Criminal behavior and accountability of artificial intelligence systems

Summary

The research deals with the subject of criminal responsibility *of* Artificial Intelligence (AI) systems, focusing on whether such a legal framework is *needed* and *feasible*.

Chapter 1 presents the main RQ (to what extent is a theoretical framework of criminal liability for non-human agents *needed* and *feasible*?) and the issues that that will be discussed throughout the research, together with a structure of the corresponding sub-questions. It then outlines the methodology of the study and provides a set of examples of "AIs going bad".

Chapter 2 tackles the issue of defining AI and adopts the AI-HLEG definition as working definition for the study. This provides foundation to the analysis. The definition will then be tested throughout the following chapters and assessed in Chapter 8.

Chapter 3 delivers an extensive literature review, which is used to situate the study amongst other scholarly outputs. The analysis of scholarly debate on AI and criminal law is based on over 100 sources written in three languages (Italian, English, and German). The authors are divided into three categories: expansionists, moderates, and skeptics. The Chapter is concluded with the identification of 10 recurring questions and 7 gaps.

Chapter 4 introduces the heart of the study: an analysis which mirrors, in structure, the classical construct of criminal offenses. Indeed, AI seems to clash with traditional notions of criminal law, and understanding how to do deal with this (apparent) conflict is one of the research's tenets. In order to discuss said issues, the Chapter presents an analogy between the field of AI and that of aviation, together with an overview of different theories of criminalization.

Chapter 5 focuses on whether AI systems could display the prerequisites of criminal liability, i.e., the characteristics that are needed in order for a subject to be a plausible addressee of a criminal norm. Such a reflection is conducted by discussing the connection between moral and illegal wrongs and by examining whether AI systems could be considered moral and/or legal agents. This Chapter advances the idea that the *comprehension* of an offense's command is an essential prerequisite for establishing criminal liability and that, as such, it represents the main obstacle to attributing liability to AI systems directly.

Chapter 6 considers whether AI behavior could fulfil the *mens rea* requirement of a criminal offense, i.e., whether an AI system could be deemed "guilty". When doing so, it also looks at humans-behind-the-machine. Specifically, it addresses situations in which the classical building blocks of negligence, i.e., risk taking, foreseeability, and awareness, struggle to identify a liable human being to whom we can attribute AI-caused harm. Then, the chapter shifts its focus to whether AI behavior could fulfil the *actus reus* requirement. In particular, it identifies three main issues which obstruct the identification of a clear causal nexus between an AI act and the realization of harm. Subsequently, it analyzes with a critical eye the differences and similarities between AI criminal liability and corporate criminal liability.

Chapter 7 provides an outline of the state of the art regarding the adoption of hard and/or soft law instruments directed at regulating AI and criminal liability. In particular, it analyzes: A – The Council of Europe's European Committee of Criminal Problems and the drafting of an "Instrument on Artificial Intelligence and Criminal Law"; B – the Singapore Penal Code Review Committee Report of 2018 and the Report on "Criminal Liability, Robotics and AI systems" drafted by the Singapore Law Commission of 2021; C – the legislative reform of the French Road Act (*Ordonnance n° 2021-443 du 14 avril 2021 relative au régime de responsabilité pénale applicable en cas de circulation d'un véhicule à délégation de conduite et à ses conditions d'utilisation*); D – the "Automated Vehicles: joint report" drafted by the Law Commission of England and Wales and by the Scottish Law Commission; E – the amendment of the German Road Traffic Act.

Chapter 8 retraces the questions and the interim conclusions posed throughout the research, and identifies avenues for further inquiry. It suggests that the two prongs of the research question are switched, i.e., that the RQ should be reformulated as follows: to what extent is a theoretical framework of criminal liability for non-human agents feasible and needed. The succinct answer to the first prong of the RQ is the following: criminal liability of AI agents is *feasible* depending on how much of our anthropomorphic notions we are willing to give up. Since criminal liability can be imputed only to individuals who are capable of understanding and following the *demands* of criminal law – i.e., to make a deliberate choice *not* to conform to normative expectations, and since it is not possible, as of now, to attribute such capacity to AI systems, holding AI systems responsible would entail either being able to create machines that are *susceptible* to the commands of a legal norm, or abandoning the concept of rationality (meant as the ability to be responsive to the law) altogether. The succinct answer to the second prong of the RQ is the following: criminal liability of AI agents is not *needed* as long as AI systems are not treated as members of our community and, as such, as subject to the same rules as us. The change in perception, i.e., the anthropomorphization of AI Systems, could happen only the moment

"true" Artificial Moral Agents (AMAs) are developed. Being an AMA would suffice to pin moral blame on AI systems. Yet, in order to be considered criminally culpable, AI systems should also be "legal" agents. And, in other to be legal agents, AI systems would have to be responsive to criminal norms. In other words, the questions of whether criminal liability of AI systems is feasible, and needed, remained intertwined.

Sintesi

La ricerca affronta il tema della responsabilità penale *dei* sistemi di Intelligenza Artificiale (IA), concentrandosi sulla sua *necessità* e *fattibilità*.

Capitolo 1. Il primo capitolo introduce la domanda di ricerca principale (*To what extent is a theoretical framework of criminal liability for non-human agents* needed *and* feasible?) e le tematiche che saranno discusse nel corso della ricerca, insieme a una sintesi delle sotto-domande correlate ai vari capitoli. La ricerca illustra poi la metodologia dello studio e fornisce una serie di esempi di "*AIs going bad*".

Capitolo 2. Per dare fondamento all'analisi, lo studio inizia affrontando la questione della mancanza di una definizione universalmente accettata di IA e adotta la definizione fornita dall'AI-HLEG quale definizione operativa per la ricerca. Tale definizione viene poi testata nei capitoli successivi e la sua adeguatezza viene valutata nel Capitolo 8.

Capitolo 3. Il capitolo offre un'ampia rassegna della dottrina in materia di IA e diritto penale, utilizzata quale base per collocare questa ricerca all'interno del dibattito. L'analisi si basa su oltre cento fonti scritte in tre lingue (italiano, inglese e tedesco). Gli autori sono suddivisi in tre categorie: espansionisti, moderati e scettici. Il capitolo si conclude con l'individuazione di dieci domande ricorrenti e sette lacune.

Capitolo 4. Il capitolo introduce il cuore dello studio: l'impatto dell'IA sui costrutti classici del diritto penale. Per fare ciò, viene proposta un'analogia fra il campo dell'IA e quello dell'aviazione, nonché una disamina delle analisi delle principali teorie poste alla base delle scelte di criminalizzazione.

Capitolo 5. Il capitolo si concentra sull'imputabilità dei sistemi di IA, ossia se questi possano possedere le caratteristiche necessarie affinché vengano trattati quali "destinatari plausibili" della sanzione penale. Tale riflessione viene condotta approfondendo la connessione tra illecito morale e illecito penale e, di conseguenza, viene discusso se i sistemi di IA possano essere considerati agenti "morali" e/o soggetti attivi ai fini della commissione di un reato (c.d. "*legal agents*"). In questo capitolo viene presentata la tesi secondo cui l'incapacità dei sistemi di IA di *comprendere* il comando espresso dalla fattispecie penale rappresenterebbe il principale ostacolo all'attribuzione diretta di responsabilità penale in capo ad essi.

Capitolo 6. Il capitolo esamina se il comportamento dell'IA possa soddisfare il requisito dell'elemento soggettivo di un reato, ossia se un sistema IA possa essere considerato "colpevole". Nel fare ciò, viene posta attenzione anche ai c.d. "*humans-behind-the-machine*". In particolare, per quanto riguarda quest'ultimi, viene discussa la possibilità di configurare una responsabilità a titolo colposo a loro carico. Uno degli snodi più problematici si rinviene nel livello di attenzione esigibile dal potenziale supervisore a causa di alcuni fenomeni diffusi nel campo dell'automazione, quali ad esempio l'*automation complacency* e l'*automation bias*. Tali fenomeni, amplificati nel caso di automatizzazione basata su IA, riducono sensibilmente la soglia di attenzione dell'agente umano alle prese con la macchina, e, di conseguenza, diminuiscono la soglia dell'esigibilità di una condotta diligente da parte dello stesso.

Il capitolo analizza poi se l'agire dell'IA possa essere considerato rilevante per la qualificazione dell'elemento oggettivo del reato. Per quanto riguarda l'accertamento del nesso causale, in particolare, vengono individuati tre ostacoli principali: il "*problem of many hands*", il problema della "*black box*" e il problema delle "scorciatoie". Successivamente il capitolo analizza con occhio critico le differenze e le affinità tra la responsabilità penale dell'IA e la responsabilità penale delle persone giuridiche.

Capitolo 7. Questo capitolo fornisce una panoramica dello stato dell'arte relativo all'adozione di strumenti di *hard* e/o *soft law* volti a regolamentare l'IA e la responsabilità penale. In particolare, analizza: A – la stesura di uno "Strumento sull'intelligenza artificiale e il diritto penale" da parte del Consiglio d'Europa; B – il rapporto del Comitato di Revisione del Codice Penale di Singapore del 2018 e il rapporto su "Responsabilità penale, robotica e sistemi di intelligenza artificiale" redatto dalla Singapore Law Commission nel 2021; C – la riforma legislativa del codice della strada francese; D – il "*Automated Vehicles: joint report*" redatto dalla Law Commission di Inghilterra e Galles e dalla Law Commission scozzese; E – la modifica del codice stradale tedesco.

Capitolo 8. Questo capitolo ripercorre le domande e le conclusioni intermedie poste nel corso della ricerca e identifica percorsi per future indagini. Suggerisce di scambiare i due elementi su cui si articola la domanda di ricerca principale, ossia di analizzare prima la fattibilità di un meccanismo di imputazione di responsabilità penale diretta in capo ai sistemi di IA, e interrogarsi dopo sulla sua necessità. La risposta sintetica alla prima parte della domanda di ricerca è la seguente: la responsabilità penale degli agenti algoritmici è possibile a seconda di quanto siamo disposti a rinunciare alle nostre nozioni giuridiche antropomorfiche. La responsabilità penale può essere attribuita solo a coloro che sono in grado di comprendere, e aderire, alle "richieste" del diritto penale – soggetti cioè capaci di fare una scelta intenzionale di non conformarsi alle aspettative normative. Ad oggi, non è possibile attribuire tale capacità ai sistemi di IA. Ne consegue che al fine di poter ritenere

i sistemi di IA penalmente responsabili sarebbe necessario o essere in grado di creare macchine che siano suscettibili ai comandi di una norma giuridica, o abbandonare del tutto il concetto di razionalità (intesa come capacità di "rispondere alla legge"). La risposta sintetica alla seconda parte della domanda di ricerca è la seguente: la responsabilità penale dei sistemi di IA non è necessaria finché i sistemi di IA non verranno ritenuti membri della nostra comunità e, in quanto tali, soggetti alle nostre stesse regole. Il cambiamento di percezione, cioè l'antropomorfizzazione dei sistemi di IA, potrà avvenire ove vengano sviluppati dei "veri" agenti morali artificiali. Tuttavia, sebben essere un "agente morale artificiale" sarebbe sufficiente per attribuire una colpa morale ai sistemi di IA, non sarebbe adeguato a considerare gli stessi agenti penalmente colpevoli. Difatti, per essere considerati soggetti attivi di un reato, i sistemi di IA dovrebbero rispondere al comando delle fattispecie di diritto penale e ciò, ad oggi, non è ancora possibile. In altre parole, fattibilità e necessità di un quadro di responsabilità penale dei sistemi di IA rimangono intrecciate.

Samenvatting

Het in dit proefschrift gepresenteerde onderzoek gaat over de strafrechtelijke aansprakelijkheid van kunstmatige intelligentie-systemen (AIs).

Hoofdstuk 1 presenteert de hoofdvraag: in hoeverre is een theoretisch kader van strafrechtelijke aansprakelijkheid van niet-menselijke actoren noodzakelijk en mogelijk? Daarnaast worden de in dit proefschrift neergelegde fundamentele kwesties geïntroduceerd, inclusief de corresponderende deelvragen. Het introductiehoofdstuk eindigt met een presentatie van de methodologie waar het proefschriftonderzoek op is gebaseerd en enkele illustratieve voorbeelden van "AIs going bad".

Hoofdstuk 2 gaat in op de kwestie hoe AI te definiëren. Concreet wordt er daarbij aangesloten bij de AI-HLEG-conceptualisering. In de resterende hoofdstukken van het proefschrift wordt deze definitie getoetst en in hoofdstuk 8 op zijn merites beoordeeld.

Hoofdstuk 3 situeert het in dit proefschrift gepresenteerde onderzoek binnen de volle reikwijdte van het wetenschappelijke debat over AI en strafrecht. Daarvoor zijn er meer dan 100 publicaties geraadpleegd in drie talen (Italiaans, Engels en Duits). De respectievelijke auteurs en hun wetenschappelijke productie wordt gegroepeerd in drie categorieën: expansionisten, gematigden en sceptici. Het hoofdstuk wordt afgesloten met de signalering van tien terugkerende vragen en zeven lacunes.

Hoofdstuk 4 presenteert het fundament van het onderzoek: een analyse die qua structuur de klassieke indeling van strafbare feiten volgt. AI lijkt immers niet in traditionele strafrechtterminologie te positioneren: een van de uitgangspunten van dit onderzoek is hoe met dit (schijnbare) conflict moet worden omgegaan. In dit hoofdstuk wordt daarom AI en de luchtvaart vergeleken en een overzicht gepresenteerd van verschillende theorieën over criminalisering.

Hoofdstuk 5 gaat in op de vraag of AI-systemen kunnen voldoen aan de voorwaarden voor strafrechtelijke aansprakelijkheid. Concreet: of AI-systemen de kenmerken herbergen voor het 'zijn' van een adressant van een strafrechtelijke norm. Deze vraag wordt beantwoord aan de hand van een presentatie van het verband tussen morele en illegale misstanden en door na te gaan of AI-systemen kunnen worden beschouwd als morele en/of rechtspersonen. In dit hoofdstuk wordt het idee verdedigd dat begrip van een strafbaar feit een essentiële voorwaarde is voor de vaststelling van strafrechtelijke aansprakelijkheid

en dat dit als zodanig het belangrijkste obstakel vormt om AI-systemen rechtstreeks aansprakelijk te stellen.

Hoofdstuk 6 beschouwt of AI-gedragingen kunnen voldoen aan de mens rea-eis van een strafbaar feit. In het kort: kunnen AI-systemen "schuldig" zijn? Daarbij wordt gekeken naar de mens achter de betreffende AI-machines. In het bijzonder wordt ingegaan op situaties waarin de klassieke bouwstenen van nalatigheid, te weten risicobereidheid, voorzienbaarheid en besef, het moeilijk maken om een aansprakelijk mens aan te wijzen aan wie door AI veroorzaakte schade kan worden toegerekend. Vervolgens gaat het hoofdstuk in op de vraag of AI-gedragingen kunnen voldoen aan het actus reus-vereiste. In het bijzonder worden de drie belangrijke kwesties geïdentificeerd die de identificatie van een evident causaal verband tussen een AI-handeling en het ontstaan van schade blokkeren. Vervolgens worden kritisch de verschillen en overeenkomsten tussen strafrechtelijke aansprakelijkheid voor AI en strafrechtelijke aansprakelijkheid van ondernemingen gepresenteerd.

Hoofdstuk 7 geeft een overzicht van de stand van zaken met betrekking tot de opname van hard en/of soft law-instrumenten ter regulering de strafrechtelijke aansprakelijkheid van AI. In het bijzonder wordt ingegaan op A – het Europees Comité voor strafrechtelijke vraagstukken van de Raad van Europa en het "Instrument on Artificial Intelligence and Criminal Law"; B – the Singapore Penal Code Review Committee Report of 2018 en het rapport over "Criminal Liability, Robotics and AI systems", zoals opgesteld door de Singapore Law Commission of 2021; C – de herziening van de Franse Wegenwet (*Ordonnance n° 2021-443 du 14 avril 2021 relative au régime de responsabilité pénale applicable en cas de circulation d'un véhicule à délégation de conduite et à ses conditions d'utilisation*), D – de "Automated Vehicles: joint report" opgesteld door de Law Commission of England and Wales en de Scottish Law Commission; E – het amendement van de Duitse *Straßenverkehrsgesetz*.

Hoofdstuk 8 bespreekt de vragen en deelconclusies die in het proefschrift zijn opgeworpen nog een laatste maal. Daarnaast worden enkele aanbevelingen voor verder onderzoek gepresenteerd. Zo wordt voorgesteld om de twee onderdelen van de onderzoeksvraag om te wisselen. De vraag leest dan als volgt: in hoeverre is een theoretisch kader van strafrechtelijke aansprakelijkheid voor niet-menselijke agenten haalbaar en nodig. Het beknopte antwoord op het eerste deel van de vraag: strafrechtelijke aansprakelijkheid van AI-agenten is mogelijkheid, maar afhankelijk van de mate waarin wij bereid zijn onze traditionele antropomorfe opvattingen los te laten. Aangezien strafrechtelijke aansprakelijkheid alleen kan worden toegeschreven aan individuen die in staat zijn de eisen van het strafrecht te begrijpen en te volgen – i.e. een afgewogen keuze te maken om zich niet aan zekere nor-

matieve verwachtingen te conformeren – en aangezien het op dit moment niet mogelijk is een dergelijk vermogen aan AI-systemen toe te kennen, zou het verantwoordelijk stellen van AI-systemen betekenen dat men ofwel in staat moet zijn machines te creëren die ontvankelijk zijn voor de bevelen van een wettelijke norm, ofwel het begrip rationaliteit (in de zin van het vermogen om te reageren op het recht) helemaal moet worden opgeven. Het beknopte antwoord op het tweede punt van de hoofdvraag luidt als volgt: strafrechtelijke aansprakelijkheid van AI-agenten is niet nodig zolang AI-systemen niet worden behandeld als leden van onze gemeenschap die aan dezelfde regels zijn onderworpen als wij. Een verandering in perceptie, i.e. de antropomorfisering van AI-systemen, zou pas plaats kunnen vinden op het moment dat er "echte" Artificial Moral Agents (AMA's) worden ontwikkeld. Het 'zijn' van een AMA zou het dan mogelijk maken om AI-systemen moreel te beschuldigen. Maar om strafrechtelijk verwijtbaar te zijn, moeten AI-systemen ook rechtspersonen zijn. En om rechtspersonen te zijn, zouden AI-systemen moeten reageren op criminele normen. Met andere woorden, de vraag of strafrechtelijke aansprakelijkheid van AI-systemen haalbaar en nodig is, blijft met elkaar verweven.