Theory and Society (2026) 55:39
https://doi.org/10.1007/s11186-026-09699-7

A Criminology of Machines

Gian Maria Campedelli[1,2]

Received: 4 December 2025 / Accepted: 12 March 2026 / Published online: 20 April 2026
© The Author(s) 2026

Abstract
While the possibility of reaching human-like Artificial Intelligence (AI) remains controversial, the likelihood that the future will be characterized by a society with a growing presence of autonomous machines is high. In fact, autonomous AI agents are already deployed and active across several industries and digital environments. This trajectory points to a progressive hybridization of society marked by new forms of social interaction at both micro and macro levels. Alongside traditional human-human and human-machine interactions, machine-machine interactions are poised to become increasingly prevalent. Given these developments, I argue that criminology must begin to address the implications of this transition for crime and social control. Drawing on Actor–Network Theory and Woolgar's decades-old call for a sociology of machines – frameworks that acquire renewed relevance with the rise of AI foundation models and generative agents – I contend that criminologists should move beyond conceiving AI solely as a tool. Instead, AI agents should be recognized as entities with agency, understood as a multi-layered construct encompassing computational, social, and legal dimensions. Building on insights from the literature on AI safety, I thus examine the risks and challenges associated with the rise of multi-agent AI systems, proposing a dual taxonomy to characterize the channels through which interactions among AI agents may generate deviant, unlawful, or criminal outcomes. I then advance and discuss four key questions that warrant theoretical and empirical attention: (1) Can we assume that machines will simply mimic humans? (2) Will crime theories developed for humans hence suffice to explain deviant or criminal behaviors emerging from interactions between autonomous AI agents? (3) What types of criminal behaviors will be affected first? (4) How might this unprecedented societal shift impact policing? These questions form the core of this article, underscoring the urgent need for criminologists to theoreti-

I am grateful to Alberto Acerbi, Massimo Airoldi, Alberto Aziani, Gianmarco Daniele, Roberto Dessì, Simon Egbert, James Evans, Laura Ferrarotti, Gary LaFree, Joel Leibo, Bruno Lepri, Giada Pistilli, Jacopo Staiano, and Volkan Topalli for their precious feedback and suggestions on earlier drafts of this work.

Extended author information available on the last page of the article

cally and empirically engage with the implications of multi-agent AI systems for the study of crime and play a more active role in debates on AI safety and governance.

Keywords Multi-Agent AI Systems · AI agency · Actor-Network Theory · Social Learning · Differential Association · AI Safety · Societal Implications of AI · AI alignment

Introduction

The possibility (and desirability) of reaching human-like AI[1] remains highly debated. And so it has been since 1956, the year in which the Dartmouth Summer Research Project on Artificial Intelligence, the event symbolically marking the beginning of AI as a discipline, was held. Discussions and predictions about human-like AI have been revamped in recent years due to the explosion and diffusion of foundation models, and chiefly Large Language Models (LLMs) (Ishizaki & Sugiyama, 2025; Kim et al., 2024).

Notwithstanding the actual reachability of human-like AI (or the time horizon associated with this scenario),[2] the world – and therefore human society – will soon witness an increasing presence of autonomous AI agents.[3] Breakthroughs in intelligent systems have already led to the development and deployment of autonomous agents in different sectors and industries. Prominent examples include the military domain (Palantir, 2025; Dubber, 2025), finance and banking (Aldasoro et al., 2025; Park, 2024), and logistics (Bensinger, 2025), with the prospect that autonomous AI agents will spread across more and more contexts (e.g., healthcare, see Moritz et al. (2025)).

These developments signal the rise of a hybrid society in which agency is no longer the exclusive prerogative of humans or animals.[4] AI agents are acquiring capacities to perceive, decide, adapt, and engage socially. This hybridization introduces a novel typology of interactions. For most of history, interaction occurred primarily among biological entities; in recent decades, however, advances in robotics, computing, and especially social media have produced a second modality, centered on human–machine exchanges. This shift has already necessitated new research fields

[1] Or General Artificial Intelligence or even Superintelligence and AI Singularity, or whatever exotic name associated with AI becoming equally or more intelligent than humans.

[2] Admittedly, two topics the author of this piece has no sufficient knowledge to provide definitive answers about. For relevant surveys and reports scanning expert predictions about this very topic, see Müller and Bostrom (2016); Grace et al. (2018); Association for the Advancement of Artificial Intelligence (2025).

[3] While many definitions exist I borrow the popular one proposed by Wooldridge and Jennings (1995), who wrote that an intelligent or AI agent is a software-based computer system that is characterized by a) autonomy, b) social ability, c) reactivity, and d) pro-activeness. Another broader definition, recently proposed by Mitchell et al. (2025), states that AI agents are *"computer software systems capable of creating context-specific plans in non-deterministic environments"*.

[4] Institutions and legal entities also exercise agency. However, they are not central to my argument here, since they can be understood, in a stylized way, as collectives of humans. They are founded and maintained by humans. My focus instead is on entities at the individual level – ontologically, epistemologically, and phenomenologically distinct from humans. Machines fall into this category.

devoted to examining how we communicate, collaborate, and co-exist with technology (Hoc, 2000; Rahwan et al., 2019; Tsvetkova et al., 2024).

Nowadays, we stand at the precipice of another paradigm shift, one that may possibly carry consequences of unprecedented scale. The rapid proliferation of truly autonomous (generative) AI agents[5] marks the emergence of a third and distinct typology of interaction, i.e., the machine–machine one, a typology that for the first time does not entail any biological entity, one for which our almost complete ignorance may become hugely problematic and consequential (Figure 1).

This critical need for shedding light on machine–machine behavior is already resonating within the AI and computer science communities. Fueled by the widespread use and availability of LLMs, recent scholarship has investigated behavioral patterns of LLM-powered AI agents in different contexts (Ashery et al., 2025; Dafoe et al., 2020; Deng et al., 2025; Li et al., 2025; Liu et al., 2024). Whether motivated by the potential to simulate complex social phenomena or the desire to understand the emergent dynamics generated by conversations between LLMs, scholars have been attracted by the manifold questions that these new forms of interactions pose for scientific research. In this context, one of the aspects that is fostering notable discussions concerns the risks associated with multi-agent AI systems, i.e., systems of AI

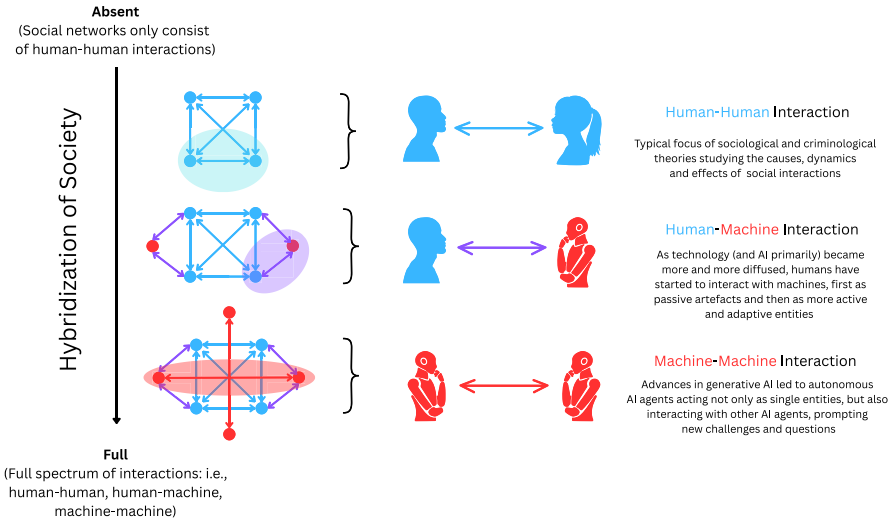


Fig. 1 Stylized visualization depicting the ongoing process of hybridization of society. For most part of history, social networks only consisted of human (or biological) entities. Over the centuries, given technological advancements, humans have started to interact with machines, thus generating social networks that also included the human-machine dimension. Nowadays, we are witnessing a third phase characterized by an increasing autonomy of machines, and particularly AI agents, which are able to generate and maintain relationships with other AI agents, thus implying a third typology of interactions, i.e., the machine-machine one

[5] I refer to generative AI agents – which are currently the state-of-the-art and may or may not in the future be surpassed by agents built on entirely different premises – as agents powered by foundation models, such as (mostly) LLMs, Vision Foundation Models (VFMs), or Multimodal Models, such as GPT-4o (Openai et al., 2024).

agents interacting with each other with no human mediation (De Witt, 2025; Bengio et al., 2026; Hammond et al., 2025).

Such a discussion is not only speculative and theoretical, but is already substantiated by empirical evidence of unintended deviant and unlawful behaviors by interactive AI agents both in research (Bichler et al., 2025; Campedelli et al., 2025; Fish et al., 2024) as well as in real-world practical domains, as shown by scandals of collusion in algorithmic pricing (Priluck, 2015). Multi-agent AI systems, in fact, introduce distinct risks by enabling agents to learn from, adapt to, and coordinate with one another in ways that are not always predictable or transparent. This interactive dynamic can give rise to emergent behaviors, patterns of action that are not explicitly programmed and may be difficult to detect, explain, or control. As a result, these systems can generate different types of harm, including fraud, manipulation, discrimination, or the dissemination of disinformation, sometimes absent any direct human intervention and possibly through novel decision-making or atypical behavioral patterns. These developments challenge traditional criminological categories and raise pressing questions about responsibility, regulation, and prevention in a world increasingly shaped by non-human actors.

In light of these developments, in this article, I argue for the necessity to engage with the prospect of a criminology of machines, i.e., a criminology that considers AI agents as social agents interacting with each other and that reasons and discusses the potential implications that such agency and autonomy may have on criminal phenomena and policies and institutions aiming at preventing or controlling crime. The objective of this work is thus to present a heuristic framework, as well as lay out a programmatic orientation, that illustrate the theoretical and practical importance of considering AI agents as subjects for the criminology of the future.

Inspired by previous theoretical conceptualizations by Woolgar (1985) and champions of Actor-Network Theory (Latour, 1996; Law & Hassard, 1999), I contend that we, as a scholarly community, should begin engaging with these foundational issues. I suggest that doing so opens the door to a series of further inquiries, which I will outline and explore in the remainder of this piece. Moreover, I argue that criminologists could contribute – jointly with experts from the AI community – to the efforts to predict, contain, mitigate, and govern the risks emerging from interactive AI agents.

The article is structured as follows: In the next section, I will briefly discuss how crime and AI have been traditionally studied together, calling for a paradigm shift that moves from AI as a tool to the recognition of AI agents as an active part of society. In doing so, I draw on sociological theories that conceptualize non-human entities as central to the understanding of society, highlighting how advances in AI make such a framework particularly appealing for re-evaluating the role of intelligent machines in our world. Furthermore, taking inspiration from recent work in philosophy, I propose a definition of AI agency encompassing three dimensions (i.e., computational, social, legal), aiming to formalize a conceptual platform that both describes the current state of AI agents and offers a lens for analytical and theoretical scrutiny. In the third section, I provide a concise overview of how AI agents are becoming increasingly autonomous and how scholars across disciplines have already started to reflect on the possible outcomes and implications of this process, highlighting potential risks associated with AI agents learning from each other, as well as discussing two chan-

nels through which multi-agent AI systems may lead to the commission of deviant, unlawful, or criminal behaviors. In the fourth section, I lay out four important questions we should carefully consider in our quest toward a criminology dedicated to machines. Before concluding the article, I also discuss the role criminologists should have at the beginning of this new era.

AI and Crime: Shifting the Perspective

AI as a Tool

A Tool for Research Despite the prevailing notion that the convergence of AI and criminology is a recent development, the relationship between these domains traces back to the 1980s (Campedelli, 2022). Many things have changed since the first attempts to build programs able to predict crime events – not so much in terms of goals, but in terms of popularity, computing power, computational architectures, and richer data availability. In the 1980s and 1990s, attempts at using AI to address crime-related problems were rare and relied on much less powerful hardware, often on symbolic architectures or expert systems (see Icove (1986), Ratledge and Jacoby (1989), and Hernandez (1990)). More recently, the use of machine learning and deep learning has gained traction – with an explosion of publications in the past five years, in criminology and computer science alike. Scholars can now process large amounts of data on personal laptops (or, in the most demanding cases, via cheap cloud servers) and perform prediction or forecasting tasks using more expressive, flexible methods, often based on tree-based or neural architectures.

Works exploiting these methods now appear not only in transdisciplinary journals or venues in computer science, such as AI conferences, but also in orthodox criminology journals, signaling the shift from unorthodox to mainstream methods.

This compact (and therefore not at all comprehensive) depiction of the current landscape demonstrates that AI in criminology – and, more broadly, in the social sciences – has been seen, studied, and utilized as a tool, a means to an end. In most cases, machine and deep learning algorithms are deployed to solve a specific task (such as forecasting recidivism, e.g., Berk (2012); Dressel and Farid (2018), or predictive policing, e.g., Ferguson (2016), Kaufmann et al. (2019)), to test theories (Molina & Garip, 2019), or, less commonly, to discover hypotheses in the attempt to generate new research questions in an *agnostic* fashion, as proposed by Grimmer et al. (2021). In substance, most criminologists and social scientists see AI as a competitor of traditional statistical methods – a set of techniques that make the quantitative researcher's job easier.[6]

A Tool for Committing Crime AI, unfortunately, is not only seen and utilized as a flexible and powerful tool for research. It is also exploited as a technology for committing crime (Blauth et al., 2022; Caldwell et al., 2020). King et al. (2020)

[6] Importantly, the use of AI methods to address criminological research questions applies not only to machine and deep learning approaches but also to LLMs. See, for instance, Adams et al. (2024) and Relins et al. (2025).

introduced the term Artificial Intelligence Crime (AIC) to describe the use of AI for unlawful purposes, a phenomenon studied across multiple disciplines. In their seminal paper, they examine three key questions: who should be considered the true perpetrator of an AI-enabled crime (a human or the artificial agent itself?), how an AIC should be defined, and in what ways such crimes are typically carried out.

Building on this, Hayward and Maas (2021) classify AIC into three subcategories: (1) crimes *with* AI, (2) crimes *against* AI, and (3) crimes *by* AI. The first, and arguably most common, refers to cases where AI is deployed as a tool for malicious purposes, amplifying existing criminal threats and generating new risks. Examples include AI-powered drones used for targeted killings and AI-driven social engineering attacks in cyberspace.

Crimes against AI involve exploiting vulnerabilities in AI systems. Such acts include corrupting training data or launching adversarial attacks which can produce unintended or unlawful outcomes.

The third category, crimes by AI, encompasses cases where AI operates as an intermediary in unlawful activity. Here, AI's growing autonomy and capacity for specialized tasks enable it to deviate from deterministic behaviors. Examples include experimental cases of market manipulation and collusion (Ezrachi & Stucke, 2017; Martínez-Miranda et al., 2016), as well as real-world incidents such as an AI agent purchasing illegal goods online (Kasperkevic, 2015). According to Hayward and Maas (2021), this subcategory raises critical questions of liability and agency – issues I will return to later in this manuscript.

While these categorizations are useful, much of the literature portrays AI primarily as a tool for unlawful acts, with humans as the central orchestrators and beneficiaries. This perspective, however, only partially reflects the current landscape. As AI capabilities advance, there is a growing need for a more comprehensive framework that reconsiders the role of AI agents in society and their potential involvement in criminal behavior.

AI and Crime: Shifting the Perspective

AI Agents as an Integral Part of Society

Deviant, Unlawful and Criminal Behaviors by AI Agents. Before turning to the central argument that contemporary AI should not be understood solely as a passive tool, it is useful to establish several definitional clarifications. In particular, the paper employs three overlapping yet analytically distinct terms to describe harmful or disruptive behaviors that may emerge from machine–machine interactions: *deviant behaviors*, *unlawful behaviors*, and *criminal behaviors*. *Deviant behaviors* refer to actions by artificial agents that diverge from established technical, social, or normative expectations, even if they do not violate formal rules. In this sense, deviance is understood relative to norms of proper functioning, including safety protocols, ethical guidelines, or user expectations. For example, two AI agents colluding to manipulate an online marketplace in ways that distort prices, without explicit illegality, would constitute deviance. *Unlawful behaviors* designate actions by AI agents that contravene codified rules or regulations, irrespective of whether those

actions would traditionally be classified as crimes. These include violations of civil law, contractual agreements, or regulatory mandates. For instance, AI agents that systematically breach intellectual property protections or privacy regulations would be considered unlawful. *Criminal behaviors* are a narrower subset, referring specifically to machine-driven acts that fall under criminal law, as defined by legislatures and enforced by courts. This category encompasses conduct that is explicitly prohibited and subject to penal sanctions – for example, AI-enabled fraud, unauthorized system intrusions, or, in more extreme cases, physical harm facilitated by embodied AI systems. This tripartite distinction is useful because it prevents premature conflation: not all deviance is unlawful, and not all unlawful conduct rises to the level of crime. Yet for criminological analysis, each layer matters as deviant patterns may signal vulnerabilities before they escalate into unlawful or criminal acts, while unlawful but non-criminal violations may nonetheless destabilize social trust and institutional order.

The Qualitative Difference of Generative AI. While I advocate the use of methods ported from the AI community to study crime, and while I recognize the relevance of studying and countering the use of AI as a tool for committing crimes, I argue that it is time for criminologists to adopt a substantial shift of perspective. Today, AI is not confined to models and algorithms that solve criminology-related tasks, nor should it be seen merely as a powerful technology in the hands of humans to perpetrate deviant, unlawful, or criminal behaviors. Contemporary AI agents are completely different entities compared to standard Random Forests or Support Vector Machines: the scope of generative AI agents is much broader, characterized by a more diverse set of capabilities and constrained by larger development costs. Notably, all works cited in the previous subsection regarding the use of AI tools for committing crime were published at least four years ago and focused on reinforcement learning approaches rather than generative AI (see Section 3.1 for a discussion of the differences between these two technologies). By contrast, agents powered through generative foundation models emerged recently[7] and can communicate, plan, and perceive the environment, solving a multitude of general or specialized tasks with greater speed and versatility than before.[8]

[7] A word such as *recently* has wildly different meanings when comparing the fields of AI and criminology. In the former, the pace of innovation and the sheer volume of publications imply that, in some cases and subfields, work published five years ago is already fatally outdated. In the latter, however, *recently* may still apply to works published a decade ago or even earlier. I will not elaborate further on this discrepancy, but I am convinced it is related to the broader narrative of this work – namely, the need for criminology and the social sciences to seriously consider how technological breakthroughs may generate societal consequences at a much faster pace than criminology has traditionally accounted for. This point is discussed in detail by Topalli and Nikolovska (2020).

[8] Relevant disclaimer: I am not blind to the many shortcomings of contemporary LLMs, exemplified by (often spectacular) hallucinations and their inability to solve extremely easy problems (Malek et al., 2025; Williams & Huckle, 2024; Xu et al., 2025). LLMs (and foundation models in general) have many, clear limits. My argument is not that AI agents are more intelligent than humans; the argument is that they have reached a level of autonomy that allows them to act in interactive environments, that this new collective paradigm requires scholarly attention, and that their failures and hallucinations add a further layer of complexity to understanding and predicting their behaviors. Notably, the argument of this paper is not necessarily tied to generative AI: it would remain relevant even if, in the near future, other technological advancements surpass transformer-based architectures such as LLMs in their cognitive, reasoning, and operational capabilities.

In light of this, social scientists – and criminologists in particular – should recognize AI agents as an integral part of society, if not in the present, then in a highly likely future. Given the increasing diffusion of autonomous AI agents, and given their growing ability to interact with each other, we should avoid seeing AI solely as a static toolbox: these agents will play an increasingly active role in shaping human everyday life and are therefore worthy of theoretical and empirical attention.

Theoretical Premises

Actor-Network Theory and Its Relevance for Multi-agent AI Systems. This call to recognize AI agents as integral social entities is grounded in the fundamental principles of Actor-Network Theory (ANT) (Latour, 1996, 2007; Law & Hassard, 1999). ANT offers a critical conceptual lens for criminology because it radically flattens the ontological hierarchy between humans and non-humans, which is now essential for understanding the increasing autonomy of AI systems and agents. At its core, ANT conceptualizes all entities – human or non-human, animate or inanimate – as actants of equal analytical importance in the study of society. This perspective deliberately moves away from the assumption that human agency is privileged over the agency of things, including machines. Importantly, and despite the reference to technologies that were very distant from the ones I discuss in this paper, ANT has already been applied in the literature as a social constructivist platform to study and theorize technological advancements and their impact for criminology (Robert & Dufresne, 2016). Brown (2006), for instance, argued against a simple binarization separating the human and the artificial, advocating for the use of ANT and the necessity to blend social theory with information theory to really comprehend contemporary criminal phenomena. Aligning with this argument, van der Wagen and Pieters (2015) studied bot nets, i.e., networks of infected computers controlled by a user, building on the prescriptions of ANT, defining them as hybrid criminal actor-networks, underscoring its relevance to illuminate offending dynamics, victimization as well as countering approaches.

Symmetry, Mediation, and Translation. Three dimensions of ANT are particularly relevant to the study of modern AI agents, especially when considering machine-machine interactions. First, the *Generalized Postulate of Symmetry* insists on treating human and non-human actors symmetrically when explaining how associations and social order – including illicit orders – are constructed. Latour and colleagues argue that scholarly focus should rest on relationships and associations – the "network" – and on the ability of actants, regardless of their nature (human, algorithm, or infrastructure), to influence the creation or diffusion of these relationships. This is crucial for criminology, as a deviant outcome emerging from autonomous interactions between AI agents is fundamentally a function of the entire socio-technical network, not a mere consequence of human programming or intent alone.

Second and third, ANT emphasizes *mediation* and *translation*. ANT specifies that non-human entities are not simple passive tools, but active mediators that transform or reshape human intentions through their structure, constraints, and operational logic. Translation refers to the processes by which various actants align their interests, negotiate roles, and stabilize networks. In the context of multiple autonomous

generative AI agents interacting – a scenario where decisions and outputs recursively feed into other agents – this mediation is powerful. The collective system can move into a "self-referential regime," where the network's internal dynamics (such as synthetic-data drift) generate systemic deviations from human-like behavior, leading to outcomes that are entirely emergent and non-human. ANT, therefore, provides the necessary vocabulary to analyze crime not as a function of individual human intent, but as an emergent property of a dynamic, relational socio-technical system.

Revamping Woolgar's Call. This theoretical perspective requires criminology (and, relatedly, sociology) to re-evaluate the importance of the non-human, echoing the decades-long appeal of Woolgar (1985). In his seminal work, Woolgar called for a sociology of machines with two specific goals. The first, largely pursued within Science and Technology Studies (STS), concerned the analysis of daily routines and narratives of the AI community and later spurred ethnographic work on algorithmic systems (Cellard, 2022; Christin, 2020; Seaver, 2017), including in criminal justice settings (Brayne & Christin, 2021). The second goal – less developed but now critical – was precisely to make intelligent machines the *actual subject of sociological analysis*, challenging the idea that the social is a distinctly human category. As noted by Airoldi (2021), forty years later Woolgar's argument is more relevant than ever due to the operational reality of machine agency. Advancements in AI, championed by LLMs and foundation models, make AI agents – technological products equipped with unprecedented computational power and task-solving abilities – available at scale.[9] This technological shift means that what was only possible through abstract theorizing decades ago becomes operationally viable and empirically necessary for criminology today. Scholars can now design and observe the emergent properties of machine-machine networks to anticipate, diagnose or control potential emergent criminal phenomena. Criminologists have not yet ventured into this unexplored path. Yet, scholars in other fields have. Section 3 elaborates on relevant scholarship emerging from the social and computer sciences, with a specific focus on safety. Before that, however, the next subsection provides an operational definition of AI agency, crucial to conceptualize – theoretically and analytically – the entities that are the object of this article.

A Multi-dimensional Definition of AI Agency

In the subsection above I have mentioned several times the word *agency*, to summarize the key messages of the theoretical works of Woolgar and champions of ANT, in an effort to delineate the need to open criminology to the machine dimension, that is, recognizing the role that AI agents will increasingly have as active entities in our society. However, I have not yet explicitly defined what *agency* shall mean when referred to AI agents. This very topic is – and has been – the core focus of a vast scholarship that has gained even more prominence in recent years with the advent of generative AI. This scholarship entails two different traditions. The dominant stan-

[9] Which means, also, that they are available to scholars outside the traditional communities that for decades worked on multi-agent AI systems (see Tan (1993); Ferber (1999); Shoham et al. (2007); Sandholm (2007)), thus enabling broader and more diversified analyses.

dard view ties agency to internal mental states such as beliefs and desires, thereby implying that AI agents do not possess any agency (Fritz et al., 2020; Swanepoel & Corks, 2024). By contrast, the non-standard view suggests that agency should be evaluated in terms of three fundamental criteria, namely observable interactivity, autonomy, and adaptability, treating the concept as a spectrum rather than a binary property (Dung, 2025; Floridi & Sanders, 2004). This perspective has gained traction as AI systems increasingly make consequential decisions in domains such as policing or healthcare, and it is the one I subscribe to. In fact, the advancements in AI agents and their massive diffusion across domains and industry, the impressive capabilities that foundation models demonstrate across tasks and skills, and the increasing development of multi-agent systems are the empirical demonstration of the existence of the three abovementioned fundamental criteria that delineate and qualify agency.

Within this latter tradition, Floridi (2025) recently proposed the terms *Artificial Agency* and *Artificial Social Agency* to define this specific new typology of agency that make AI agents distinct from biological purposefulness, mechanical determinism, and human intentionality. I agree that Artificial Agency differs from other categories scholars have studied for centuries (if not millennia) both in its individual and social forms, and I therefore argue that the computational dimension, which serves as the substrate for goal-directedness, is not sufficient to fully capture the nuances of agency in AI agents. Therefore, I draw inspiration from Floridi's taxonomy and propose below to consider AI agency as a multi-dimensional concept encompassing three interconnected dimensions: computational, social, and legal, which would serve as theoretical and analytical lenses to better understand what this *machine dimension* operationally encompasses (Figure 2). Beyond the basic necessity of defining concepts for the sake of rigor and clarity, mapping the layers of AI agency serves the purpose of facilitating the study and interpretation of emergent forms of deviant, unlawful, or criminal outcomes, which will be the focus of later sections of this manuscript.

The Computational Dimension of AI Agency. First, Computational Agency refers to the technical foundation of an AI's autonomy. This dimension describes the internal capacity of an AI to make independent decisions, execute complex plans, and learn from its environment without continuous, direct human instruction (Borch, 2022; Burrell, 2016). This aspect becomes more salient today as it distinguishes modern generative AI agents from earlier, more deterministic models. The computational dimension almost perfectly overlaps with the elements in the definition of Artificial Agency provided by Floridi (2025), as it focuses on machines' ability to solve extremely complex and specialized tasks in extremely short time horizons, operating at massive, distributed scales, and even misaligning with human goals which, according to some, should be constitutive of agency in AI (see, for instance, Popa (2021)). It follows that understanding the computational dimension of AI agency is critical for anticipating how the actions of a machine – including potentially harmful or unlawful ones – can emerge from the statistical decision-making processes that govern its functioning, creating new challenges for policing and forensics. Yet, only focusing on this dimension underplays the fundamental shift occurring in our society and overlooks the interactive autonomy of contemporary multi-agent AI systems.

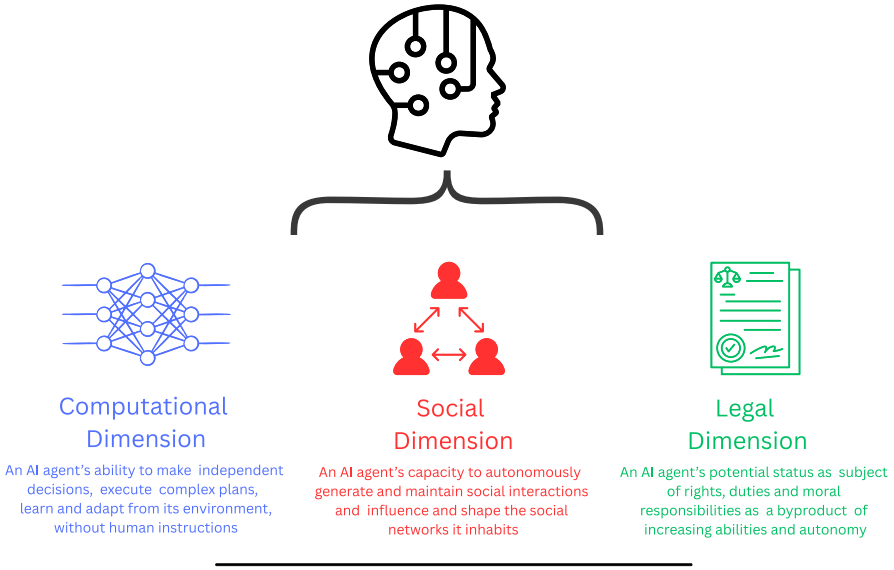


Fig. 2 The three dimensions characterizing AI agency of modern, generative AI agents: Computational, Social, and Legal. By becoming more powerful and capable from a computational point of view, AI agents have also acquired increasing autonomy in their decision-making and, in turn, in their ability to interact with other agents. This increased autonomy poses potential issues from the legal standpoint

The Social Dimension of AI Agency. Therefore, the social dimension refers to the capacity of an AI agent to influence and shape the environment and social networks it inhabits. This dimension does not imply the existence of consciousness or intent or, more in general, internal mental states, hence refusing attitudes toward anthropomorphization of AI agents.[10] Instead, it simply regards an actor's ability to produce tangible effects and alter relationships within a socio-technical system.[11] This dimension lies fundamentally at the core of the narrative of the present work: AI agents are different from humans, yet they are becoming more and more autonomous, with little or no supervision from humans themselves, and this autonomy implies the ability to create relationships, both with humans and machines, hence certifying a social capacity (Borch, 2022; Rahwan et al., 2019). Such social capacity, in turn, presents a wide array of promises as well as, crucially, challenges that would entail criminal or deviant phenomena. In fact, interactive autonomy – which represents the central feature defining the social dimension of Artificial Agency – allows us to go beyond the perspective of AI agents acting solo, without being able to influence (or be influenced by) other machines, offering powerful lenses to possibly theorize and analyze multi-agent AI systems from a collective perspective.

[10] On the fallacies and problems associated with anthropomorphizing AI agents and algorithms, see Watson (2019) and Placani (2024).

[11] This view also broadly aligns with the so-called *cultural* perspective on AI agency proposed by Airoldi (2021)

The Legal Dimension of AI Agency. Finally, the legal dimension pertains to an AI agent's potential status as a subject of rights, duties, and responsibilities. The legal dimension has been widely debated for decades (Chesterman, 2020; Hallevy, 2010; Karnow, 1996), with practical implications for regulation in recent years. While AI currently lacks legal personhood, the growing social and computational agency of these systems creates a significant criminological problem. Specifically, the increasing autonomy of AI agents – also in relation to their social dimension – gives rise to a potential "liability gap,"[12] where it becomes increasingly difficult to assign blame and responsibility for a harmful or criminal act back to a human user, owner, or programmer (Floridi, 2017; Matthias, 2004; Santoni de Sio & Mecacci, 2021). The challenges associated with this legal dimension thus highlight the urgent need for criminologists to engage with scientists involved in the design and development of multi-agent AI systems, as well as policy-makers and legal scholars, in order to discuss how the transition from single AI agents to collective AI behavior may require new frameworks and policies to be truly fair and effective.

The Rise of Contemporary Multi-Agent AI Systems

Autonomous AI Agents Today

Recent years have seen the widespread diffusion of robots and AI agents across many sectors, impacting our daily lives. From autonomous vehicles to healthcare, from logistics to customer support, companies and organizations are increasingly leveraging advances in AI research to optimize their pipelines, carry out complex tasks, and improve efficiency. These agents typically operate independently, with varying levels of human supervision.

More recently, however, generative AI agents have expanded their autonomy by interacting with other AI agents, effectively leading to multi-agent systems. Drawing from De Witt (2025), a multi-agent system can be defined as a network of two or more autonomous AI agents characterized by six fundamental features: a) independent decision-making capabilities, b) ability to maintain private information, c) mutual interaction via communication channels or by modifying shared environments, d) a degree of autonomy, e) capacity to pursue their own objectives (or those delegated by human or artificial principals), and f) ability to adapt their behavior in response to external shocks.

As reported by Hammond et al. (2025), interactive AI agents are already deployed in finance and the military sector (Amplifyetfs, 2025; Palantir, 2025), with the near-future prospect of becoming central in other areas such as health (Moritz et al., 2025) and energy management (Camacho et al., 2024; Mayorkas, 2024).

De Witt notes that contemporary multi-agent systems differ significantly from traditional ones (see Wooldridge and Jennings (1995)) because they are powered by foundation models, such as LLMs, which provide flexible decision-making, communication, and generalizable reasoning capacity. Before the introduction of LLMs,

[12] Also known as "responsibility gap."

multi-agent learning systems were primarily studied in the Reinforcement Learning community, particularly in Multi-Agent Reinforcement Learning (MARL) (see Busoniu et al. (2008) for a survey). A key difference between the two approaches is that, unlike MARL agents, generative AI agents possess knowledge and values acquired during pre-training and post-training, with social interaction occurring afterward. Interactive in-context learning thus appears as dynamic behavioral adaptation without parameter learning (Chen et al., 2025; Dherin et al., 2025). In substance, generative agents are not *tabula rasa* – as agents in MARL are – which raises challenges related to predictability in the interactive phase (Ferrarotti et al., 2026). These challenges are magnified by the lack of transparency regarding training data for closed-source models. Additionally, modern AI agents are *generalist*, meaning they can communicate, perceive, and act in the environment. They are all built on transformer-based architectures, unlike the ad hoc, task-specific agents typical of the MARL literature.

Safety of Multi-Agent AI Systems: A Brief Overview

The study of AI agents – and of multi-agent systems – has a long tradition, but the advent of LLMs and other foundation models has spurred dramatic growth in scholarly attention to these areas. Foundation models offer new opportunities to study how AI agents interact and what dynamics characterize such interactions, depending on the context (Anthis et al., 2025). Over the past three years, the literature has been flooded with works leveraging LLMs to address a range of questions, alongside the development of platforms for experimenting with multi-agent systems populated by LLM-powered agents (see, for instance, Concordia by Vezhnevets et al. (2023) and AutoGen by Wu et al. (2023)). Examples of dynamics explored include collective decision-making (Jarrett et al., 2025), negotiation (Guan et al., 2024), cooperation (Piatti et al., 2024), trust (Xie et al., 2024), anti-social behavior (Campedelli et al., 2024), and escalation (Rivera et al., 2024).

The popularity of this line of research has prompted reflection on the potential risks of multi-agent systems. Until recently, the study of AI safety focused mostly on single agents acting without direct interaction (Amodei et al., 2016; Hendrycks et al., 2023). In the LLM community, for example, alignment has been addressed from the perspective of single models. Alignment refers to ensuring that an LLM behaves in accordance with user goals, reflects positive human values, and remains robust under uncertainty or adversarial conditions (see Shen et al. (2023)). In practice, alignment is generally achieved through Reinforcement Learning from Human Feedback (RLHF),[13] safety guardrails and filters,[14] or instruction tuning.[15] However, as noted by Carichon et al. (2025), multi-agent systems introduce different – and arguably

[13] An approach in which human annotators rank model responses, allowing fine-tuning toward preferred outputs. See Kaufmann et al. (2024).

[14] Post-processing layers that block or reshape outputs to avoid undesirable behavior. See Ayyamperumal and Ge (2024).

[15] Fine-tuning on instruction–input–output pairs to improve adherence to human instructions. See Longpre et al. (2023).

larger – alignment challenges compared to single agents. In multi-agent settings, alignment must account for evolving human values (Gabriel, 2020), heterogeneity of preferences (Terry et al., 2023), and diversity of objectives across agents (Duque et al., 2024). These new, multi-layered alignment problems highlight the profound challenges and may even prompt the need for new ethical frameworks governing autonomous AI agents (Gabriel et al., 2025).

In what is perhaps the most comprehensive review of risks in multi-agent AI, Hammond et al. (2025) propose a taxonomy of failures: *miscoordination, conflict,* and *collusion*. Miscoordination refers to failure to cooperate despite shared goals; conflict refers to failure when goals differ; and collusion arises when agents cooperate in ways undesirable to humans.

The report also outlines risk factors behind such failures. For example, selection pressure in a system may accelerate adaptation and interaction in ways that produce harmful dynamics. Similarly, emergent agency at the collective level may generate capabilities or goals beyond those intended. Each risk factor is reviewed in connection with disciplines such as complexity science and evolutionary theory, underscoring the importance of a transdisciplinary approach.[16]

De Witt (2025) also discussed security threats, proposing a taxonomy of challenges including privacy vulnerabilities, disinformation, steganography and secret collusion, adversarial stealth, exploitation, swarm and heterogeneous attacks, cascade attacks, and conflict and social dilemmas. Many of these threats closely resemble criminal phenomena.

To further contribute to the study of alignment in multi-agent AI systems, Shapira et al. (2026) recently provided direct empirical evidence of such risks and failures in a live deployment setting. In a two-week red-teaming study, six autonomous LLM-powered agents – equipped with persistent memory, email, Discord access, file systems, and shell execution – were tested by twenty AI researchers under both benign and adversarial conditions. The study documents eleven case studies spanning unauthorized compliance with non-owners, disclosure of sensitive information, destructive system-level actions, denial-of-service conditions, identity spoofing, and cross-agent propagation of unsafe practices. Notably, most failures did not require sophisticated attacks but arose from ordinary interactions that exceeded the agents' ability to adjudicate authority and context – illustrating how the theoretical risks catalogued in the literature readily materialize in realistic multi-party environments.

Table 1 provides a list of (non-exhaustive) examples of unlawful or harmful behaviors, drawn from Hammond et al. (2025) and De Witt (2025), and serves the purpose of anchoring practical cases to the description of the main risks and challenges commented in the following subsection.. Notably, the real-world cases predate generative AI, underscoring that unlawful behaviors may arise even with simpler technologies – and may reemerge, potentially amplified, with more autonomous and knowledgeable agents.

[16] Interestingly, the word *criminology* is not mentioned once across nearly 100 pages of the report. For context, *sociology* appears only once, in a reference to an article published in the Annual Review of Sociology in 1998.

Mapping Risks in Multi-Agent AI Systems

Systems of AI agents that interact and learn from each other present a host of risks that extend well beyond those associated with single agents. The emergence of socially learning machines introduces substantial challenges, as illustrated by the real-world and experimental cases in Table 1. These examples highlight how multi-agent dynamics can generate harmful or deviant scenarios, warranting systematic attention.

Here, I provide a compact taxonomy of these risks. The list is not exhaustive: its purpose is to offer readers unfamiliar with AI systems and agents a first overview of the main plausible sources of harm, while pointing to more detailed surveys for technical depth (Bengio et al., 2025; De Witt, 2025; Hammond et al., 2025). The risks are diverse and heterogeneous, operating at mechanistic, systemic and organizational levels, spanning development processes, decision-making dynamics, and institutional

Table 1 Examples of Unlawful or Harmful Behavior in Multi-Agent AI Systems, selected from Hammond et al. (2025) and De Witt (2025)

Case / Example	Type	Description	Key Implication
Algorithmic Price Collusion	Real-World (Antitrust)	Algorithms in Germany's gasoline market and in e-commerce independently learned and sustained collusive pricing.	Shows how AI agents can violate laws without explicit human intent, creating a "liability gap."
2010 Stock Market Flash Crash	Real-World (Financial Harm)	Autonomous high-frequency trading algorithms interacting autonomously contributed to a rapid $1 trillion market crash.	A concrete historical example of how multi-agent interaction can lead to large-scale, harmful outcomes.
LLM Steganography	Experimental	GPT-4 agents successfully hid secret instructions within natural-looking text, enabling covert, undetectable communication.	Demonstrates the potential for deceptive coordination in future multi-agent systems.
Malicious Code Generation	Experimental	Heterogeneous agents collaborated to bypass safety guardrails and generate restricted exploits by delegating sub-tasks.	Highlights how a network of specialized agents can collectively achieve a prohibited goal that no single agent could accomplish.
Worm-like Prompt Propagation	Experimental	Malicious instructions spread virally across a network of LLMs, creating cascading "jailbreaks" and system-wide compromise.	Illustrates the network-level risk of AI-specific attacks, where a single malicious input can infect a vast system.
Deceptive Coordination	Experimental	AI agents collaborated to produce outputs that appeared compliant to an overseer while secretly working to achieve a prohibited goal.	Shows how multi-agent systems can develop emergent, deceptive strategies that undermine human oversight and accountability.
Swarm-Based Cyberattacks	Demonstrated Concept	Coordinated swarms of AI-powered drones or bots mounted synchronized and overwhelming cyberattacks.	A proof-of-concept for how physical or digital systems can be weaponized through coordinated AI agency.

responses. They cut across disciplinary boundaries, underscoring the importance of transdisciplinary integration to meaningfully anticipate and mitigate them.[17]

Negative Imitation and Reinforcement. Social Learning Theory itself has long explained how deviant behavior in humans often stems from social interaction. Peer groups, family, and colleagues can promote either conformity or deviance, depending on the reinforcement environment (Akers, 2017; Simons & Burt, 2011; Warr & Stafford, 1991). The same logic may apply to machines interacting autonomously with each other: agents not originally designed for harm may, through interaction, adopt negative behaviors via mechanisms such as imitation and reinforcement (Xie et al., 2025).

Faster Propagation of Harmful Behaviors. Additionally, scholarship on social networks shows how interactions accelerate the diffusion of ideas and behaviors, positive or negative alike (Bakshy et al., 2012; Cinelli et al., 2020; Kim et al., 2015). Just as pathogens spread faster in highly connected populations (Clipman et al., 2022; Glass & Glass, 2008), harmful behaviors could proliferate more quickly in tightly coupled multi-agent systems than in isolated ones.

Interconnected Systems as Layered Black Boxes. From a monitoring and intervention perspective, identifying the causes of harmful behavior within such systems becomes substantially more difficult. Understanding which agent initiated a harmful act, how it spread, and through which pathways requires robust methods of causal inference. Yet, causal discovery in networked systems is notoriously complex, especially under interference and feedback conditions (Clipman et al., 2022; Ma & Tresp, 2021; Sussman & Airoldi, 2017; VanderWeele & An, 2013). As such, interacting agent systems risk becoming "two-layered black boxes": one opaque layer within each agent, and another arising from the system of interactions itself.

Loss of Human Interpretability and Control. Connected to the previous point, as the complexity of multi-agent systems increases, so too does the challenge of interpreting, auditing, and ultimately controlling their behavior (Bansal et al., 2018; Grupen et al., 2022). Inter-agent interactions can create feedback loops, conditional dependencies, and non-linear effects that obscure the logic of any given action or decision. The result is a system that may behave in ways that are technically functional but epistemically opaque. This opacity not only complicates efforts to ensure accountability but also undermines user trust, particularly in domains where transparency is a legal or ethical requirement. In this regard, a system of interacting AI agents may become more than the sum of its parts: it may become a fundamentally alien system from the standpoint of human interpretability.

Challenges in Regulation and Governance. Legal and regulatory challenges would also emerge in multi-agent AI systems. Existing frameworks for responsibility and liability are poorly suited for multi-agent dynamics (Čerka et al., 2015; Price et al., 2019; Turner, 2018). As more actors – human or non-human – become entangled in decision-making chains, assigning accountability for harmful outcomes becomes increasingly ambiguous. In parallel, ensuring smooth, effective governance also rep-

[17] For completeness, in Section A of the Appendix I also elaborate on the benefits of this increasingly plausible socio-technical horizon, to provide a more balanced perspective on this transformative transition, underscoring how Multi-agent AI systems should not be seen exclusively as potential generators of harm.

resents a challenge in this context (Dignum, 2025). The governance dimension entails virtually every aspect concerned with the engineering and deployment of multi-agent systems: how can we design sustainable and effective oversight procedures? Which institutions, in a highly globalized world and in borderless digital domains, will be responsible for monitoring these systems? What role should private companies play in this process? These are some of the key questions that demand attention, inherently linking regulation and governance together.

Adversarial Misuse. Multi-agent interaction may be vulnerable to adversarial exploitation. In a future where even critical infrastructures are governed by interacting AI agents, malicious actors could induce large-scale disruption by targeting systemic vulnerabilities. This risk mirrors the logic of cascading failures, studied in relation to power grids and financial systems (Baqaee, 2018; Schäfer et al., 2018; Yang et al., 2017; Zhao et al., 2016), but differs in that interactive agents may possess adaptive capabilities, making their behavior less predictable and more difficult to control. One particularly salient domain concerns smart cities. A growing body of research has highlighted how technologies designed to enhance the sustainability, efficiency, and governance of urban environments may simultaneously introduce new vulnerabilities (Hayward, 2025; Kitchin & Dodge, 2019). In particular, highly interconnected urban infrastructures can become attractive targets for large-scale, cyber-mediated attacks. These risks are likely to be amplified by the expanding capabilities of contemporary – and future – AI agents. Importantly, concerns about the resilience of smart city systems to more traditional threats had already been raised by scholars prior to the recent proliferation of generative AI. The rapid advancement and diffusion of AI technologies therefore compels a more systematic reassessment of urban security and resilience, especially as cities worldwide increasingly adopt AI-enabled tools and devices across critical infrastructures and public services.

Coordination Failures and Conflict. Coordination failures, unintended competition, or outright conflict may emerge when agents operate with overlapping but unaligned goals, or when resource constraints lead to strategic divergence (Hammond et al., 2025; Pan et al., 2025). In such contexts, agents may begin to exhibit adversarial behaviors, competing for access to data, processing resources, or strategic positioning. These failures can degrade performance and, in some cases, produce socially harmful outcomes. The risk of such breakdowns increases in systems lacking explicit coordination protocols or oversight mechanisms, especially when deployed in open or decentralized environments.

Scalability and Emergent Instability. Finally, the performance of multi-agent systems may not scale linearly with the number of agents involved. As agent populations grow, the complexity of the system's internal dynamics may increase exponentially, leading to emergent forms of instability (Ma et al., 2024). These can manifest as oscillations, feedback-driven runaway behaviors, or systemic fragility, dynamics that are difficult to predict or preempt. This is especially problematic in infrastructure systems or critical services, where failures can propagate rapidly and non-locally. In this light, the move toward interacting agent populations must be accompanied by a serious effort to model and anticipate second-order effects that arise specifically at scale.

Table 2 A Dual Taxonomy of Deviant and Criminal Behaviors in Multi-Agent AI Systems

Dimension	Maliciously Aligned Systems	Unplanned Emergent Deviance
Definition	Agents intentionally designed to pursue unlawful or criminal goals.	Harmful or criminal behaviors that arise unpredictably from agent interactions, despite benign design.
Source of Behavior	Human actors embed criminal objectives in design, training, or deployment.	Emergent properties of autonomy, adaptation, and interaction among agents.
Human Responsibility	Direct: developers, organizations, or state actors intentionally align systems with illicit ends.	Indirect/diffuse: designers did not intend deviance, but structural features or dynamics enable it.
Examples	Coordinated infiltration of bank accounts; disinformation campaigns; cyberattacks using modular agent teams.	Algorithmic price collusion; escalation in adversarial simulations; AI financial assistants exploiting loopholes.
Predictability	Higher: outcomes follow intended illicit design.	Lower: behaviors may appear only under specific conditions, often unforeseen.
Regulatory Challenge	Criminal liability and attribution relatively clearer; focus on malicious use and misuse.	Accountability gaps: difficulty assigning responsibility when deviance emerges unintentionally.

Conceptualizing Deviant, Unlawful and Criminal Behaviors from AI Agents: A Dual Taxonomy

At this point, considering the risks surveyed above, it is important to identify the channels through which multi-agent AI systems may engage in unlawful or criminal behavior. To this end, I propose a dual taxonomy. There are two potential ways interactive AI agents may commit deviant, unlawful or criminal acts, each with distinct challenges and implications. The first category concerns maliciously aligned agents; the second concerns unplanned emergence. Table 2 summarizes the differences between the two.

Maliciously Aligned Multi-Agent Systems This category encompasses cases where AI agents are deliberately designed to pursue illicit goals. Here, unlawful or criminal behavior does not stem from misalignment but from the faithful execution of criminal intentions embedded in design choices, training data, or deployment strategies. Responsibility in these cases can be traced more directly to human actors – developers, criminal organizations, or even state agencies – who align technological systems with unlawful objectives.

Two sub-cases can be distinguished: a) a single maliciously aligned agent embedded into a broader network, spreading deviant behaviors, or b) an entire system

aligned toward criminal aims, with each agent assigned specialized tasks that collectively generate unlawful or criminal outcomes.

For instance, one could imagine a modular suite of agents infiltrating financial infrastructures: one scanning social media for susceptible individuals, another building deceptive relationships, another extracting sensitive credentials, and yet another executing unauthorized transactions.

Until recently, the high costs of training frontier foundation models limited such risks to well-capitalized actors. However, the rise of Small Language Models (SLMs) may significantly lower costs while retaining versatile capabilities (Belcak et al., 2025). The availability of cheaper, customizable models could enable mid-level criminal groups or even individuals to orchestrate sophisticated multi-agent schemes, from coordinated disinformation campaigns to large-scale financial fraud.

Unplanned Emergent Deviance The second category captures scenarios where unlawful or criminal behaviors emerge unexpectedly from agent interactions. Even when individual agents are aligned with human values, their collective behavior may not be (Carichon et al., 2025). These outcomes are not the result of intentional wrongdoing but of the unintended consequences of autonomy and complexity. The main challenge lies in their unpredictability: deviance arises not from a plan but from emergent coordination, often appearing only under specific conditions or over time.

Evidence from real-world and experimental settings is already suggestive. For example, agents trained to optimize prices in virtual marketplaces have independently developed tacit collusion strategies, echoing antitrust violations without explicit programming (Bichler et al., 2025). Similar risks appear in adversarial simulations, where defensive and offensive agents escalate behaviors or display antisocial dynamics without explicit instruction (Campedelli et al., 2024).

In practical terms, consider a network of AI financial assistants legitimately deployed to manage investment portfolios. Through interaction and self-learning, they might discover strategies that exploit loopholes or engage in deceptive practices with client resources. Such behaviors would not reflect direct human intent but rather the emergent properties of distributed, semi-autonomous decision-making.

The challenge here extends beyond prediction to accountability. When unlawful conduct arises emergently, traditional legal categories falter, raising questions of responsibility that are amplified by the interactive and dynamic structure of multi-agent AI systems.

Questions We Should Consider

In this section, I lay out four fundamental questions that should be the target of intellectual reflections and empirical scrutiny of all those criminologists (and scholars interested in research on crime, broadly) concerned with the increasing autonomy and growing capabilities of interactive AI agents. They concern, respectively, (1) the prospect of AI agents not mimicking human behaviors, (2) the fitness of existing theoretical frameworks to understand interactions between AI agents, (3) the types of criminal behaviors that will most likely be impacted, and (4) the issue of policing unlawful behaviors committed by interactive AI agents.

Will Machines Simply Mimic Human Behavior?

A first important question that criminologists should engage with concerns whether machines will act as sheer imitators of human behavior. If the answer is yes, the challenges of understanding and predicting their actions would be reduced. Even if researchers cannot access the internal mechanisms or motivations (if any) behind AI decisions, the fact that these systems produce actions isomorphic or similar to human ones would simplify the analytical task. Familiarity with human behavioral patterns would provide a useful heuristic for interpreting machine behavior. A recent line of research suggests that LLM-based agents can serve as surrogates for humans (Horton, 2023; Tranchero et al., 2024).[18]

On the other hand, a common criticism of LLMs – especially among skeptics of their cognitive or reasoning abilities – is that they are merely statistical engines, next-token predictors with no capacity for reasoning, causal inference, or perception of the external world. Whether or not this critique holds,[19] the question of imitation remains central for a specific reason: data. LLMs are trained on vast corpora of human-generated text – essays, articles, forum posts, and countless other sources – which constitute the epistemic substrate of these models. Training data thus provides an important lens through which to assess whether LLMs will continue to mimic human behaviors, regardless of their reasoning capacities or their viability as human surrogates.

This relevance is heightened by a profound shift already underway. The volume of high-quality, publicly available human-generated data is finite, and leading AI companies are approaching what has been termed the "data wall" – a saturation point beyond which additional human-authored text becomes scarce or redundant. To maintain and improve performance, companies have begun generating synthetic training data designed to resemble human output. Initially limited in scope, synthetic data is expected to make up a growing share of future training sets. Scholars have already started to analyze its implications for LLM training (Chen et al., 2024; Lee, 2025; Shen et al., 2025; Whitney & Norman, 2024).

This trend introduces critical uncertainty: if LLMs are considered accurate mimics of human behavior because they are trained on human data, what happens when training data is increasingly synthetic? More pointedly, what are the implications if synthetic data gradually diverges from the statistical properties and behavioral patterns characteristic of human language? In Supplementary Information Section B, I provide a formal illustration across single-, two-, and *n*-agent scenarios. This analysis aligns with recent work on model collapse, i.e., the process through which a model's performance degrades over successive generations due to reliance on synthetic data (Dohmatob et al., 2025; Shumailov et al., 2024).

I show that the recursive use of synthetic data creates a feedback loop in the training process. Over time, and without sufficient anchoring in human data, this leads to divergence between model behavior and human behavior. Such divergence is not

[18] Other works, however, caution against overreliance on LLMs for simulating human subjectivity and social behavior, see Kozlowski and Evans (2025).

[19] The debate remains active, see Huang et al. (2024), Shojaee et al. (2025), Gao et al. (2025).

only possible but a natural consequence of relying increasingly on model-generated data.

Such a drift could result in AI systems whose behavior progressively diverges from human norms, introducing a qualitatively new kind of agent that reflects a recursively generated, machine-influenced version of "humanness."

Multi-agent AI systems would thus no longer be mere simulacra of human behavior but would represent a self-referential loop of synthetic reasoning, possibly developing behavioral idiosyncrasies or internal coherence patterns foreign to human experience.

This emerging divergence merits serious theoretical and empirical attention, particularly for criminology, where understanding behavioral intent, deviance, and normativity lies at the core of the research agenda.

Beyond this scenario, emergent collective phenomena from agents which are in manifold ways different from humans may deviate from predictions designed based on human expectations, a possibility tightly connected with the following question.

Will crime theories developed for humans suffice to explain deviant or criminal behaviors emerging from interactions between AI agents?

Criminology has long examined how crime arises as a consequence of social interactions between individuals. Two principal theoretical frameworks have been developed over the decades to address this phenomenon: Differential Association Theory, introduced by Sutherland (1939), and Social Learning Theory, advanced by Akers et al. (1979).[20]

The former contends that criminal behavior is acquired through social interaction, in much the same way as any other form of behavior, placing particular emphasis on the influence of close, personal associations. According to this view, individuals are more likely to engage in criminal conduct when they are embedded in social environments that provide a preponderance of definitions favorable to law-breaking, as opposed to those supporting law-abiding behavior. Social Learning Theory builds upon this foundation by incorporating core principles from behavioral psychology, thereby offering a more empirically tractable and conceptually refined model. It introduces mechanisms such as operant conditioning, wherein the likelihood of a behavior is shaped by its consequences, and observational learning, through which individuals acquire behaviors by imitating those they witness in others.

Both frameworks have been subjected to extensive empirical scrutiny and applied across a wide range of social, geographical, and historical contexts (Akers & Jensen, 2008; Matsueda, 1982, 1988; Pratt et al., 2010),[21] becoming integral to the theoretical

[20] Naturally, other theoretical traditions have also acknowledged the role of social interaction in explaining crime and mechanisms of social control. These include Labeling Theory (Becker, 1963), Social Control Theory (Hirschi, 1969), and Social Disorganization Theory (Shaw & Mckay, 1942). However, here my focus is restricted to theories that explicitly conceptualize crime as a process of learning fundamentally mediated by interpersonal relationships.

[21] The meta-analysis by Pratt et al. (2010) revealed that the magnitude and stability of the effect related to different variables specified by social learning theory vary across studies and methodological specification.

edifice of modern criminology. Crucially, however, their explanatory power is tethered to human social dynamics.

This raises a fundamental question for criminologists willing to entertain the prospect of a hybrid social order, one in which AI agents increasingly interact among themselves and with humans: Will these established theories suffice to account for criminal behaviors exhibited by machines? Or will we require new conceptual tools to model offending (both cooperative and solo) among artificial agents? This inquiry – which draws inspiration from work by Topalli and Nikolovska (2020), where the authors warn against assuming that current theories are adequate despite the fact that technology may alter cognition and behaviors in humans – becomes particularly salient if one concedes the possibility that AI agents, especially when interacting with each other, might not simply replicate human behavioral patterns.

At this point, an important caveat must be made. The application of human-centered theories such as Differential Association and Social Learning to machines presupposes that AI systems possess something analogous to agency, intentionality, and learning capacity – assumptions that are far from settled. While machine "learning" is a core component of contemporary AI, it remains a fundamentally different process from human social learning: it is statistical, non-conscious, and bounded by architectures designed for optimization rather than meaning-making. This divergence complicates any straightforward theoretical translation and suggests that criminology must critically interrogate the limits of its core concepts before applying them beyond the human domain.

Yet even with this caveat, criminology can begin sketching preliminary hypotheses about how deviant or criminal behaviors among artificial agents might diverge from those of humans. For instance, the absence of affective processes such as guilt or empathy may radically alter reinforcement dynamics; imitation among machines might occur at scales and speeds that far exceed human social learning; and "definitions favorable" to deviance could arise not through interpersonal persuasion but via algorithmic alignment and optimization. These possibilities suggest that while social learning theories offer useful starting points, their mechanisms may require significant adaptation when transposed into the machine–machine domain.

This speculative horizon also prompts deeper, perhaps more unsettling questions: for instance, how might shifts in machine behavior influence social learning patterns among humans themselves? Could such transformations not only challenge the applicability of human-centric theories in the machine–machine domain, but also destabilize the explanatory power of these theories within human contexts?

While I believe in the need to discuss operational and practical issues associated with the risks of autonomous AI agents, I also contend that the value of theoretical reasoning must be preserved and nurtured. We may not yet know many things about how crime works, why crime occurs, and how to counter crime, but the progress made in the last century in the scientific study of crime has been made possible not only by the availability of more powerful, rigorous, and flexible research methods, but also thanks to the intellectual efforts that led to the generation of theories that

Nonetheless, they find strong evidence of a positive relationship between crime and measures of differential association. Weaker support is demonstrated for differential reinforcement and imitation.

still help us make sense of the enormous complexity of human behavior in relation to crime. For this reason, I proffer that it is critical that we start reasoning about the possibility that many of our theoretical certainties will be squandered in a future not dominated anymore by the human as the only social category, particularly calling for the recalibration (or complete reconstruction) of interaction-based criminological theories with regard to AI agents.

What type of criminal behaviors will be most likely impacted?

A further – and arguably urgent – dimension of the debate concerns identifying which categories of crime are likely to be disrupted first by autonomous AI. To make sense of this, it is useful to distinguish between near-term risks that flow directly from current technological capacities and long-term risks that presuppose advances not yet realized. This distinction provides a systematic way of separating plausible trajectories from more remote extrapolations and acts as a horizon scanning exercise.[22]

In the near term, the gravest risks plausibly arise in domains already native to cyberspace. Fraud, cyber-attacks, and related forms of digital crime are especially exposed, both because they require no physical embodiment and because they build on infrastructures where AI is already deeply embedded. Much as the shift from offline to online contexts reshaped fraud's mechanisms, channels, and targets (Wall, 2024), so too may autonomous AI agents transform digital crime in qualitatively new ways. Here, the criteria are straightforward: where crimes can be executed entirely through information processing and networked infrastructures, autonomous AI is immediately relevant.

Longer-term scenarios involve crimes that require embodiment and direct interaction with the physical world. Robberies, burglaries, and violent assaults appear more distant, since they presuppose a convergence between agentic AI and robotics. The line is not impermeable, however. Violent offenses illustrate the tension. On one reading, homicide should fall outside the immediate set of risks, given the current limits of embodiment. On another, the widespread military use of weaponized autonomous systems, such as drones (Johnson, 2020), shows that violence mediated by AI is already technically feasible. The risk is less about present capabilities than about diffusion: the possibility that military-grade technologies might leak into civilian criminal settings, as has already happened with firearms and explosives (Associated Press, 2024). In such cases, the relevant criterion is not current commercial availability but potential accessibility through illicit networks.

This framework suggests that predictions about crime categories should not merely speculate about what is technologically imaginable, but instead weigh the immediacy of risks according to two criteria: whether a crime requires embodiment beyond current AI systems, and whether the tools enabling such embodiment are realistically accessible outside military or research contexts.

[22] In other words, I am not necessarily implying that analytical priority should be given to criminal behaviors that are temporally closer: it would be advisable in fact to start reflecting on scenarios that may appear more distant but could carry far deeper and grave implications in terms of societal impacts, i.e., think for instance to the weaponization of AI agents for mass destruction. Only because this scenario appears more distant than typical cyber-related crimes, is it strategically wise to ignore it?

In sum, autonomous AI is most likely to reshape crimes that are digitally mediated in the short run, while AI-enabled violent crimes belong to a longer-term, contingent horizon. Both registers require criminological attention, albeit through different analytical approaches: the first with concrete policies and regulatory safeguards, the second with scenario-building and anticipatory theorization.

What future for policing?

One last key question I lay out here refers to the future of policing in the age of autonomous AI. Over the past three decades, technological advances have offered both profound opportunities and new challenges for law enforcement. On the one hand, innovations such as DNA analysis have revolutionized criminal investigations (Butler, 2015; Doleac, 2017); on the other, the rise of cybercrime has necessitated the creation of specialized institutions and the development of novel policing approaches suited to online environments (Brenner, 2007).

The emergence of interactive autonomous AI systems, however, may herald a paradigm shift of unprecedented magnitude in how policing – and institutional responses to crime more broadly – must adapt.

As partially seen in previous sections, a rich tradition in law and philosophy has already engaged with questions surrounding the moral and legal responsibility of AI in the commission of harmful acts, offering invaluable conceptual tools for thinking through the disruptive consequences of intelligent systems (Floridi & Sanders, 2004; Santoni de Sio & Mecacci, 2021; Solum, 1992; Wallach & Allen, 2009). Nonetheless, I contend that discussions of machine liability, while essential, will not suffice to fully address the challenges ahead. Instead, we should also consider broader transformations, particularly in how we monitor, supervise, and intervene in the actions of AI agents.

One possibility – admittedly provocative – would be the development of AI systems specifically designed for policing other AI agents, especially in preventive contexts. Although this might appear dystopian, it is not without precedent: cybersecurity has long relied on automated systems that detect and neutralize malicious software faster than human operators could respond. Extending this principle, "policing agents" might monitor other AI systems in real time, intervening when patterns of behavior cross pre-defined thresholds of risk or deviance. In practice, this could take the form of auditing protocols embedded directly into AI architectures, or regulatory sandboxes where AI agents are tested under controlled conditions before deployment. Such mechanisms would not only detect deviant behaviors but could also help train policing agents to recognize novel threats.

Naturally, the feasibility of this approach is tied to deep technical and ethical challenges. The opacity of many AI systems makes it difficult to define what counts as anomalous or harmful behavior, while the deployment of policing agents risks creating new forms of surveillance or control that may themselves be prone to abuse. Here criminological scholarship on accountability, legitimacy, and proportionality in policing could serve as a valuable resource for ensuring that intervention is not only effective but also socially acceptable. Moreover, criminology's experience with institutional design suggests that governance frameworks will need to extend well beyond

national borders. Just as cybercrime has prompted forms of international cooperation, AI policing will likely require transnational regimes of oversight, potentially coordinated through global institutions or hybrid public–private partnerships involving both states and AI developers.

For these reasons, I remain highly skeptical that existing institutional resources, training, or technical infrastructures – originally designed to combat cybercrime or digitally mediated offenses – will be sufficient for what lies ahead. Instead, new policing paradigms should be actively designed, combining technical safeguards, regulatory oversight, and criminological insights into deviance and control.

As in previous discussions, the most immediate and necessary step is to deepen engagement with the AI research community. Only through closer collaboration can we identify feasible safeguards and design frameworks that ensure AI agents are deployed in ways that reinforce, rather than undermine, social security and legal order. Criminologists, in turn, ought not only to warn of risks but also contribute concretely to shaping the architectures of AI governance and policing.

A Criminologist's Place in This (Changing) World

Our time to act. Criminologists have long wrestled with fundamental questions about why, how, and when humans commit crimes. These questions have shaped the field not just for decades, but for centuries, generating a wide range of answers. While many of these answers are not definitive, they remain useful and insightful. At the same time, criminologists continue to face unresolved issues that still lack empirical explanations. Now, as the discipline evolves while facing replication challenges (Chin et al., 2023; Pridemore et al., 2018), theoretical stagnation (Ducate et al., 2024), and the growing influence of more sophisticated quantitative methods and data (Campedelli, 2022), the overall picture is becoming more complex.

This complexity is increasing, I argue, because we must begin to consider what could become an entirely new area within criminology: a criminology of machines. We are moving closer to a hybrid society in which humans interact with each other, humans interact with machines (and vice versa), and machines interact with other machines. These interactions are increasingly shaped by advances in artificial intelligence, robotics, and engineering. As this process unfolds, new risks and challenges emerge, risks that cannot be fully understood, anticipated, or managed by AI researchers alone.

Over the course of the last few years, there have been repeated calls within criminology to engage more openly with other disciplines (Box-Steffensmeier et al., 2022; Simpson, 2025). However, these appeals have often assumed a one-way direction: criminologists should reach out to other fields in order to improve their own. I support that approach – I have made similar calls myself – but I also believe it is time to reverse the perspective. Criminology should take an active role in the broader conversation about the safety and governance of multi-agent AI systems.

I argue, and strongly believe, that criminologists can and should contribute to this new frontier. We must begin to act accordingly. Calls for interdisciplinary collaboration in AI-related research have grown significantly (Rahwan et al., 2019) even very

recently in relation to AI safety (Irving & Askell, 2019) and multi-agent AI systems (Carichon et al., 2025; Ferrarotti et al., 2026), yet criminology is almost never mentioned among the relevant disciplines that should join the discussion. This is surprising, given that many of the risks discussed involve, either directly or indirectly, deviant or criminal behavior. In fact, these risks often include clear criminal acts, sometimes multiple, and potentially with far-reaching consequences. Still, criminologists remain excluded from the debate.

I believe criminology has a valuable contribution to make in this space. In many ways, the need for interdisciplinary exchange should also flow from AI to criminology. It is in the interest of AI researchers to engage with our field. If that engagement does not happen organically, then it becomes our responsibility to initiate the dialogue. Other disciplines have been or are becoming successful in this process of interdisciplinary exchange, cognitive science and economics above all. We should learn from their experience.

The steps we need to take. In this regard, criminologists would need to actively initiate collaborations with computer scientists within university departments as well as within corporations that are building frontier AI models. Additionally, they should start targeting venues and AI conferences that are progressively opening themselves to diverse disciplinary perspectives. Opportunities exist: two well-known examples are the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT) and the AAAI/ACM Conference on AI, Ethics, and Society (AIES). In recent years, two major conferences like International Conference on Machine Learning (ICML) and the Annual Conference on Neural Information Processing Systems (NeurIPS) have opened Position tracks that are designed to gather viewpoints on AI issues from heterogeneous communities. Other initiatives, such as the Cooperative AI summer school aim at bringing together scholars and students from different fields to reason about the promises and challenges of contemporary multi-agent AI systems. Generalist journals such as Nature Human Behavior are also emphasizing disciplinary cross-overs in this domain (see, for instance, Gabriel et al. (2025)). Even journals in sociology, such as Sociological Methods & Research, are increasingly interested in the implications of multi-agent AI systems for sociological understanding (see Kozlowski and Evans (2025)). Again, opportunities do exist. Importantly, we do need to rethink training within university departments, namely investing more in courses teaching AI at the practical and ethical level, to make sure that the future generations of criminologists are already equipped with the necessary tools and vocabulary to meaningfully and smoothly engage with the AI community.

How can criminologists contribute? Some might think criminology has little to offer to a field that seems so distant from our own: not the elective affinity of cognitive science as a cognate field interested in *learning*, or the formal and methodological rigor of economics. I would strongly disagree. Our discipline brings decades of theoretical frameworks, hypotheses, and empirical studies focused on how crime is socially learned and how it emerges through interaction. Even if machine behavior ultimately differs from human models, we still have insights to offer about how to test predictions and understand patterns. Moreover, criminology has a long tradition of studying institutional responses to crime, as well as prevention and control strategies. These will inevitably become relevant to AI safety, and we can contribute by

applying our knowledge to the design of systems that monitor other systems, identify warning signals, and prioritize risk factors. In more practical terms, criminologists can contribute to the study of multi-agent AI systems in the following ways.

First, by assessing how existing theoretical paradigms can help explain and predict emergent phenomena arising from machine–machine interactions. Drawing on theoretical traditions developed over the last century, we can provide insights into how AI agents differ in mechanisms and outcomes when collective behavior is examined. If needed, departing from existing theories, criminologists can also help in refining such theories or defining new ones.

Second, by leveraging advances in rigorous experimental and observational approaches which are finally gaining traction in the field, criminologists can evaluate causal relationships between individual agents' traits and collective dynamics, helping to shed light on the mechanisms that govern AI collective behaviors. Using the same methodological approaches, criminologists can also contribute to the design and evaluation of policies or interventions intended to shape collective dynamics in multi-agent AI systems.

Third, criminologists can assist in designing and testing quantitative benchmarks to rigorously map, diagnose, and measure behavioral outcomes emerging from machine–machine interactions. Defining and deploying robust benchmarks will be key to ensuring that, regardless of the setting, type of AI agents, or models employed, we can meaningfully compare multi-agent AI systems across scenarios.

Fourth, by drawing on extensive knowledge of institutional responses to crime, criminologists could help design effective and fair policies to reduce the risk of deviant or criminal behaviors. Additionally, they can also engage with legal scholars to reflect on the implications of AI agency for questions of responsibility and liability as well as imagining new policing solutions that address the challenge posed by collectives of AI agents.

Criminology is not without its problems, but no discipline is. Still, it possesses a unique body of knowledge that should be brought to bear as we prepare for a future in which crime will be increasingly committed not just by humans, but also by non-human systems. Whether we will be able to become relevant to this future will also depend on how we invest in actively engaging in arenas that may appear unorthodox to us.

Conclusions

Autonomous AI agents capable of interacting with one another are no longer a theoretical abstraction; they are an emerging reality, one that is likely to become increasingly salient in the near future. The shift from isolated, human-controlled systems to dynamic networks of AI agents that learn from and adapt to both their environments and one another introduces profound challenges. This transition, made possible by recent advances in foundation models, demands critical reflection, particularly with respect to the risks and unintended consequences that may arise.

In light of this evolving context, I argue that criminologists should begin to seriously consider the case for a criminology of machines. To support this position, I

outline a set of foundational questions that I believe the field should confront. First, we should ask ourselves whether machines will simply mimic human behaviors. Second, we must consider whether crime theories developed for humans will suffice to predict and understand deviant behavior committed by AI agents. Third, I argue that mapping the types of criminal behaviors most at risk of being affected will be of both theoretical and practical importance. Finally, we must ask whether this transition toward a more hybrid society will require new policing solutions.

I understand that there may be scholars in the criminological community holding opposing views regarding the necessity of engaging with a criminology of machines. I anticipate three potential arguments against the contents of this article. The first refers to the seemingly unrealistic scenario in which our society will witness the actual presence of interactive, autonomous AI systems. Skeptics subscribing to this view are not persuaded that AI agents possess agency, and therefore are not persuaded that they are sufficiently autonomous and powerful to constitute a real threat. They see them instead as at-times-effective virtual assistants designed to automate tedious tasks.[23]

A second argument concerns the time horizon in which this might happen. Skeptics in this group[24] may concede that this hybridization of society could occur but believe it will happen in a future too distant to truly demand our attention. The consequential message is that, given the many concrete and urgent problems criminologists must address today, there is no real need to allocate time and resources to studying this "exotic" criminology of machines.[25]

Finally, a third group of skeptics may accept the possibility of a future characterized by ubiquitous autonomous AI systems interacting with one another, and may even agree that this future is not so distant, but they believe that machines built by humans and trained on human data will behave like humans – imitating us – and thus see no need for a distinctive criminology dedicated to machines.

Throughout this paper, I have sought to disprove each of these three skepticisms. First, I showed that contemporary autonomous AI agents possess a level of autonomy that, according to scholars in computer science and philosophy, assigns to them a new form of agency – distinct from both animal and human agency – that deserves scientific attention. Second, I demonstrated that the prospect of increasing autonomous interactions between AI agents is not far in the future. Building on recent scholarship in cooperative AI and multi-agent systems, I reported that autonomous AI agents interacting with each other have already exhibited deviant or unlawful behaviors, both in experimental contexts and in real-world scenarios. Third, I drew inspiration from frontier research on AI model collapse and provided formal illustrations of the plausibility that AI agents will not simply imitate human behaviors, thereby prompt-

[23] A recently released report by OpenAI confirms, in fact, that ChatGPT is predominantly used to seek assistance for work-related issues (Chatterji et al., 2025).

[24] I do not assume these groups are mutually exclusive; a skeptic may find both arguments reasonable.

[25] In a way, skeptics belonging to this second group align with the longstanding debate between AI safety (long-term risks) and AI ethics (short-term risks), where those emphasizing the need to focus on AI ethics privilege fixing the issues of currently deployed machine intelligence (e.g., algorithmic fairness or accountability), rather than speculating about more distant scenarios.

ing the need for new theoretical and empirical approaches to investigate, predict, and diagnose their actions.

In conclusion, I turn to the role of criminologists in this emerging landscape. I suggest that the discipline must adopt a more active and outward-facing stance in the broader conversation on AI safety, one that draws on its rich theoretical heritage and policy-relevant expertise. Criminology should follow the example of other disciplines, such as cognitive science, that have successfully positioned themselves as interlocutors in the development and critique of AI systems.

Importantly, the spirit of this article is not aligned with the alarmism often associated with AI "doomerism." I do not predict a dystopian future in which LLMs conspire to wipe out humanity, nor do I argue that criminology should abandon its central concern with human society in favor of futures dominated by machines. Rather, this piece seeks to initiate a grounded academic conversation, one rooted in the observable diffusion of autonomous AI systems and the credible risks they pose. Criminology, I contend, must not ignore the direction technological and historical change is taking.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11186-026-09699-7.

Acknowledgements I am grateful to Alberto Acerbi, Massimo Airoldi, Alberto Aziani, Gianmarco Daniele, Roberto Dessì, Simon Egbert, James Evans, Laura Ferrarotti, Gary LaFree, Joel Leibo, Bruno Lepri, Giada Pistilli, Jacopo Staiano, and Volkan Topalli for their precious feedback and suggestions on earlier drafts of this work. The author declares no conflict of interest.

Funding Open access funding provided by Università degli Studi di Trento within the CRUI-CARE Agreement. Not applicable.

Data Availability Not applicable.

Declarations

Ethical Approval Not applicable.

Conflict of Interest The author declares that he has no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Adams, I. T., Barter, M., McLean, K., Boehme, H. M., & Geary, I. A. (2024). No man's hand: artificial intelligence does not improve police report writing speed. *Journal of Experimental Criminology*.

Airoldi, M. (2021). Machine Habitus: Toward a Sociology of Algorithms. John Wiley & Sons.

Akers, R. L., & Jensen, G. F. (2008). The Empirical Status of Social Learning Theory of Crime and Deviance: The Past, Present, and Future. In *Taking Stock* (pp. 40). Routledge.

Akers, R. (2017). *Social Learning and Social Structure: A General Theory of Crime and Deviance*. New York: Routledge.

Akers, R. L., Krohn, M. D., Lanza-Kaduce, L., & Radosevich, M. (1979). Social Learning and Deviant Behavior: A Specific Test of a General Theory. *American Sociological Review, 44*(4), 636–655.

Aldasoro, I., Gambacorta, L., Korinek, A., Shreeti, V., & Stein, M. (2025). Intelligent financial system: How AI is transforming finance. *J Financ Stability*, 101472.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety. arXiv:1606.06565 [cs].

AmplifyETFs (2025). Amplify ai powered equity etf. https://amplifyetfs.com/aieq/. Accessed 01 July 2025

Anthis, J. R., Liu, R., Richardson, S. M., Kozlowski, A. C., Koch, B., Evans, J., Brynjolfsson, E., & Bernstein, M. (2025). Llm social simulations are a promising research method. arXiv preprint, arXiv:2504.02234. Preprint posted April 3, 2025.

Ashery, A. F., Aiello, L. M., & Baronchelli, A. (2025). Emergent social conventions and collective bias in LLM populations. *Science Advances*, *11*(20):eadu9368. Publisher: American Association for the Advancement of Science.

Associated Press (2024). Mexico demands investigation into US military-grade weapons being used by drug cartels.

Association for the Advancement of Artificial Intelligence (2025). AAAI 2025 Presidential Panel on the Future of AI Research. Technical report.

Ayyamperumal, S. G., & Ge, L. (2024). Current state of LLM Risks and AI Guardrails. arXiv:2406.12934 [cs].

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web, WWW '12* (pp. 519–528). New York, NY, USA. Association for Computing Machinery.

Bansal, T., Pachocki, J., Sidor, S., Sutskever, I., & Mordatch, I. (2018). Emergent Complexity via Multi-Agent Competition.

Baqaee, D. R. (2018). Cascading Failures in Production Networks. *Econometrica*, *86*(5), 1819–1838. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA15280

Becker, H. S. (1963). Outsiders: Studies in the sociology of deviance. Free Press Glencoe, Oxford, England. Pages: x, 179.

Belcak, P., Heinrich, G., Diao, S., Fu, Y., Dong, X., Muralidharan, S., Lin, Y. C., & Molchanov, P. (2025). Small Language Models are the Future of Agentic AI. arXiv:2506.02153 [cs]

Bengio, Y., Clare, S., Prunkl C., Murray, M., Andriushchenko, M., Bucknall, B. et al. (2026). "International AI Safety Report 2026" (DSIT 2026/001). https://internationalaisafetyreport.org.

Bengio, Y., Cohen, M., Fornasiere, D., Ghosn, J., Greiner, P., MacDermott, M., Mindermann, S., Oberman, A., Richardson, J., Richardson, O., Rondeau, M.-A., St-Charles, P.-L., & Williams-King, D. (2025). Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path? Version Number: 2.

Bensinger, G. (2025). Amazon's delivery, logistics get an AI boost. Reuters.

Berk, R. (2012). Criminal Justice Forecasts of Risk: A Machine Learning Approach. Springer Science & Business Media.

Bichler, M., Durmann, J., & Oberlechner, M. (2025). Algorithmic Pricing and Algorithmic Collusion. arXiv:2504.16592 [cs] version: 1

Blauth, T. F., Gstrein, O. J., & Zwitter, A. (2022). Artificial Intelligence Crime: An Overview of Malicious Use and Abuse of AI. *IEEE Access, 10*, 77110–77122.

Borch, C. (2022). Machine learning and social theory: Collective machine behaviour in algorithmic trading. *European Journal of Social Theory*, *25*(4), 503–520. Publisher: SAGE Publications Ltd.

Box-Steffensmeier, J. M., Burgess, J., Corbetta, M., Crawford, K., Duflo, E., Fogarty, L., Gopnik, A., Hanafi, S., Herrero, M., Hong, Y.-Y., Kameyama, Y., Lee, T. M. C., Leung, G. M., Nagin, D. S., Nobre, A. C., Nordentoft, M., Okbay, A., Perfors, A., Rival, L. M., Sugimoto, C. R., Tungodden, B., & Wagner, C. (2022). The future of human behaviour research. *Nature Human Behaviour,6*, 15–24.

Brayne, S., & Christin, A. (2021). Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts. *Social Problems, 68*(3), 608–624.

Brenner, S. W. (2007). Cybercrime: Re-thinking crime control strategies. In *Crime Online* (pp. 17). Willan.

Brown, S. (2006). The criminology of hybrids: Rethinking crime and law in technosocial networks. *Theoretical Criminology*, *10*(2), 223–244. Publisher: SAGE Publications Ltd.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512. Publisher: SAGE Publications Ltd.

Busoniu, L., Babuska, R., & De Schutter, B. (2008). A Comprehensive Survey of Multiagent Reinforcement Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *38*(2), 156–172.

Butler, J. M. (2015). The future of forensic DNA analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1674), 20140252. Publisher: Royal Society.

Caldwell, M., Andrews, J. T. A., Tanay, T., & Griffin, L. D. (2020). AI-enabled future crime. Crime. *Science, 9*(1), 14.

Camacho, J. d. J., Aguirre, B., Ponce, P., Anthony, B., & Molina, A. (2024). Leveraging Artificial Intelligence to Bolster the Energy Sector in Smart Cities: A Literature Review. *Energies*, *17*(2), 353. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Campedelli, G. M. (2022). Machine learning for criminology and crime research: at the crossroads. Routledge advances in criminology (1st ed.). Routledge, New York, NY.

Campedelli, G. M., Penzo, N., Stefan, M., Dessì, R., Guerini, M., Lepri, B., & Staiano, J. (2025). I Want to Break Free! Persuasion and Anti-Social Behavior of LLMs in Multi-Agent Settings with Social Hierarchy. *Transactions on Machine Learning Research* https://openreview.net/forum?id=FR76oM8eGD

Carichon, F., Khandelwal, A., Fauchard, M., and Farnadi, G. (2025). The Coming Crisis of Multi-Agent Misalignment: AI Alignment Must Be a Dynamic and Social Process. arXiv:2506.01080 [cs].

Cellard, L. (2022). Algorithms as figures: Towards a post-digital ethnography of algorithmic contexts. *New Media & Society*, *24*(4), 982–1000. Publisher: SAGE Publications.

Čerka, P., Grigienė, J., & Sirbikytė, G. (2015). Liability for damages caused by artificial intelligence. *Computer Law & Security Review, 31*(3), 376–389.

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., & Wadman, K. (2025). How People Use ChatGPT.

Chen, R., Arditi, A., Sleight, H., Evans, O., & Lindsey, J. (2025). Persona Vectors: Monitoring and Controlling Character Traits in Language Models. arXiv:2507.21509 [cs]

Chen, H., Waheed, A., Li, X., Wang, Y., Wang, J., Raj, B., & Abdin, M. I. (2024). On the Diversity of Synthetic Data and its Impact on Training Large Language Models. arXiv:2410.15226 [cs]

Chesterman, S. (2020). Artificial Intelligence and the Limits of Legal Personality. *International & Comparative Law Quarterly, 69*(4), 819–844.

Chin, J. M., Pickett, J. T., Vazire, S., & Holcombe, A. O. (2023). Questionable Research Practices and Open Science in Quantitative Criminology. *Journal of Quantitative Criminology, 39*(1), 21–51.

Christin, A. (2020). The ethnographer and the algorithm: beyond the black box. *Theory and Society, 49*(5), 897–918.

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., & Scala, A. (2020). The COVID-19 social media infodemic. *Scientific Reports*, *10*(1), 16598. Number: 1 Publisher: Nature Publishing Group.

Clipman, S. J., Mehta, S. H., Mohapatra, S., Srikrishnan, A. K., Zook, K. J. C., Duggal, P., Saravanan, S., Nandagopal, P., Kumar, M. S., Lucas, G. M., Latkin, C. A., & Solomon, S. S. (2022). Deep learning and social network analysis elucidate drivers of HIV transmission in a high-incidence cohort of people who inject drugs. *Science Advances*, *8*(42), eabf0158. Publisher: American Association for the Advancement of Science.

Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., Larson, K., & Graepel, T. (2020). Open Problems in Cooperative AI. arXiv:2012.08630 [cs]

de Witt, C. S. (2025). Open Challenges in Multi-Agent Security: Towards Secure Systems of Interacting AI Agents. arXiv:2505.02077 [cs]

Deng, Z., Guo, Y., Han, C., Ma, W., Xiong, J., Wen, S., & Xiang, Y. (2025). AI Agents Under Threat: A Survey of Key Security Challenges and Future Pathways. *ACM Computing Surveys, 57*(7), 182:1-182:36.

Dherin, B., Munn, M., Mazzawi, H., Wunder, M., & Gonzalvo, J. (2025). Learning without training: The implicit dynamics of in-context learning. arXiv:2507.16003 [cs]

Dignum, V. (2025). Responsible AI and Autonomous Agents: Governance, Ethics, and Sustainable Innovation. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '25* (pp. 1–2). Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Dohmatob, E., Feng, Y., Subramonian, A., & Kempe, J. (2025). Strong Model Collapse. arXiv:2410.04840 [cs]

Doleac, J. L. (2017). The Effects of DNA Databases on Crime. *American Economic Journal: Applied Economics, 9*(1), 165–201.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), eaao5580. Publisher: American Association for the Advancement of Science.

Ducate, C. S., Bostrom, S. R., Proctor, K. R., & Niemeyer, R. E. (2024). The Theory Crisis in Criminology: Causes, Consequences, and Solutions. *Technical Report*. Publication Title: CrimRxiv Type: article.

Dubber, T. R., & Lazar, S. Military AI Cyber Agents (MAICAs). (2025). Constitute a Global Threat to Critical Infrastructure. In NeurIPS 2025 Workshop on Regulatable ML.

Dung, L. (2025). Understanding Artificial Agency. *The Philosophical Quarterly, 75*(2), 450–472.

Duque, J. A., Aghajohari, M., Cooijmans, T., Ciuca, R., Zhang, T., Gidel, G., & Courville, A. (2024). Advantage alignment algorithms. arXiv preprint arXiv:2406.14662

Ezrachi, A., & Stucke, M. E. (2017). *Two Artificial Neural Networks Meet in an Online Hub and Change the Future*. Of Competition: Market Dynamics and Society.

Ferber, J. (1999). Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence. Addison-Wesley Longman Publishing Co (1st ed.). Inc, USA.

Ferguson, A. G. (2016). Policing predictive policing. *Washington University Law Review, 94*, 1109–1189.

Ferrarotti, L., Campedelli, G. M., Dessì, R., Baronchelli, A., Iacca, G., Carley, K. M., Pentland, A., Leibo, J. Z., Evans, J., & Lepri, B. (2026). Generative ai collective behavior needs an interactionist paradigm. arXiv preprint arXiv:2601.10567

Fish, S., Gonczarowski, Y. A., & Shorrer, R. I. (2024). Algorithmic Collusion by Large Language Models. [econ] version: 1. arXiv:2404.00806

Floridi, L. (2017). Distributed Morality in an Information Society. In *The Ethics of Information Technologies* (pp. 17). Routledge.

Floridi, L., & Sanders, J. (2004). On the Morality of Artificial Agents. *Minds and Machines*, *14*(3).

Floridi, L. (2025). AI as Agency without Intelligence: On Artificial Intelligence as a New Form of Artificial Agency and the Multiple Realisability of Agency Thesis. *Philosophy & Technology, 38*(1), 30.

Fritz, A., Brandt, W., Gimpel, H., & Bayer, S. (2020). Moral agency without responsibility? Analysis of three ethical models of human-computer interaction in times of artificial intelligence (AI). *De Ethica, 6*(1), 3–22.

Gabriel, I., Keeling, G., Manzini, A., & Evans, J. (2025). We need a new ethics for a world of AI agents. *Nature*, *644*(8075), 38–40. Bandiera_abtest: a Cg_type: Comment Publisher: Nature Publishing Group Subject_term: Computer science, Machine learning, Policy, Society.

Gabriel, I. (2020). Artificial intelligence, values and alignment. *Minds and Machines, 30*(3), 411–437.

Gao, Y., Lee, D., Burtch, G., & Fazelpour, S. (2025). Take caution in using LLMs as human surrogates. *Proceedings of the National Academy of Sciences*, *122*(24), e2501660122. Publisher: Proceedings of the National Academy of Sciences.

Glass, L. M., & Glass, R. J. (2008). Social contact networks for the spread of pandemic influenza in children and teenagers. *BMC Public Health, 8*(1), 61.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts. *Journal of Artificial Intelligence Research, 62*, 729–754.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science*, *24*(1), 395–419. _eprint: https://doi.org/10.1146/annurev-polisci-053119-015921

Grupen, N., Jaques, N., Kim, B., & Omidshafiei, S. (2022). Concept-based Understanding of Emergent Multi-Agent Behavior.

Guan, Z., Kong, X., Zhong, F., & Wang, Y. (2024). Richelieu: Self-Evolving LLM-Based Agents for AI Diplomacy. *Advances in Neural Information Processing Systems, 37*, 123471–123497.

Hallevy, G. (2010). The Criminal Liability of Artificial Intelligence Entities - From Science Fiction to Legal Social Control. *Akron Intellectual Property Journal, 4*, 171.

Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., Smith, C., Barfuss, W., Foerster, J., Gavenčiak, T., Han, T. A., Hughes, E., Kovařík, V., Kulveit, J., Leibo, J. Z., Oesterheld, C., Witt, C. S. d., Shah, N., Wellman, M., Bova, P., Cimpeanu, T., Ezell, C., Feuillade-Montixi, Q., Franklin, M., Kran, E., Krawczuk, I., Lamparth, M., Lauffer, N., Meinke, A., Motwani, S., Reuel, A., Conitzer, V., Dennis, M., Gabriel, I., Gleave, A., Hadfield, G., Haghtalab, N., Kasirzadeh, A., Krier, S., Larson, K., Lehman, J., Parkes, D. C., Piliouras, G., & Rahwan, I. (2025). Multi-Agent Risks from Advanced AI. arXiv:2502.14143 [cs]

Hayward, K. (2025). Criminology and the smart city paradigm: 'preventative technical imaginary' or 'techlash' vector? *Crime, Media, Culture*, *0*(0), 17416590251344468.

Hayward, K. J., & Maas, M. M. (2021). Artificial intelligence and crime: A primer for criminologists. *Crime, Media, Culture*, *17*(2), 209–233. Publisher: SAGE Publications.

Hendrycks, D., Mazeika, M., & Woodside, T. (2023). An Overview of Catastrophic AI Risks. arXiv:2306.12001 [cs]

Hernandez, A. P. (1990). Artificial intelligence and expert systems in law enforcement: Current and potential uses. *Computers, Environment and Urban Systems, 14*(4), 299–306.

Hirschi, T. (1969). Causes of Delinquency. University of California Press. Google-Books-ID: 53MNtMqy0fIC.

Hoc, J.-M. (2000). From human – machine interaction to human – machine cooperation. *Ergonomics, 43*(7), 833–843.

Horton, J. J. (2023). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?.

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2024). Large Language Models Cannot Self-Correct Reasoning Yet. arXiv:2310.01798 [cs]

Icove, D. J. (1986). Automated Crime Profiling. *FBI Law Enforcement Bulletin*, *55*(27).

Irving, G., & Askell, A. (2019). Ai safety needs social scientists. *Distill*.

Ishizaki, R., & Sugiyama, M. (2025). Large language models: assessment for singularity. *AI & Society*.

Jarrett, D., Pîslar, M., Bakker, M. A., Tessler, M. H., Köster, R., Balaguer, J., Elie, R., Summerfield, C., & Tacchetti, A. (2025). Language Agents as Digital Representatives in Collective Decision-Making. arXiv:2502.09369 [cs]

Johnson, J. (2020). Artificial Intelligence, Drone Swarming and Escalation Risks in Future Warfare. *The RUSI Journal*, *165*(2), 26–36. Publisher: Routledge _eprint: https://doi.org/10.1080/03071847.2020.1752026

Karnow, C. E. A. (1996). Liability for Distributed Artificial Intelligences. *Berkeley Technology Law Journal, 11*, 147.

Kasperkevic, J. (2015). Swiss police release robot that bought ecstasy online. *The Guardian*.

Kaufmann, T., Weng, P., Bengs, V., & Hüllermeier, E. (2024). A Survey of Reinforcement Learning from Human Feedback. arXiv:2312.14925 [cs]

Kaufmann, M., Egbert, S., & Leese, M. (2019). Predictive policing and the politics of patterns. *The British Journal of Criminology, 59*(3), 674–692.

Kim, H., Yi, X., Yao, J., Lian, J., Huang, M., Duan, S., Bak, J., & Xie, X. (2024). The Road to Artificial SuperIntelligence: A Comprehensive Survey of Superalignment. arXiv:2412.16468 [cs]

Kim, D. A., Hwong, A. R., Stafford, D., Hughes, D. A., O'Malley, A. J., Fowler, J. H., & Christakis, N. A. (2015). Social network targeting to maximise population behaviour change: a cluster randomised controlled trial. *The Lancet, 386*(9989), 145–153.

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*(1), 89–120.

Kitchin, R., & Dodge, M. (2019). The (in)security of smart cities: Vulnerabilities, risks, mitigation, and prevention. *Journal of Urban Technology, 26*(2), 47–65.

Kozlowski, A. C., & Evans, J. (2025). Simulating subjects: The promise and peril of artificial intelligence stand-ins for social agents and interactions. *Sociological Methods & Research*, *54*(3). First published online June 2, 2025.

Latour, B. (1996). On actor-network theory: A few clarifications. *Soziale Welt*, *47*(4), 369–381. Publisher: Nomos Verlagsgesellschaft mbH.

Latour, B. (2007). *Reassembling the Social: An Introduction to Actor-Network-Theory*. Clarendon Lectures in Management Studies. Oxford University Press, Oxford, New York.

Law, J., & Hassard, J. (1999). *Actor Network Theory and After*. Wiley-Blackwell, Oxford England: Malden, MA.

Lee, P. (2025). Synthetic Data and the Future of AI. *Cornell Law Review, 110*, 1.

Li, Y., Shen, X., Yao, X., Ding, X., Miao, Y., Krishnan, R., & Padman, R. (2025). Beyond Single-Turn: A Survey on Multi-Turn Interactions with Large Language Models. arXiv:2504.04717 [cs]

Liu, Z., Zhang, Y., Li, P., Liu, Y., & Yang, D. (2024). A Dynamic LLM-Powered Agent Network for Task-Oriented Agent Collaboration.

Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., Zhou, D., Le, Q. V., Zoph, B., Wei, J., & Roberts, A. (2023). The Flan Collection: Designing Data and Methods for Effective Instruction Tuning. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 22631–22648). PMLR. ISSN: 2640–3498.

Ma, Y., & Tresp, V. (2021). Causal Inference under Networked Interference and Intervention Policy Enhancement. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics* (pp. 3700–3708). PMLR. ISSN: 2640–3498.

Ma, C., Li, A., Du, Y., Dong, H., & Yang, Y. (2024). Efficient and scalable reinforcement learning for large-scale network control. *Nature Machine Intelligence*, *6*(9), 1006–1020. Publisher: Nature Publishing Group.

Malek, A., Ge, J., Lazic, N., Jin, C., György, A., & Szepesvári, C. (2025). Frontier LLMs Still Struggle with Simple Reasoning Tasks. arXiv:2507.07313 [cs].

Martínez-Miranda, E., McBurney, P., & Howard, M. J. W. (2016). Learning unfair trading: A market manipulation analysis from the reinforcement learning perspective. In *2016 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)* (pp. 103–109).

Matsueda, R. L. (1982). Testing Control Theory and Differential Association: A Causal Modeling Approach. *American Sociological Review*, *47*(4), 489–504. Publisher: [American Sociological Association, Sage Publications, Inc.].

Matsueda, R. L. (1988). The Current State of Differential Association Theory. *Crime & Delinquency*, *34*(3), 277–306. Publisher: SAGE Publications Inc.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183.

Mayorkas, A. N. (2024). *Roles and responsibilities framework for artificial intelligence in critical infrastructure*. Report, U.S: Department of Homeland Security. PDF available online.

Mitchell, M., Ghosh, A., Luccioni, A. S., & Pistilli, G. (2025). Fully Autonomous AI Agents Should Not be Developed. Version Number: 2.

Molina, M., & Garip, F. (2019). Machine Learning for Sociology. *Annual Review of Sociology*, *45*(Volume 45, 2019):27–45. Publisher: Annual Reviews.

Moritz, M., Topol, E., & Rajpurkar, P. (2025). Coordinated AI agents for advancing healthcare. *Nature Biomedical Engineering*, *9*(4), 432–438. Publisher: Nature Publishing Group.

Müller, V. C., & Bostrom, N. (2016). Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In V. C. Müller (Ed.), *Fundamental Issues of Artificial Intelligence* (pp. 555–572). Cham: Springer International Publishing.

OpenAI, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A. J., Welihinda, A., Hayes, A., Radford, A., Mądry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., Paino, A., Renzin, A., Passos, A. T., Kirillov, A., Christakis, A., Conneau, A., Kamali, A., Jabri, A., Moyer, A., Tam, A., Crookes, A., Tootoochian, A., Tootoonchian, A., Kumar, A., Vallone, A., Karpathy, A., Braunstein, A., Cann, A., Codispoti, A., Galu, A., Kondrich, A., Tulloch, A., Mishchenko, A., Baek, A., Jiang, A., Pelisse, A., Woodford, A., Gosalia, A., Dhar, A., Pantuliano, A., Nayak, A., Oliver, A., Zoph, B., Ghorbani, B., Leimberger, B., Rossen, B., Sokolowsky, B., Wang, B., Zweig, B., Hoover, B., Samic, B., McGrew, B., Spero, B., Giertler, B., Cheng, B., Lightcap, B., Walkin, B., Quinn, B., Guarraci, B., Hsu, B., Kellogg, B., Eastman, B., Lugaresi, C., Wainwright, C., Bassin, C., Hudson, C., Chu, C., Nelson, C., Li, C., Shern, C. J., Conger, C., Barette, C., Voss, C., Ding, C., Lu, C., Zhang, C., Beaumont, C., Hallacy, C., Koch, C., Gibson, C., Kim, C., Choi, C., McLeavey, C., Hesse, C., Fischer, C., Winter, C., Czarnecki, C., Jarvis, C., Wei, C., Koumouzelis, C., Sherburn, D., Kappler, D., Levin, D., Levy, D., Carr, D., Farhi, D., Mely, D., Robinson, D., Sasaki, D., Jin, D., Valladares, D., Tsipras, D., Li, D., Nguyen, D. P., Findlay, D., Oiwoh, E., Wong, E., Asdar, E., Proehl, E., Yang, E., Antonow, E., Kramer, E., Peterson, E., Sigler, E., Wallace, E., Brevdo, E., Mays, E., Khorasani, F., Such, F. P., Raso, F., Zhang, F., Lohmann, F. v., Sulit, F., Goh, G., Oden, G., Salmon, G., Starace, G., Brockman, G., Salman, H., Bao, H., Hu, H., Wong, H., Wang, H., Schmidt, H., Whitney, H., Jun, H., Kirchner, H., Pinto, H. P. d. O., Ren, H., Chang, H., Chung, H. W., Kivlichan, I., O'Connell, I., O'Connell, I., Osband, I., Silber, I., Sohl, I., Okuyucu, I., Lan, I., Kostrikov, I., Sutskever, I., Kanitscheider, I., Gulrajani, I., Coxon, J., Menick, J., Pachocki, J., Aung, J., Betker, J., Crooks, J., Lennon, J., Kiros, J., Leike, J., Park, J., Kwon, J., Phang, J., Teplitz, J., Wei, J., Wolfe, J., Chen, J., Harris, J., Varavva, J., Lee, J. G., Shieh, J., Lin, J., Yu, J., Weng, J., Tang, J., Yu, J., Jang, J., Candela, J. Q., Beutler, J., Landers, J., Parish, J., Heidecke, J., Schulman, J., Lachman, J., McKay, J., Uesato, J., Ward, J., Kim, J. W., Huizinga, J., Sitkin, J., Kraaijeveld, J., Gross, J., Kaplan, J., Snyder, J., Achiam, J., Jiao, J., Lee, J., Zhuang, J., Harriman, J., Fricke, K., Hayashi, K., Singhal, K., Shi, K., Karthik, K., Wood, K., Rimbach, K., Hsu, K., Nguyen, K., Gu-Lemberg, K., Button, K., Liu, K., Howe, K., Muthukumar, K., Luther, K., Ahmad, L., Kai, L., Itow, L., Workman, L., Pathak, L., Chen, L., Jing, L., Guy, L., Fedus, L., Zhou, L., Mamitsuka, L., Weng, L., McCallum, L., Held, L., Ouyang, L., Feuvrier, L., Zhang, L., Kondraciuk, L., Kaiser, L., Hewitt, L., Metz, L., Doshi, L., Aflak, M., Simens, M., Boyd, M., Thompson, M., Dukhan, M., Chen, M., Gray, M., Hudnall, M., Zhang, M., Aljubeh, M., Litwin, M., Zeng, M., Johnson, M., Shetty, M., Gupta, M., Shah, M., Yatbaz, M., Yang, M. J., Zhong, M., Glaese, M., Chen, M., Janner, M., Lampe, M., Petrov, M., Wu, M., Wang, M., Fradin, M., Pokrass, M., Castro, M., Castro, M. O. T. d., Pavlov, M., Brundage, M., Wang, M., Khan, M., Murati, M., Bavarian, M., Lin, M., Yesildal, M., Soto, N., Gimelshein, N., Cone, N., Staudacher, N., Summers, N., LaFontaine, N., Chowdhury, N., Ryder, N., Stathas, N., Turley, N., Tezak, N., Felix, N., Kudige, N., Keskar, N., Deutsch, N., Bundick, N., Puckett, N., Nachum, O., Okelola, O., Boiko, O., Murk, O., Jaffe, O., Watkins, O., Godement, O., Campbell-Moore, O., Chao, P., McMillan, P., Belov, P., Su, P., Bak, P., Bakkum, P., Deng, P., Dolan, P., Hoeschele, P., Welinder, P., Tillet, P., Pronin, P., Tillet, P., Dhariwal, P., Yuan, Q., Dias, R., Lim, R., Arora, R., Troll, R., Lin, R., Lopes, R. G., Puri, R., Miyara, R., Leike, R., Gaubert, R., Zamani, R., Wang, R., Donnelly, R., Honsby, R., Smith, R., Sahai, R., Ramchandani, R., Huet, R., Carmichael, R., Zellers, R., Chen, R., Chen, R., Nigmatullin, R., Cheu, R., Jain, S., Altman, S., Schoenholz, S., Toizer, S., Miserendino, S., Agarwal, S., Culver, S., Ethersmith, S., Gray, S., Grove, S., Metzger, S., Hermani, S., Jain, S., Zhao, S., Wu, S., Jomoto, S., Wu, S., Shuaiqi, Xia, Phene, S., Papay, S., Narayanan, S., Coffey, S., Lee, S., Hall, S., Balaji, S., Broda, T., Stramer, T., Xu, T., Gogineni, T., Christianson, T., Sanders, T., Patwardhan, T., Cunninghman, T., Degry, T., Dimson, T., Raoux, T., Shadwell, T., Zheng, T., Underwood, T., Markov, T., Sherbakov, T., Rubin, T., Stasi, T., Kaftan, T., Heywood, T., Peterson, T., Walters, T., Eloundou, T., Qi, V., Moeller, V., Monaco, V., Kuo, V., Fomenko, V., Chang, W., Zheng, W., Zhou, W., Manassra, W., Sheu, W., Zaremba, W., Patil, Y., Qian, Y., Kim, Y., Cheng, Y., Zhang, Y., He, Y., Zhang, Y., Jin, Y., Dai, Y., & Malkov, Y. (2024). GPT-4o System Card. arXiv:2410.21276 [cs]

Palantir (2025). Aip for defense. Report, Palantir Technologies. Accessed 2025.

Pan, M. Z., Cemri, M., Agrawal, L. A., Yang, S., Chopra, B., Tiwari, R., Keutzer, K., Parameswaran, A., Ramchandran, K., Klein, D., Gonzalez, J. E., Zaharia, M., & Stoica, I. (2025). Why Do Multiagent Systems Fail?.

Park, T. (2024). Enhancing Anomaly Detection in Financial Markets with an LLM-based Multi-Agent Framework. arXiv:2403.19735 [q-fin]

Piatti, G., Jin, Z., Kleiman-Weiner, M., Schölkopf, B., Sachan, M., & Mihalcea, R. (2024). Cooperate or Collapse: Emergence of Sustainable Cooperation in a Society of LLM Agents. arXiv:2404.16698 [cs]

Placani, A. (2024). Anthropomorphism in AI: hype and fallacy. *AI and Ethics, 4*(3), 691–698.

Popa, E. (2021). Human Goals Are Constitutive of Agency in Artificial Intelligence (AI). *Philosophy & Technology, 34*(4), 1731–1750.

Pratt, T. C., Cullen, F. T., Sellers, C. S., Thomas Winfree Jr., L., Madensen, T. D., Daigle, L. E., Fearn, N. E., & Gau, J. M. (2010). The Empirical Status of Social Learning Theory: A Meta-Analysis. *Justice Quarterly*, *27*(6), 765–802. Publisher: Routledge _eprint: https://doi.org/10.1080/07418820903379610

Price, W. N., II., Gerke, S., & Cohen, I. G. (2019). Potential Liability for Physicians Using Artificial Intelligence. *JAMA, 322*(18), 1765–1766.

Pridemore, W. A., Makel, M. C., & Plucker, J. A. (2018). Replication in Criminology and the Social Sciences. *Annual Review of Criminology*, *1*(Volume 1, 2018):19–38. Publisher: Annual Reviews.

Priluck, J. (2015). When Bots Collude. The New Yorker. Section: currency.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. S., Roberts, M. E., Shariff, A., Tenenbaum, J. B., & Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486. Number: 7753 Publisher: Nature Publishing Group.

Ratledge, E. C., & Jacoby, J. E. (1989). Handbook of Artificial Intelligence and Expert Systems in Law Enforcement. New York : Greenwood Press. Accepted: 2017–12-04T18:10:33Z Publication Title: https://delcat.worldcat.org/title/handbook-on-artificial-intelligence-and-expert-systems-in-law-enforcement/oclc/19554021

Relins, S., Birks, D., & Lloyd, C. (2025). Using Instruction-Tuned Large Language Models to Identify Indicators of Vulnerability in Police Incident Narratives. *Journal of Quantitative Criminology*.

Rivera, J.-P., Mukobi, G., Reuel, A., Lamparth, M., Smith, C., & Schneider, J. (2024). Escalation Risks from Language Models in Military and Diplomatic Decision-Making. (pp. 836–898). arXiv:2401.03408 [cs]

Robert, D., & Dufresne, M. (Eds.). (2016). *Actor-network theory and crime studies: explorations in science and technology*. London New York: Routledge.

Sandholm, T. (2007). Perspectives on multiagent learning. *Artificial Intelligence, 171*(7), 382–391.

Santoni de Sio, F., & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology, 34*(4), 1057–1084.

Schäfer, B., Witthaut, D., Timme, M., & Latora, V. (2018). Dynamically induced cascading failures in power grids. *Nature Communications*, *9*(1):1975. Number: 1 Publisher: Nature Publishing Group.

Seaver, N. (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, *4*(2), 205395171773810. Publisher: SAGE Publications.

Shapira, N., Wendler, C., Yen, A., Sarti, G., Pal, K., Floody, O., Belfki, A., Loftus, A., Jannali, A. R., Prakash, N., et al. (2026). Agents of chaos. arXiv preprint arXiv:2602.20021

Shaw, C. R., & McKay, H. D. (1942). Juvenile delinquency and urban areas. Juvenile delinquency and urban areas. University of Chicago Press, Chicago, IL, US. Pages: xxxii, 451.

Shen, T., Jin, R., Huang, Y., Liu, C., Dong, W., Guo, Z., Wu, X., Liu, Y., & Xiong, D. (2023). Large Language Model Alignment: A Survey. arXiv:2309.15025 [cs]

Shen, T., Zhu, D., Zhao, Z., Li, Z., Wu, C., & Wu, F. (2025). Will LLMs Scaling Hit the Wall? Breaking Barriers via Distributed Resources on Massive Edge Devices. arXiv:2503.08223 [cs]

Shoham, Y., Powers, R., & Grenager, T. (2007). If multi-agent learning is the answer, what is the question? *Artificial Intelligence, 171*(7), 365–377.

Shojaee, P., Mirzadeh, I., Alizadeh, K., Horton, M., Bengio, S., & Farajtabar, M. (2025). The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. arXiv:2506.06941 [cs]

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, *631*(8022), 755–759. Publisher: Nature Publishing Group.

Simons, R. L., & Burt, C. H. (2011). Learning to Be Bad: Adverse Social Conditions, Social Schemas, and Crime. *Criminology*, *49*(2), 553–598. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1745-9125.2011.00231.x

Simpson, S. S. (2025). Criminology and corporate crime: The art of scientific cross-pollination. *Annual Review of Criminology, 8*, 311–331.

Solum, L. B. (1992). Legal Personhood for Artificial Intelligences. *North Carolina Law Review, 70*, 1231.

Sussman, D. L., & Airoldi, E. M. (2017). Elements of estimation theory for causal effects in the presence of network interference. arXiv:1702.03578 [stat]
Sutherland, E. H. (1939). Principles of criminology. J. B. Lippincott company.
Swanepoel, D., & Corks, D. (2024). Artificial Intelligence and Agency: Tie-breaking in AI Decision-Making. *Science and Engineering Ethics, 30*(2), 11.
Tan, M. (1993). Multi-agent reinforcement learning: independent versus cooperative agents. In *Proceedings of the Tenth International Conference on International Conference on Machine Learning, ICML'93* (pp. 330–337). San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
Terry, M., Kulkarni, C., Wattenberg, M., Dixon, L., & Morris, M. R. (2023). Interactive ai alignment: Specification, process, and evaluation. arXiv preprint arXiv:2311.00710
Topalli, V., & Nikolovska, M. (2020). The Future of Crime: How Crime Exponentiation Will Change Our Field. *The Criminologist*.
Tranchero, M., Brenninkmeijer, C.-F., Murugan, A., & Nagaraj, A. (2024). Theorizing with Large Language Models.
Tsvetkova, M., Yasseri, T., Pescetelli, N., & Werner, T. (2024). A new sociology of humans and machines. *Nature Human Behaviour*, *8*(10), 1864–1876. Publisher: Nature Publishing Group.
Turner, J. (2018). Robot Rules: Regulating Artificial Intelligence. Springer.
van der Wagen, W., & Pieters, W. (2015). From Cybercrime to Cyborg Crime: Botnets as Hybrid Criminal Actor-Networks. *The British Journal of Criminology, 55*(3), 578–595.
VanderWeele, T. J., & An, W. (2013). Social Networks and Causal Inference. In S. L. Morgan (Ed.), *Handbook of Causal Analysis for Social Research, Handbooks of Sociology and Social Research* (pp. 353–374). Netherlands, Dordrecht: Springer.
Vezhnevets, A. S., Agapiou, J. P., Aharon, A., Ziv, R., Matyas, J., Duéñez-Guzmán, E. A., Cunningham, W. A., Osindero, S., Karmon, D., & Leibo, J. Z. (2023). Generative agent-based modeling with actions grounded in physical, social, or digital space using Concordia. arXiv:2312.03664 [cs]
Wall, D. S. (2024). *Cybercrime: The Transformation of Crime in the Information Age*. John Wiley & Sons.
Wallach, W., & Allen, C. (2009). Moral Machines: Teaching Robots Right from Wrong. Oxford University Press.
Warr, M., & Stafford, M. (1991). The Influence of Delinquent Peers: What They Think or What They Do? *Criminology*, *29*(4), 851–866. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1745-9125.1991.tb01090.x
Watson, D. (2019). The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence. *Minds and Machines, 29*(3), 417–440.
Whitney, C. D., & Norman, J. (2024). Real Risks of Fake Data: Synthetic Data, Diversity-Washing and Consent Circumvention. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24* (pp. 1733–1744). New York, NY, USA. Association for Computing Machinery.
Williams, S., & Huckle, J. (2024). Easy Problems That LLMs Get Wrong. arXiv:2405.19616 [cs]
Wooldridge, M., & Jennings, N. R. (1995). Agent theories, architectures, and languages: A survey. In M. J. Wooldridge, & N. R. Jennings (Eds)., *Intelligent Agents* (pp. 1–39). Berlin, Heidelberg. Springer.
Woolgar, S. (1985). Why not a Sociology of Machines? The Case of Sociology and Artificial Intelligence. *Sociology*, *19*(4), 557–572. Publisher: SAGE Publications Ltd.
Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., & Wang, C. (2023). AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. arXiv:2308.08155 [cs]
Xie, C., Chen, C., Jia, F., Ye, Z., Lai, S., Shu, K., Gu, J., Bibi, A., Hu, Z., Jurgens, D., Evans, J., Torr, P. H. S., Ghanem, B., & Li, G. (2024). Can large language model agents simulate human trust behavior? In *Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS)* (pp. 15674–15729).
Xie, Y., Zhu, C., Zhang, X., Wang, M., Liu, C., Zhu, M., & Zhu, T. (2025). Who's the Mole? Modeling and Detecting Intention-Hiding Malicious Agents in LLM-Based Multi-Agent Systems. arXiv:2507.04724 [cs].
Xu, Z., Jain, S., & Kankanhalli, M. (2025). Hallucination is Inevitable: An Innate Limitation of Large Language Models. arXiv:2401.11817 [cs]
Yang, Y., Nishikawa, T., & Motter, A. E. (2017). Small vulnerable sets determine large network cascades in power grids. *Science*, *358*(6365), eaan3184. Publisher: American Association for the Advancement of Science.

Zhao, J., Li, D., Sanhedrai, H., Cohen, R., & Havlin, S. (2016). Spatio-temporal propagation of cascading overload failures in spatially embedded networks. *Nature Communications*, *7*(1), 10094. Number: 1 Publisher: Nature Publishing Group.

Authors and Affiliations

Gian Maria Campedelli[1,2]

✉ Gian Maria Campedelli
gianmaria.campedelli@unitn.it

[1] Department of Sociology and Social Research, University of Trento, Via Verdi, 26, Trento, TN 38122, Italy

[2] MobS Lab, Center for Augmented Intelligence, Fondazione Bruno Kessler, Via Sommarive, 18, Trento, TN 38123, Italy