risks and violates general safety requirements because the car must drive its lane. Thus, if policy makers rely on industry standards as the base for criminal liability, a developer will be guilty of negligence in case of a violation of rules that cause negative consequences. If the car kills or seriously injures the child, it could be very socially dangerous, but developer avoids criminal liability. Standards are established to decrease potential harm in general, but intellectual systems can find more optional way to decide in a critical situation.

The problem of making decision in the sphere of autonomous driving is widely discussed. Usually it is mentioned as Trolley problem (Wu, 2020). Developer may save a lot of lives if he develops cars not strictly upholding the law and standards but being able to make decisions to minimize amount of fatalities. It is endorsed by society. As questionnaire that imitates situations on the road in which an autonomous vehicle is in dangerous situation shows that more people choose that autonomous cars must save lives than follow the law. The variety of reactions on dangerous situations provided on the mentioned resource demonstrates that even for moral beings it is difficult to choose what decision should be made. Especially when the circumstances have additional parameters of prepositive alternative victims (age, social value, gender etc.)[5]. The question still exists what a developer will

---

[5] http://moralmachine.mit.edu/

choose for his products.

Another strategy for the developer to avoid liability for crime with negligence is to notify consumers about possible risks. If the developer puts respective notification for consumers on foreseeable risk during application of the product, he/she would not be liable. Duty to notify the consumers about possible risks is described in national legislations in different ways but has similar content. The consumers have to be informed by the developer about serious risks of harm or death that the product may cause. For example, the General Product Safety Regulations 2005 issued in the United Kingdom claim that "*within the limits of his activities, a producer shall provide consumers with the relevant information to enable them—(a)to assess the risks inherent in a product throughout the normal or reasonably foreseeable period of its use, where such risks are not immediately obvious without adequate warnings, and (b)to take precautions against those risks*" (The General Product Safety Regulations, 2005, Article 7). A challenge here is not only to anticipate risk but provide appropriate information. Artificial intelligence system is software that has very complicated system of responding to environment as well as this system is dynamic in case of self-learning. Thus, it is difficult to describe such system in a way understandable to everybody. Consequently, it is difficult to develop an adequate notice to the customer.

Systems that are capable to self-learning may create additional challenges. In fact, the developer is able to predict some future risks that will emerge during self-learning by the system. But Tay example demonstrates that not all risks could be predicted for self-learning software. Moreover, some people may intentionally intervene in the process of system learning to provoke negative consequences. There is no answer how the degree of liability of a programmer and a third party would be shared. As some authors fairly comment, "*concerning the knowledge threshold, in some cases the mens rea could actually be missing entirely. The potential absence of a knowledge-based mens rea is due to the fact that, even if it is understood that an autonomous agent can perform the actus reus autonomously, the complexity of the autonomous agent's programming makes it possible that the designer, developer, or deployer (i.e., a human agent) will neither know nor predict the autonomous agent's criminal act or omission*" (King, 2020).

If we punish unguilty person or punishment is more severe than a level of criminal's participation deserves, it could provoke a lot of social concern. For example, as Fenwick fairly underlines, "*punishment without fault or, alternatively, punishment that is disproportionate to the degree of fault is correctly regarded as inappropriate and unjust*" (Fenwick, 2009, 113). It implies that the court faces a new challenge. If an offender does not recognize all risks in case of AI application but be punished as he does, it would be considered as unfair. If the owner, the commander, the designer is strictly liable for all acts of AI that they own it would be also unfair. As well as it is unfair, if the owner (the commander, the designer) avoids criminal liability just due to AI application. Thus, it is necessary to find balance between two extremes.

## 5. Conclusion

The aim of the paper is to highlight key legal considerations related to the AI and conception of culpability in criminal law. The result of the analysis does not imply that we need to stop development of this promising type of technology. This brief overview demonstrates that culpability issue becomes more complicated due to AI. It just indicates that not simple questions arise in that sphere. Of course, this paper has more questions than answers, but the authors hope that it provokes future studies.

Nevertheless, an obvious proposal to any legislator is to reconsider the conventional concept of culpability in criminal law. Probably, new criteria for assessment of the degree of culpability of a person in case AI applied as a tool of a crime may be developed. Such new culpability paradigm could not be universal since the new conception has to comply with a national legal tradition. In spite of solid variety of legal traditions, developed conception will have some general points. The common challenge here is that a person who is in charge of serious consequences has to be imposed with a relevant degree of liability as well as an innocent person could be recognized as liable. Moreover, new rules of culpability assessment have to be designed in respect of sufficient features of AI technologies. For instance, the new legal concept has to take into consideration the fact that AI has in some sense its own cognition and will.

The main challenge which could not be overcome by the traditional theory of criminal law is that the artificial intelligence system is a "gun" with its own cognition and will. That very sophisticated gun may learn and be developed by itself and thus, it shoots in different targets depending on input signals and its experience. Often its behavior is so complicated that could not be predicted by the developer. If a criminal uses AI as a tool of crime, it is not obvious to detect level of awareness, degree of negligence and other features of the criminal's attitude to negative consequences of his/her actions.

## References

Alexander, L., Ferzan, K. K., & Morse, S. J. (2009). *Crime and culpability a theory of criminal law*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511804595

Asaro, P. M. (2012). A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics*. MIT Press, London.

Brożek, B., & Jakubiec, M. (2017). On the legal responsibility of autonomous machines. *Artificial Intelligence and Law, 25*(3), 293-304. https://doi.org/10.1007/s10506-017-9207-8

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183-186. https://doi.org/10.1126/science.aal4230

Dremliuga, R. I., Mamychev, A. Yu., & Dremliuga, O. A. (2019) Crimes in Virtual Reality. *Dilemas Contemporaneos-Educacion Politica Y Valores, 7*(1).

Dremliuga, R., Kuznetcov, P., & Mamychev, A. (2019). Criteria for Recognition of AI as a Legal Person. *Journal of Politics and Law, 12*(3), 105-112. https://doi.org/10.5539/jpl.v12n3p105

Duxbury, S. W., & Haynie, D. L. (2017). The Network Structure of Opioid Distribution on a Darknet Cryptomarket. *Journal of Quantitative Criminology, 34*(4), 921-941. https://doi.org/10.1007/s10940-017-9359-4

Fenwick, M. (2009). *Corporate wrongdoing and the limits of the criminal law in Facing the limits of the law*. Publisher: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-79856-9_7

Fletcher, G. (1998). *Basic concepts of Criminal law, Oxford University Press*. New York. 240 p.

Kerr, I. R., & Bornfreund, M. (2005). Buddy bots: How Turing's fast friends are under-mining consumer privacy. *Presence: Teleoperators and Virtual Environments, 14*(6), 647–655. https://doi.org/10.1162/105474605775196544

King, T. C., Aggarwal, N., Taddeo, M., & Floridi, L. (2020) Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions. *Science and Engineering Ethics, 26*, 89–120. https://doi.org/10.1007/s11948-018-00081-0

Lin, P. (2012). Introduction to Robot Ethics. In P. Lin, K. Abney, & G. A. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics*. MIT Press, London.

Model Penal Code developed by the American Law Institute. Retrieved from https://ia800102.us.archive.org/29/items/ModelPenalCode_ALI/MPC%20full%20%28504%20pages%29.pdf

Mohamed, E. B., & Marchuk, I. (2013). A comparative study of the principles governing criminal responsibility in the major legal systems of the world (England, United States, Germany, France, Denmark, Russia, China, and Islamic legal tradition). *Criminal Law Forum, 24*(1), 1-48. https://doi.org/10.1007/s10609-012-9187-z

Pergolizzi, J. V., LeQuang, J. A., Taylor, R., & Raffa, R. B. (2017). The "Darknet": The new street for street drugs. *Journal of Clinical Pharmacy and Therapeutics, 42*(6), 790-792. https://doi.org/10.1111/jcpt.12628

Rome Statute of the International Criminal Court. (1998). Text of the Rome Statute circulated as document A/CONF.183/9 of 17 July 1998 and corrected by process-verbaux of 10 November 1998, 12 July 1999, 30 November 1999, 8 May 2000, 17 January 2001 and 16 January 2002. The Statute entered into force on 1 July 2002. Retrieved from https://www.icc-cpi.int/NR/rdonlyres/EA9AEFF7-5752-4F84-BE94-0A655EB30E16/0/Rome_Statute_English.pdf

The General Product Safety Regulations. (2005). The United Kingdom. Retrieved from http://www.legislation.gov.uk/uksi/2005/1803/regulation/7/made

Vellinga, N. E. (2017). From the testing to the deployment of self-driving cars: Legal challenges to policymakers on the road ahead. *Computer Law and Security Review, 33*(6), 847-863. https://doi.org/10.1016/j.clsr.2017.05.006

Wallerstein, S. (2014). Oblique intent in English and Jewish law. *Oxford Journal of Law and Religion, 3*(2), 258-285. https://doi.org/10.1093/ojlr/rwt044

Wu, S. S. (2020). Autonomous vehicles, trolley problems, and the law. *Ethics and Information Technology, 22*(1), 1-13. https://doi.org/10.1007/s10676-019-09506-1

**Copyrights**