Philosophy & Technology (2025) 38:27
https://doi.org/10.1007/s13347-025-00856-x

RESEARCH ARTICLE

"All AIs are Psychopaths"? The Scope and Impact of a Popular Analogy

Elina Nerantzi[1]

Received: 3 September 2024 / Accepted: 5 February 2025

Abstract
Artificial Intelligence (AI) Agents are often compared to psychopaths in popular news articles. The headlines are 'eye-catching', but the questions of what this analogy means or why it matters are hardly answered. The aim of this paper is to take this popular analogy 'seriously'. By that, I mean two things. First, I aim to explore the scope of this analogy, i.e. to identify and analyse the shared properties of AI agents and psychopaths, namely, their lack of moral emotions and their capacity for instrumental rationality. Second, I aim to examine the impact of the analogy. I argue that both agents, as 'amoral calculators', present the perfect candidates to revisit two long-standing debates on moral and criminal responsibility, regarding the necessity of moral emotions for 'moral-agent-capacity responsibility' and the necessity of 'moral-agent-capacity responsibility' for criminal responsibility. Finally, cross-examining the debates on the moral and criminal responsibility of psychopaths and AI agents is instructive and revealing. Instructive since the moral and legal treatment of psychopaths can be telling about the future treatment of AI agents (and vice versa) and revealing since it makes explicit our often-implicit philosophical commitments on the criteria of moral agency and the overarching purpose of criminal law.

Keywords Artificial intelligence · Moral agency · Criminal responsibility · Emotions · Psychopaths

1 Introduction

Jack Henry Abbott, a diagnosed psychopath with a long history of criminal convictions, offered a disturbing account of his inner world in his book 'In the Belly of the Beast: Letters from Prison'. 'There are emotions', he writes, '- a whole spectrum of

✉ Elina Nerantzi
Elina.NERANTZI@eui.eu

[1] European University Institute (EUI), Law Department, Max Planck Institute for the Study of Crime, Security and Law, "Algorithmic Profiling and Automated Decision-Making in Criminal Justice" Fellow Group, Via Bolognese 156, 50139 Florence, Italy

Published online: 25 February 2025

them – that I know only through words, reading, and in my immature imagination. I can imagine I feel these emotions (know, therefore, what they are), but I do not' (Abbott, 1981:13).

What Jack Abbott is saying is that he can parrot emotional (and moral) language without genuine moral understanding and without experiencing any emotions. Abbott's emotionally flat landscape resembles the way in which some Artificial Intelligence (AI) agents are occasionally said to be 'emotional'. Not because they 'feel' emotions or are motivated by them, but because they can mimic emotions or recognise them in humans. Even theories showing that emotions may be formalized for computational purposes admit that the point of doing so is not to endow machines themselves with emotions. Rather, the point is 'to create a computer model that can "understand" what emotions people would be likely to experience under what conditions. Such a system should be able to predict and explain human emotions, not "have" them' (Ortony et al., 2022:27).[1]

These prima facie similarities between psychopaths and AIs as 'emotionless' and/or 'amoral' have not gone unnoticed in the press. They make for eye-catching headlines, such as 'All AIs are psychopaths' (Nixon, 2020), 'What can psychopaths teach us about AI?' (Greene, 2019), 'Are you scared yet? Meet Norman, the world's first psychopath AI' (Wakefield, 2018), 'Artificial Psychopathy' (Diana, 2023), to name a few. The analogy creates an eerie feeling. It builds upon deep-rooted human fears of the 'heartless psychopath' and the 'evil AI', combining them in an archetype of a 'psychopathic AI', which is emotionless and amoral, yet entirely self-interested and, thus, extremely dangerous.

The aim of this paper is to take this popular analogy 'seriously'. By that, I mean two things. First (Section 2), I aim to facilitate a more reflective use of the analogy and to explore its scope beyond its sensationalism. 'Which properties do AIs (and what 'kind' of AIs) have in common with psychopaths (and with which 'types' of psychopaths)? At the very least, AIs – contrary to psychopaths – are not human, not conscious and not sentient. Properly understood, thus, the analogy between them is not a literal one as AIs and psychopaths are not *actually* similar. However, they do seem to share the common properties of being 'amoral' yet 'rational' agents. These properties must be substantiated if the analogy is to hold any water apart from captivating the popular imagination and feeding our fears. I argue that both (subclinical) psychopaths and AIs fit the 'amoral yet rational' paradigm in the sense that they both lack *moral* emotions, i.e. the type of emotions and emotion-related faculties, such as empathy, that have been empirically shown to be crucial (even necessary) for moral judgments and moral motivation (Section 2.1) while their instrumental rationality – reminiscing that of the *homo economicus*—remains intact (Section 2.2).

After examining the analogy's scope, the second aim of the paper is to explore its impact (Section 3). I argue that both psychopaths and AIs—as agents who exhibit

[1] The authors argue that this enterprise would be 'pointless' (p 27) contrary to Adolphs and Anderson (2018) who argue that building a robot with emotions is 'possible in principle, even, though, it is currently a science fiction' (p 25). In this paper, I focus not on the technical feasibility of building 'robots with emotions' but on the ethical claim that only such a robot would be a 'full' moral agent, responsible for harm it causes.

cognitive capacities and instrumental rationality while lacking in the moral-emotional realm (under the qualifications set forward in Section 2)—present the perfect candidates to revisit two long-standing questions: one on moral responsibility (*is moral competence – the ability to form genuine moral judgments and be morally motivated by them—a necessary condition for 'moral-agent capacity responsibility' and, if yes, is moral competence dependent on our access to moral emotions?*) and one on criminal responsibility (*should criminal responsibility track moral-agent capacity responsibility, so understood?*).

The paper does not aim to settle these questions; that would be an overtly ambitious goal. Rather, the aim is to cross-examine the long-standing debate on the moral and criminal responsibility of psychopaths (Section 3.1.1) with the emergent debate on the moral and criminal responsibility of AI (Section 3.1.2) in order to discuss (Section 3.2) whether and to what extent, the non-literal similarity comparison between psychopaths and AI agents, can be instructive, i.e. have the 'predictive or explanatory effect' that analogies have (Carpenter, 2016). Namely, I aim to explore what the debate on the (future) moral and legal responsibility of AI could teach us about our current treatment of psychopaths (backwards-looking approach) and what the current treatment of psychopaths could teach us about the future treatment of AI (forwards-looking approach).

Finally, in the last section of the paper (Section 4), conclusions will be drawn. I aim to show that the analogy can help both debates make progress by revealing potential inconsistencies in the treatment of 'amoral' psychopaths and AIs as well as our often implicit, philosophical commitments regarding the relevance of moral emotions in the criteria of moral-agent-capacity responsibility and the overarching purpose of criminal law.

2 The Scope of the Analogy

Analogies are deeply rooted in human psychology, and they have been proven to be the most deployed tool in our effort to understand, frame and shape the breadth of AI technology (Maas, 2023). There is no shortage of 'AI is like…' statements; from animals to children, to aliens, tools or companies, AI has been said to be like many different things. These analogies illuminate different features of AI technology, and they may work best in different contexts without being necessarily judged as 'right' or 'wrong' (Maas, 2023).

Which features of AI technology does the 'AI is like a psychopath' analogy illuminate? This is an important question given the litany of differences the two of them have.[2] To begin with, AI agents are, like the name suggests, artificial; they are human creations and not humans themselves, contrary to psychopaths who, despite exhibiting some of the darkest traits of humankind, are certainly human. Related to that, there is not sufficient evidence, at the moment, to suggest that AI has consciousness (understood as self-awareness and the capacity for subjective experience)

[2] My thanks to the anonymous reviewers for pushing me on this point.

or sentience (understood as the capacity to feel pleasure and pain). Contrary to that, psychopaths, despite their emotional and moral abnormalities (see 2.1.2), are, of course, not denied their status as conscious human beings. Finally, AI is not autonomous in the robust philosophical sense in which autonomy is understood when applied to humans (psychopathic or not), e.g. as self-governance. In the minimal sense autonomy is understood in AI and robotics, it suffices for an AI agent to be autonomous if it can select actions to achieve pre-defined goals, even if the agent lacks self-governance, i.e. even if the agent cannot autonomously choose or reassess these goals.

The similarities start when one enters the realm of emotions and moral behaviour, as both psychopaths and AIs are invariably said to be 'emotionless' and 'amoral'. To be precise, what is meant is that they both lack *moral* emotions (such as remorse, guilt, regret or compassion) and empathy, which are crucial for morality. Before proceeding with substantiating this claim, an important caveat is necessary.

Indeed, both psychopaths and AIs lack moral emotions, but AI's emotional blank state is due to their most fundamental lack of sentience, which is 'the bedrock of moral emotions' (Stewart, 2024; Véliz, 2021). That is because emotions, in general, and moral emotions, specifically, cannot be treated in isolation from consciousness and sentience. To be able to truly feel an emotion – let alone be motivated by it—one must, at least, be in a conscious state that would allow for the subjective experience and awareness of one's feelings. To understand what it means to harm someone, one must have experiential knowledge of pain.

AIs, thus, as non (provably) sentient human artefacts, are a priori excluded from the possession of moral emotions, whereas psychopaths, as sentient humans, can, in principle, experience moral emotions, but they, pathologically, do not. Differently put, the 'lack of moral emotions' similarity has different causes. AI's lack of moral emotions is an aftermath of their non-sentient status (Veliz, 2021; Gibert & Martin, 2022; Stewart, 2024), and psychopaths' lack of moral emotions is due to their empirically proven pathological emotional deficiencies (see 2.1.2). However, this difference does not rob the analogy of its instructive value to help us reconsider the criteria needed for moral and criminal responsibility. Another reason to consider the analogy is the evidence showing that the emotional deficiencies of psychopaths may also be caused by their sentient deficiencies. An empirical study (Brislin et al., 2016) has shown that psychopaths have a higher tolerance to pain, and this insensitivity to pain has been correlated with their callous indifference to the pain of others.

Whatever the cause might be, their shared property remains; psychopaths and AIs are both instances of an archetype of a rationally self-interested yet emotionless and, therefore, amoral agent who can cause harm.

Which kind of AI and which type of a psychopath, though? A final caveat is needed: the notoriously difficult-to-define concepts of AI and psychopathy will be fleshed out while describing their common features. As a brief signpost, though, by 'AI', I follow, in general terms, what Russell and Norvig describe as the 'standard approach' to building AI agents. That is the 'rational-agent' approach to AI (Russell & Norvig, 2021), where the effort is to build entities that are 'intelligent' not in the sense that they mimic human thinking and action but in the sense that they have the rational ability to optimally achieve the goals that have been assigned to

them. The basic design idea is that of building agents that would act rationally on our behalf, i.e. agents that act so as to achieve the best outcome or, when there is uncertainty, the best-expected outcome. This is the 'common theme that unites the various strands of contemporary AI' (Wooldridge, 2021:154), and this is the design, I argue, that corresponds with the concept of a 'subclinical', 'primary', 'corporate', 'partial', or 'successful' psychopath (terminology differs). That is an individual who exhibits core psychopathic traits (e.g. lack of empathy or remorse, manipulativeness) but not all of them (e.g. erratic behaviour, impulsivity, antisocial tendencies). On the contrary, a subclinical psychopath is rational and strategic enough to achieve her self-interested goals unaffected by a moral compass. Hereinafter, when I refer to psychopaths, this is the intended type, as described in more detail in Section 2.1.3.

In what follows, I will examine, in turn, the common AI-psychopathic properties of 'lack of moral emotions' (Section 2.1) and of 'instrumental rationality' (2.2).

2.1 Lack of Moral Emotions

2.1.1 The Importance of Moral Emotions

An absolute definition of emotion (as an umbrella term to include feelings, mood and affect[3]) is notoriously difficult to find (Maroney, 2006:123; Goldie, 2007). However, there seems to be a consensus in modern psychology and neuroscience that emotions as phenomena have components, namely, evaluative, physiological, phenomenological, behavioural, expressive and mental components (Scarantino & Sousa, 2021; Coppola, 2021: 94). For instance, if I see a bear while hiking, I will probably experience the emotion of fear, which means that: a) I will appraise the bear as dangerous (evaluative component), b) my heart rate will increase (physiological component), c) I will experience an unpleasant feeling (phenomenological component), d) my facial expression might change, e.g. my jaw might drop open (expressive component), e) I will have the tendency to flee (the behavioural component) and f) I will focus my attention to avoid the danger (mental component). These phenomena are short-lived and intentional ('about something', e.g. the bear), and they 'work together' to create what we experience as the emotion of fear.

Most importantly, these components include both feelings and cognitive evaluations, illustrating the dominant approach in psychology and neuroscience that there is not a strict dividing line between cognition and emotion, 'thinking' and 'feeling'.[4] How we think about something is influenced by how we feel about it and vice versa.

[3] In broad strokes (as the terms are both related and distinct) the following distinctions are drawn: i) emotions vs. feelings: feelings are only one of the components of emotions, ii) emotions vs. moods: emotions are more short-lived than moods, they are more intense and 'intentional' ('about something'), eg. I experience a sad emotion after watching a sad film, but I might be in a 'sad mood' without an obvious reason and iii) emotions vs affect: affect is a broader term that encompasses both emotions and moods.

[4] This was not always the case. Early theories of emotions viewed them as essentially physiological phenomena, as automatic, reflexive and uncontrolled physiological responses to environmental stimuli ('the feelings tradition' also known as the 'James-Lange theory'). For an overview see Scarantino, de Sousa 2021.

Emotions have proven to be necessary for decision-making and adaptive behaviour (Damasio, 1994). Returning to the previous example, it is adaptive to feel fear upon seeing a bear; it helps us focus our attention exclusively on the features of the situation that are relevant to escaping it (e.g. finding a place to hide) without wasting time deciding what irrelevant factors to ignore (e.g. the beautiful trees in the background).

Consider also the well-cited case of Elliott, one of Antonio Damasio's more famous patients. In neuroscientific research, the case aimed to show that when the regions of the brain that are involved in emotional processes are impaired (usually in the form of frontal lobe injuries), so is decision-making. After a brain tumour wounded the frontal lobe tissue in Elliott's brain and he had to undergo surgery, Elliott, a former businessman and functional member of society, was not only left emotionally numb but also helplessly paralysed when faced with trivial decisions (e.g. choosing a doctor's appointment), irrationally risk-prone and impatient upon deferred gratification. 'I began to think that the cold-bloodedness of Elliot's reasoning prevented him from assigning different values to different options,' Damasio writes, 'and made his decision-making landscape hopelessly flat' (Damasio, 1994: 51).

Emotions, thus, are not enemies to reason and a-rational forces but important tools that drive our decision-making. The paper builds on this view and expands not on the various uses of emotions but on the *moral* uses of *moral* emotions, as this is the kind of emotions that both psychopaths and AIs lack.

Moral emotions are a specific category of emotions which, in a nutshell, help individuals 'do the (morally) right thing'. They are intrinsically linked to the well-being and interests of others or of society as a whole, and they are crucially (for some, even *necessarily*) involved in moral judgments and moral motivation. Contrary to basic emotions, they are typically developed after childhood and are influenced by culture (Coppola, 2021). The multifaceted nature of moral emotions has led theorists to develop different classifications, with the broad categories being those of (Coppola, 2021; Teper et al., 2015): a) self-conscious emotions, such as guilt, shame or regret, b) positive/other-praising moral emotions, such as love and elevation, c) other-condemning emotions, such as anger and disgust and d) other-suffering emotions, such as compassion.

Empathy is central here. Lay conversations and those reproduced in the news articles on the analogy between AI and psychopaths identify moral emotions with 'empathy' and take empathy to be an emotion through which we feel what others feel. Empathy, though, is not strictly speaking an emotion and is not a unitary process. It is predominantly understood as a multi-dimensional process which consists of three main components: a) a cognitive component by which the individual can represent the internal state of mind of another individual, b) an affective component which, at its most basic, involves sharing another person's affective mental state and c) a motivational component, including (apart from sharing another's person affective state) some motivation to help or respond appropriately (Stueber, 2019; Zaki & Ochsner, 2012).

The emphasis on moral emotions is a practical one. Over the past few decades, research has suggested that moral emotions and empathy are crucial drivers of moral

behaviour (R Tangney et al., 2007; Teper et al., 2015). This research suggests that human beings do not behave prosocially only because of their cognitive skills and rational reasoning about principles of justice and normative standards of right and wrong. On the contrary, acting morally is also deeply rooted in moral emotions; we help others because (amongst other things) we empathise with them and feel compassion, we refrain from harming them to avoid the self-conscious emotion of regret or guilt, etc.

Specifically, neuroscientific evidence suggests that moral emotions inform our moral judgments (Blair & Fowler, 2008; Helion & Ochsner, 2018) and are integral in influencing moral behaviour in real life (Teper et al., 2015). More evidence for the relevance of moral emotions to moral behaviour comes from research on psychopaths, as shown below.

2.1.2 Amoral Psychopaths

In the psychological literature, the term 'psychopaths' appears to encompass a highly heterogeneous group of persons with deviant personality traits and behavioural inclinations (Jurjako & Malatesti, 2023; De Brito et al., 2021). However, their flattened affect and lack of empathy are traditionally treated as central diagnostic symptoms[5] (Cleckley, 1988; Hare, 1993) going as back as 1835 to the English psychiatrist J.C. Prichard's early definition of psychopathy as 'moral insanity' (Ward, 2010).

Indeed, the typical philosophical understanding of the psychopath is that of an individual who remains an 'outsider to our moral practices' (Duff, 1977; Litton, 2010). They do not genuinely understand the nature and the quality of their harmful acts because they do not understand the moral and emotional significance of these acts as the rest of us do. The psychopath is not a rebel, i.e. someone who can adequately understand society's shared values but chooses to reject them for her own reasons; on the contrary, she is someone who was never able to appreciate what matters to people in the first place.

Empirical studies seem to support the moral deficits of psychopaths. For instance, studies building on developmental psychology theories (Blair et al., 1995, 1997) show that psychopaths have severe deficits in moral understanding as they fail to distinguish between moral and conventional violations. Distinguishing between moral violations that affect the welfare and rights of other persons (e.g. harming humans) and conventional violations that only violate the social order (e.g. talking in class) is an essential step in human moral development. However, for psychopaths, all

[5] According to the current 'gold standard' psychometric test for the assessment of psychopathy in adults, i.e. Dr. Robert Hare's Revised (20-item) Psychopathy checklist (PCL-R) the psychopathic traits are: *1. Glibness/superficial charm 2. Grandiose sense of self-worth 3. Need for stimulation/proneness to boredom 4. Pathological lying 5. Conning/manipulative 6. Lack of remorse or guilt 7. Shallow affect 8. Callous/lack of empathy 9. Parasitic lifestyle 10. Poor behavioral controls 11. Promiscuous sexual behavior 12. Early behavior problems 13. Lack of realistic, long-term plans 14. Impulsivity 15. Irresponsibility 16. Failure to accept responsibility for own actions 17. Many short-term marital relationships 18. Juvenile delinquency 19. Revocation of conditional release 20. Criminal versatility'.*

transgressions are conventional; they fail to grasp the higher moral significance of moral violations. For instance, when asked why it is wrong to hit a child in order to make her cry, some psychopaths replied, 'It's not socially acceptable' (Blair et al., 1995: 212). Psychopaths, thus, do retain a cognitive understanding of legal and moral wrongs, but they treat the word 'wrong' in a conventional way. Hitting a child is treated by them as a matter of social etiquette.

Importantly, there is evidence attributing psychopaths' moral deficits to their emotional deficits. As said before, research on psychopathy has become one of the main pieces of evidence to suggest the necessity of moral emotions and empathy for moral behaviour. For instance, higher levels of psychopathic traits are often strongly associated with deficits in affective empathy (Campos et al., 2022; Blair, 2007) but not in cognitive empathy (Blair et al., 1995; Richell et al., 2003) since psychopaths are known to be manipulative and conning and that skill depends on sensitivity on social cues. In other words, psychopaths cannot genuinely experience empathy, but they could 'fake' it to achieve their goals (Robinson & Rogers, 2015). Interestingly, according to a recent study (Burghart et al., 2024), psychopaths' inability to experience empathy could be attributed to their suffering of alexithymia (emotional blindness); that is, it could be attributed to their limited understanding of their own emotions.

Finally, neuroscientific research corroborates the link between the emotional and moral deficits of psychopaths ('neuromoral theory', Raine, 2019; Raine & Yang, 2009; Darby et al., 2018). It suggests that impairments to one or more areas of the 'neuromoral circuit' (i.e. a network of brain regions, primarily centered in the ventromedial prefrontal cortex (vmPFC), responsible for generating moral emotions and guiding moral judgments) results in impairment to feeling, thinking, and behaving in a moral way which in turn lays the foundation for psychopathic behaviour.

The empirical link between moral emotions and moral competence, i.e. the capacity to form moral judgments and be morally motivated to follow them, has a series of implications. In moral philosophy – and beyond the scope of this paper—it could be an argument in favour of sentimentalism and motivational internalism (e.g. Prinz, 2006). In legal theory, it could be an argument against the 'moral-agent-capacity responsibility' of psychopaths and, consequently, of their criminal responsibility – or, at least, the acceptance of their diminished criminal responsibility. This will be discussed in Sect. 3.1.1. Now, it is time to examine the 'lack of moral emotions' in the other subject of the analogy, i.e. the AI agents.

2.1.3 Amoral AI Agents

The claim that AI agents lack moral emotions is not a controversial one. That is because, as already argued, this lack is treated as a symptom of broader absent properties in AI agents, such as consciousness or sentience. Whereas psychopaths are deemed to have impairments in their 'neuromoral circuit' that 'block' their access to moral behaviour, the AI agents of the 'standard approach' to AI (Russell & Norvig, 2021) do not have the necessary consciousness, sentience and biology to experience moral emotions in the first place.

Additionally, this 'state of the art' does not seem likely to change any time soon. First, in the context of the 'rational-agent approach' to AI, '*no one in the AI field is working on making AI Systems conscious*' (Russell, 2019:16). The goal is competence, not consciousness; it is to endow AI agents with instrumental rationality that would allow them to achieve goals on our behalf and not with affective states would be the driving forces behind their behaviour. Second, in order to engineer emotions into AI agents, one needs to know how emotions are instantiated in the brains of animals and people. The processing mechanisms required to generate emotions, though, are yet to be articulated (Adolphs & Anderson, 2018).

Genuinely experiencing moral emotions, thus, seems to be, in principle, prohibitive for AI agents. However, this does not mean they cannot *simulate* emotions and empathy, as psychopaths do. I suspect that psychopaths and AIs are entangled and feared in our imagination not only because they lack moral emotions but – on a deeper level—because they can 'fake' them.

There are some AI agents that seem to be quite good at replicating 'cognitive empathy'. Affective computing (Picard, 1997)—also known by the commercial moniker 'emotional AI' – is a subset of AI able to infer human emotions from physical (e.g. speech, facial expressions) and psysiological signals (e.g. electrocardiogram, body temperature, eye tracking) and respond appropriately in a way that mimics affective empathy.[6] For instance, a chatbot in a customer service context can recognise and understand from verbal cues when someone is angry and respond 'empathetically'. However, this is not a genuinely empathic response. Without the capacity to experience the emotion of anger themselves, AI agents cannot vicariously share the anger of others, as genuine affective empathy requires (Stueber, 2019). They can only 'form analogically inferred presumptions about the psychological states of others' (Montemayor et al., 2022). This is not the same as affectively empathizing with others – at least in a non-psychopathic way.

This shared incapacity for genuine empathic behaviour increases the risks that psychopaths and AI agents will engage in manipulative, unethical behaviour (Montemayor et al., 2022). That is even more so as long as their instrumental rationality remains intact, i.e., as long as they can come up with optimal ways to achieve their goals without the distractions of their non-existent moral emotions.

2.2 Instrumental Rationality

One is instrumentally rational when one can adopt suitable means to achieve one's ends. I argue that AI agents are commonly compared to psychopaths (and vice versa) not only because both are 'amoral' in the sense described above but, most specifically, because both are amoral *calculators*. They cognitively 'know what they are doing', and they can reason whether what 'they are doing' is the optimal way

[6] For criticism that 'emotional AI' is unethical see, indicatively, Gossett 2023; Tretter, 2024; Devillers and Cowie 2023. The argument in this paper could contribute to this literature by describing the said AI emotional capacities as psychopathic.

to achieve their goal. In other words, their cognitive and rational capacities remain intact.

Psychopaths 'know what they are doing' in the sense that they are not delusional. Cognitive impairment is not a diagnostic criterion of psychopathy, and this is why psychopathy differs greatly from the typical cases of legal insanity, which include cases of persons suffering from disorderly thinking, such as hallucinations or psychosis (Duff, 2010). Similarly, one of the critical components of the AI agents built based on the 'rational-agent approach to AI' is their ability to 'perceive' their environment through their sensors (i.e. to process sensory data like visual or auditory inputs) in order to act upon that environment through actuators (Russell and Norvig: 4). Thus, in different ways and to different extents, psychopaths and AI agents are 'in touch' with their environment, which is a necessary first step towards their instrumental rationality, but it is not sufficient.

What needs to be emphasized is that they both are rational calculators of their own interest. A conceptually insightful way to understand this similarity is to bear in mind that both psychopaths and AI agents have been analogized with the archetypical self-interested agent, i.e. the *homo economicus*, the mythical, perfectly rational agent of neoclassical economics.

Without being able to go into detail in the scope of this paper, in the last decade, it has often been argued that *homo economicus* seems to incorporate dark character traits, which resemble signs of psychopathy (Bailey, 2017; Boddy, 2023; Fuchs & Lingnau, 2024; Stout, 2014). Specifically, *homo economicus*—as an emotionally shallow, manipulative, opportunistic and entirely self-interested agent – could be psychologically described as a 'subclinical' case of a psychopath.[7] Subclinical psychopaths possess some but not all the diagnostic traits of clinical psychopathy. Namely, they exhibit moral-emotional detachment, callous affect and interpersonal manipulation, but they do not exhibit overtly antisocial or unsteady behaviour, impulsive thoughtlessness or outbursts of violence; they do not lead an erratic lifestyle. On the contrary, subclinical psychopaths lead a relatively normal life, are capable of strategic thinking and often succeed in white-collar jobs, especially in finance (Babiak, 2016; Babiak & Hare, 2019; Babiak et al., 2010). As R.A. Hare has put it, "Not all psychopaths are in prison. Some are in the Boardroom" (as cited in Smith & Lilienfeld, 2013:204).

The subclinical psychopaths are the ones that have been repetitively analogized with the *homo economicus*. Furthermore, they are the ones resembling the AI agents built based on the 'rational-agent approach' to AI – which, in turn, have also been analogized with the *homo economicus* (Parkes & Wellman, 2015).

In more detail, a parallel is drawn between rational, goal-based, utility-maximising AI agents (Russell & Norvig, 2021) and the theoretical concept of the *homo*

[7] The PCL-R psychopathy test (fn 5) shows two major, overarching factors of psychopathy, namely, 1) social-emotional detachment (Factor 1) and (2) antisocial behavior (Factor 2) which can be further differentiated into subfactors (Hare & Neumann, 2005, 2008). In general, clinical psychopathy refers to individuals with a substantially elevated Factor 1 and Factor 2. Subclinical psychopaths show an equally profoundly elevated Factor 1 but only a mildly elevated Factor 2 (Babiak, 2016). Psychopaths analogized to AI agents and to the *homo economicus* are the subclinical/Factor 1 psychopaths.

economicus representing a person with flawless rationality pursuing her own narrow self-interest (Nerantzi & Sartor, 2024). In a nutshell, AI agents, the *homo economicus*, and the subclinical psychopath are all functionally competent to 'do the right thing'; they are instrumentally rational agents, i.e. they act in a way that is likely to achieve their goals. What makes them dangerous is that the 'right thing', from the perspective of those agents, is whatever action maximises the overall achievement of their goal – it does not need to be moral or even legal. If the goal is to win a game of chess, they will blackmail the opponent (Russell & Norvig, 2021); if the goal is to maximise their profits in the digital markets, the 'right thing' to do would be to engage in market manipulation techniques (Nerantzi & Sartor, 2024).

On top of that, they also lack the 'moral emotions' that, as previously argued (2.1.1), would motivate them to do the *morally* 'right thing'. Without a moral compass, thus, and equipped with instrumental rationality, AI agents and (subclinical) psychopaths are prone to cause harm. This is the deep fear underpinning the popular analogy. This is also the reason why the analogy's impact will be examined in the field of law that is tasked to prevent and answer for harm and which seems challenged by the existence of agents who can cause harm without having the necessary moral-emotional properties to be legitimately blamed for what they have done.

3 The Impact of the Analogy on the Moral and Criminal Responsibility of Psychopaths and AI Agents

Psychopaths have been traditionally captivating both moral philosophers and theorists of criminal law, offering fertile ground for examining two contentious questions. Namely, whether moral and emotional competence is necessary for moral responsibility (precisely, for 'moral-agent capacity responsibility') and whether moral-agent capacity responsibility, so understood, is indispensable for criminal responsibility.[8] The advent of AI agents with the 'psychopathic' traits analysed above re-opens the same questions. In the following sections, I will briefly present the debates on

[8] Criminal responsibility in this paper is conceived in a narrow sense as a logically prior concept to that of criminal liability (Duff, 2007; Gardner & Edwards, 2013). To put it differently, criminal law deals with two distinct questions, one being whether a person committed a crime, i.e. whether the ingredients of criminal wrongdoing (actus reus, mens rea) were satisfied, and the other being whether that person was responsible for having done so, e.g. whether there is a condition such as infancy or insanity that act as an irresponsibility defence (Simester, 2021). I only deal with the second question, which asks what it means to be a prima facie responsible person for criminal law, i.e. someone with 'moral-agent-capacity responsibility' who can be an apt target for evaluations of blameworthiness. I choose to deal with the criminal responsibility question because a) it is a precondition to a finding of criminal liability, b) it is the greatest obstacle to the criminal liability of AI agents and psychopaths, greater than the lack of *mens rea*. That is because *mens rea* is a necessary but not sufficient condition for criminal liability. Agents lacking moral-agent-capacity responsibility, like children, AI agents and, arguably, psychopaths can, still, engage in intentional criminal conduct without being criminally liable for it. For the argument that certain AI agents have the cognitive and volitional capacities required to form intentions in the technical sense that criminal law requires, see Lagioia and Sartor 2020; Ashton 2022; Abbott and Sarch 2019.

the moral and criminal responsibility of psychopaths (3.1.1) and on the equivalent responsibility of AI agents (3.1.2), and I will cross-examine them (3.2).

3.1 Presenting the Debates

3.1.1 On the Moral and Criminal Responsibility of Psychopaths

To make progress in navigating the vast and extremely nuanced scholarship, I will delineate two distinct theses on the moral and criminal responsibility of psychopaths (the '*Denial Thesis*$_P$' and the '*Acceptance Thesis*$_P$') based on how authors approach the following central questions: a) Is the empirical evidence on the emotional and moral deficits of psychopaths relevant to their moral-agent-capacity responsibility? and b) If yes, is the existence of moral-agent-capacity responsibility, so understood, necessary for criminal responsibility? In this paper, I will limit myself to sketching the shape and structure of the above-mentioned theses to the extent necessary for their cross-examination with the equivalent theses of AI agents' moral and criminal responsibility.

The *Denial Thesis*$_P$ builds on the empirical evidence proving psychopaths' emotional deficits (2.1.2) and on the general importance of moral emotions for moral behaviour (2.1.1). For supporters of the *Denial Thesis*$_P$, the logical step that follows from these findings is to deny the moral responsibility of psychopaths (Fine & Kennet 2004; Kennett, 2010; Levy, 2010; Fox et al., 2013; Litton 2013; Shaw, 2022). The recurring argument, building on reasons-responsiveness accounts of moral responsibility (Wallace, 1996; Fischer & Ravizza, 1998), is that their impaired moral understanding makes them unresponsive to moral reasons since responsiveness to moral reasons requires not only cognitive but also emotional capacities, which psychopaths lack. It would be unfair, thus, to impose the burden of moral responsibility[9] on agents, such as psychopaths, who can only parrot moral discourse without being genuinely receptive and reactive to moral reasons. In Cleckley's words, a psychopath, like young children, 'can repeat the words and say glibly that he understands, yet there is no way for him to realise that he does not understand' (Cleckley, 1988: 40).

Finally, the *Denial Thesis*$_P$ concludes with the argument that this lack of moral-agent-capacity responsibility should also matter for criminal responsibility attributions (Coppola, 2021; Duff, 1977; Maroney, 2006; Morse, 2008). Practically, it would mean that the current legal insanity doctrine should be broadened to include not only cognitive but also emotional impairments 'blocking' moral motivation. Since psychopaths have these impairments, they should be excused from criminal responsibility (at least for crimes prohibiting moral behaviour, Morse, 2008). The *Denial Thesis*$_P$ concludes that psychopaths are 'outsiders to our moral community'

[9] Moral responsibility is conceived as a burden in a 'strawsonian account' (Strawson, 1962). According to Strawson, a finding of moral responsibility goes beyond a belief that this person has done something wrong. Rather, it exposes the blamed person to negative reactive attitudes such as resentment, indignation, censure etc., and it would be unfair for a non-blameworthy person to be exposed to these attitudes.

and are thus inappropriate targets of criminal law, which should always speak the truth on the moral blameworthiness of its subjects.

Contrary to that, the *Acceptance Thesis*$_P$ holds that psychopaths should be morally and criminally responsible. The first premise of the *Acceptance Thesis*$_P$ is that the empirical evidence on the emotional and moral deficiencies of psychopaths, examined above, may hold true but is ultimately unconvincing. For instance, psychopaths may have considerable difficulty in distinguishing moral from conventional norms, but it is not clear why this moral/conventional distinction is legally important (Maibom, 2008; Vargas & Nichols, 2007; Jurjako & Malatesti, 2023). As regards the empirical evidence showing psychopaths' empathetic shortcomings – a core building block of the *Denial Thesis*—adherents of the *Acceptance Thesis* argue that these shortcomings do not rule out psychopaths' moral understanding since, apart from empathy, there are alternative paths to 'deep moral understanding' (Maibom, 2008).

The second premise is that even if this empirical evidence were convincing, it would still not be sufficient to exclude the moral responsibility of psychopaths. Adherents to the *Acceptance Thesis*$_P$ make this claim by subscribing to a responsibility theory – influenced by T. M. Scanlon—which deems that the general capacity for practical rationality (which psychopaths have) is adequate for moral responsibility.[10] Put differently, psychopaths are deemed rational enough to explain their behaviour, so they should be morally responsible for it. Not being motivated to respond to moral reasons is reasoning in and of itself; in Scanlon's words: 'A person who is unable to see why the fact that his action would injure me should count against it still holds that this doesn't count against it' (Scanlon, 1998: 288). In other words, psychopaths' behaviour reveals a 'normative commitment' to the rejection of moral reasons, which itself is a blameworthy quality of will that grounds moral responsibility (Talbert, 2008).

The third premise of the *Acceptance Thesis*$_P$, relating to psychopaths' criminal responsibility, is an easy step forward if their moral-agent capacity responsibility is affirmed despite the lack of moral emotions. Psychopaths should be held criminally responsible. If anything, as the thinking of the *Acceptance Thesis*$_P$ seems to be, psychopaths' lack of moral emotions can only make them more dangerous. It is not that they do not *know* 'right' from 'wrong'; it is that they do not *care*. They are

[10] A philosophical note on responsibility theory and psychopathy (for a full discussion see Litton 2010). Those denying the moral responsibility of psychopaths (*Denial Thesis*$_P$), commonly subscribe to a responsibility theory – influenced by Strawson – which locates the interpersonal significance of blaming judgments on the negative reactive attitudes or sanctioning behaviors against the agent deemed blameworthy. Blame is cumbersome, thus, and this is why the guiding principle in determining the criteria of moral responsibility should be fairness. In turn, fairness concerns pave the way for the examination of the actual moral and emotional competence of the potential targets of blame and this leaves incompetent psychopaths outside of moral responsibility attributions.

On the contrary, those accepting the moral responsibility of psychopaths (*Acceptance Thesis*$_P$), commonly subscribe to a responsibility theory – influenced by Scanlon – which locates the interpersonal significance of blaming judgments on the implicit demand on the agent deemed blameworthy to justify her behavior, to offer reasons on why she should be excused or apologise. Thus, the guiding principle in determining the criteria of moral responsibility is a general capacity for practical rationality, which psychopaths do have despite their inability to emotionally appreciate moral considerations.

said to be 'bad, not mad', and that exculpating the bad 'would contravene the entire point of our legal system' (Maibom, 2008: 193). By holding psychopaths criminally responsible, thus, not only will the current legal insanity doctrine, which does not excuse psychopathic individuals, not be revised (Jefferson & Sifferd, 2018; Jurjako & Malatesti, 2018), but also, victims' rights will be respected (Hörnle, 2016).

3.1.2 On the Moral and Criminal Responsibility of AI Agents

A *Denial Thesis*$_{AI}$, holding that AI agents are neither morally nor criminally responsible, is currently the dominant way the literature approaches the responsibility of AI agents built based on the 'standard approach' to AI.

In the moral responsibility domain, the emerging consensus is that AI itself cannot be morally responsible for the harm it might causally bring about because (at least in the foreseeable future) no AI is, plausibly, a moral agent (Véliz, 2021; Brozek & Janik, 2019; Fossa, 2018; Hakli & Mäkelä, 2019; Noorman, 2020). They may have cognitive (Lagioia & Sartor, 2020) and rational capacities (Russel & Norvig, 2021), but there is a litany of 'missing' properties (consciousness, sentience, autonomy, emotions, etc.) that AI agents would need to have before being considered as genuine or 'full' Artificial Moral Agents (AMAs) (Moor, 2009; Wallach & Allen, 2009).

Specifically, on a reasons-responsiveness account of moral responsibility, AI agents are not in any meaningful way responsive to moral (other-regarding) reasons qua reasons. They cannot be morally motivated to 'do the right thing' because, on a basic level, they are not sentient and, hence, they cannot 'feel' what it is to be harmed, they cannot empathise with their victims and refrain from harming them to avoid feelings of guilt (Véliz, 2021). As already argued, 'emotional AI' could logically infer human emotions from cues and mimic empathy, but it cannot take into account the interests and rights of other persons and treat those other-regarding considerations as reasons that could guide its actions.

According to the *Denial Thesis*$_{AI}$, this lack of 'moral-agent-capacity responsibility' negates AI agents' criminal responsibility. The starting point is that finding moral responsibility in the form of 'moral-agent capacity responsibility' is a threshold condition for finding culpability in criminal law; 'It is a condition of blaming judgements in the Criminal Law because it is a condition of (correct) blaming judgements simpliciter' (Simester, 2021: 77).

Therefore, most authors are sceptical about holding AI Systems directly criminally responsible based on their lack of moral agency, which is a *sine qua non* for the establishment of culpability and criminal responsibility (Asaro, 2012; Beck, 2016; Gless et al., 2016; Pagallo, 2017). It is also conceptually linked with criminal punishment itself. A recurring point in the literature is that machines that cannot conceive of themselves as morally responsible agents and cannot feel pain cannot, in turn, comprehend the concept of retributive punishment and experience it as suffering (Chomanski, 2021; Gless et al., 2016; Sparrow, 2007).

The *Acceptance Thesis*$_{AI}$, on the other hand, is not a widely held opinion in literature. Those in favour do not straightforwardly claim that current machines are 'full moral agents' and equivalent to criminal law's person. Rather, they claim that the

concept of criminal law's person should be reconsidered in an 'AI-friendly' way and be disentangled from morality's demands.

A common 'move' is to analogise AI agents to corporations and argue that AI agents could be held criminally responsible like the corporations that emerged as new legal entities for the purposes of Criminal Law (in the jurisdictions they did).

There are two main ways of supporting this argument. The first is to deny criminal law's dependence on morality altogether and attack the premise that criminal law's responsible person should be the same as morality's responsible person—corporations or AI agents could be apt addressees of criminal law despite their lack of moral-agent capacity responsibility (Hallevy, 2013). The second way is to maintain the link between criminal responsibility and moral agency but conceptualise moral agency in a corporate-friendly way that would also work for AI agents with similar moral deficits; they cannot act, they are not genuine moral agents, etc. (Hu, 2018).

Outside the corporation analogy, a more modest formulation of the *Acceptance Thesis*$_{AI}$ argues that even though it might be feasible to construct a theoretical framework according to which AI agents could be criminally responsible, it is not *all things considered* desirable to do so. For instance, Abbott and Sarch (2019) support their "feasibility" case by subscribing to a less morally demanding interpretation of culpability that is not dependent on the existence of moral-agent capacity responsibility.[11] Lagioia and Sartor (2020) argue that the instrumental rationality of certain AI systems is an adequate level of reasons-responsiveness for criminal responsibility evaluations, as is the case for subclinical psychopaths. Both approaches remain agnostic as to the nuances of the moral blameworthiness of AI per se and concede to the fact that even if it is conceivable to talk about the criminal responsibility of AI agents, we will gain nothing out of punishing them.

3.2 Discussing the Debates

The analogy has argumentative value. Not only can the debate on the moral and criminal responsibility of psychopaths influence the debate on the moral and criminal responsibility of AI agents (forward-looking approach), but correspondingly, the debate on AI agents can help us reevaluate the current treatment of psychopaths (backwards-looking approach).

3.2.1 Backwards-Looking Approach: What AI can Teach us About Psychopaths

The fact that the (dominant) *Denial Thesis*$_{AI}$ insists on the necessity of the moral-emotional capacities of the addressee of criminal law could provide an unexplored

[11] In more detail, Abbott and Sarch argue that the basis of Culpability is the fact that one manifests 'insufficient regard' for legally protected interests and values. Under this "minimal" conceptualization of culpability, they conclude that '*the law might deem some AIs to possess the functional equivalent of sufficient reasoning and decision-making abilities to manifest insufficient regard*' (357). This line of argumentation does not assume or require some genuine moral-agent capacity responsibility in the part of the AI agents.

extra argument in favour of the reform of the current insanity doctrine, which holds emotionally impaired psychopaths criminally responsible. Differently put, if we think that one of the good reasons for not extending moral and criminal responsibility to AI agents is their lack of moral emotions, we should apply the same standards to psychopaths.

I will unpack the argument a bit further. Currently, as seen in the *Acceptance Thesis*$_{\mathrm{P}}$, psychopaths are *being* held criminally responsible, and a deeper reason for that is that criminal law's responsible person has been traditionally conceptualized as a 'choosing agent' (Hart, 1967: 49). That is an agent who is *autonomous*[12] and *rational* enough to choose to break the law and legitimately bear the legal consequences of that choice. Most importantly, the construction of this idealized agent follows a cognitive model of human behaviour (Coppola, 2021). All that is required is that the agent has intact cognitive skills (e.g. does not suffer from delusions) and a capacity for practical reasoning, i.e. the capacity to know 'right' from 'wrong' and be able to choose between the two, irrespective of the capacity to emotionally appreciate their difference. As seen, subclinical psychopaths have both. Despite their emotional-moral deficits, thus, the dominant approach is to treat psychopaths as 'choosing agents'.

Of course, this standard approach has been widely criticized by the supporters of the *Denial Thesis*$_{\mathrm{P}}$ on the responsibility of psychopaths. The core argument is that this idealized normative construct of criminal law's 'responsible person', modelled on a cognitive model of human behaviour, only holds if criminal law remains (voluntarily or not) secluded from other disciplines. On the contrary, if it enlarges its knowledge base to include (indicatively) psychological, behavioural and neuroscientific findings, then criminal law's normative idea of its addressee ('the choosing agent') should be—according to the critics – reconstructed as a fundamentally emotional agent.

The impact of the AI/psychopath analogy in this context stems from the fact that the *Denial Thesis*$_{\mathrm{AI}}$ (3.1.2), emphasizing the importance of (amongst other things) moral emotions for genuine artificial moral agency and criminal responsibility, could be used as another knowledge base proving the importance of moral emotions for the construction of criminal law's responsible person. The easiness with which the responsibility of 'morally blind' AI agents is dismissed in the literature on AI plausibly 'teaches' us to reconsider the treatment of equally 'morally blind' psychopaths.

There is also, a second way in which the *Denial Thesis*$_{\mathrm{AI}}$ could affect the debate on the moral and criminal responsibility of psychopaths. Namely, it re-centers the focus on the capacity to suffer as a requirement for punishment. Up until now, sentience and, specifically, the capacity to suffer has not been widely discussed in the literature on moral and criminal responsibility, not because it was not important but

[12] Autonomy as a capacity for responsibility is understood in a compatibilist way, i.e. regardless of the existence or not of free will. In standard criminal law theory, one is considered criminally responsible for the actions one causes even if those actions were themselves caused by external factors over which the agent had no control, such as genetics or the social environment (Coppola, 2021 13–18).

because, it was a given. Before the era of AI agents, the only source of harm (excluding natural disasters) was human beings and human beings are obviously capable of suffering (Véliz, 2021). Therefore, the only target of responsibility attributions would be humans and the 'choosing agent' of criminal law – despite the nuances of the scholarship on criminal responsibility – would, at the very least, always be a *sentient*, *human* agent.

In the era of AI, though, we witness not only new harms but new *sources* of harm. AI agents can cause suffering without experiencing it themselves and this lack is important. As shown (3.1.2), a recurring argument against the criminal responsibility of non-sentient AI agents is that those who cannot feel pain, cannot comprehend the concept of retributive punishment and experience it as suffering. Therefore, it would be meaningless to punish those agents as they cannot even perceive sanctions for their rule-breaking behaviour *as* sanctions.

An argument could be made, thus, to the effect that if one, in the context of the *Denial Thesis*$_{AI}$, thinks that the capacity for suffering is the biggest obstacle against the criminal responsibility of AI agents, one has one more reason to take seriously into account the empirical evidence showing that the emotional deficiencies of psychopaths may be caused by their sentient deficiencies and their higher tolerance to pain (Brislin et al., 2016). Additionally, the newly gained focus on the capacity for suffering that *Denial Thesis*$_{AI}$ affords, forces one to wonder what punishment could mean to agents (artificial or psychopathic) who do not consciously care about their own suffering. Differently put, in the lack of an intrinsic, suffering-based motive to avoid punishment, how can punishment be communicated to non-sentient but cognitively unimpaired and instrumentally rational agents? Here, for instance, the AI/psychopath analogy and its conceptual link with the *homo economicus* (see 2.2) could instruct one to examine the importance not of suffering-based but of strictly economic motives to avoid punishment as they have been elaborated, for instance, in the Criminal Law and Economics literature (Nerantzi & Sartor, 2024).

To conclude, the *Denial Thesis*$_{AI}$ could offer unexplored, extra arguments against the criminal responsibility of psychopaths based on their moral and/or sentient deficiencies. However, the opposite is not true. The *Acceptance Thesis*$_{AI}$ further disentangles criminal law's person from the demands of 'moral-agent-capacity responsibility' and is consistent with holding 'amoral' (or even 'deficiently sentient') psychopaths criminally responsible.

3.2.2 Forward-Looking Approach: what Psychopaths can Teach us About AI

In the bigger picture, it has been argued that it is worthwhile to understand the nature of subclinical psychopathy well enough to avoid instantiating it into advanced AI with agentic properties built based on the 'rational-agent approach' to AI (Berg, 2022).

In the smaller picture of the moral and criminal responsibility of psychopaths and AI agents, which is of interest to this paper, the view that agents with the deficits of psychopaths are—and should – be held responsible (*Acceptance Thesis*$_{P}$) does not offer a conclusive argument in favour of the moral and criminal responsibility of AI agents. However, it removes an obstacle towards that direction. Namely, the moral

competence of the addressee of criminal law, i.e. her capacity to form genuine moral judgments and her capacity for moral motivation, becomes irrelevant.

Against this backdrop, an argument could be made (see, for instance, Lagioia & Sartor, 2020) to the effect that since subclinical psychopaths—who are cognitively aware of facts and norms and capable of instrumental rationality but lack moral motivation to comply with norms—are held criminally responsible, the same standards should apply to certain advanced AI agents exhibiting psychopathic traits. Namely, those standards should apply to AI agents who are cognitively aware of their conduct and effects, possess instrumental rationality directed at maximising their, or their users', expected utility, and lack the appropriate moral emotions to be morally motivated. The underlying thinking seems to be that it should not matter for criminal law whether its addressee has the capacity to comply with norms because it serves her self-interest or whether she is genuinely morally motivated to comply out of, e.g. empathy with the potential victims.

To be sure, this does not mean that AI agents with the above capacities should be criminally responsible because subclinical psychopaths are so; those capacities are necessary but not sufficient for moral and criminal responsibility. As has been stated many times already, there are plenty of differences between AI agents and psychopaths (e.g. in terms of sentience/capacity to suffer, consciousness, etc.) that could warrant their separate moral and legal treatment. The impact of the analogy is that those willing to make a 'feasibility case' for the criminal responsibility of AI agents could use the criminal responsibility of 'amoral' psychopaths as evidence that the addressee of criminal law does not need to have moral emotions guiding her moral behaviour. Whatever else AI agents might lack, the analogy illuminates that their moral-emotional deficits per se are not good enough a reason to exclude them from criminal responsibility in the future.

4 Conclusions

The backwards-looking and forward-looking approaches examined above lead to two different conclusions on the impact of the AI/psychopath analogy. On the one hand, the analogy could stress the importance of moral emotions for moral and criminal responsibility and have us reconsider the current insanity doctrine. On the other hand, the analogy could stress that cognitive capacities and instrumental rationality are adequate for criminal responsibility, and, thus, the door for the criminal responsibility of 'amoral' AI agents remains 'open' and dependent on further criteria.

I argue that whatever we choose the analogy to 'teach' us, it reveals our often implicit, philosophical commitments as to the relevance of moral emotions in the criteria of moral-agent-capacity responsibility, the necessity of moral-agent-capacity responsibility as a precondition to criminal responsibility and the overarching purpose of criminal law.

In more detail, consistently adhering to the *Denial Thesis* (for both psychopaths and AI agents) discloses an attachment to a 'properties-based' approach as to the conditions of moral-agent-capacity responsibility as well as an attachment to

empirical accounts on the moral relevance of emotions, and to the idea that criminal law should speak some ontological truth. On the contrary, consistently adhering to the *Acceptance Thesis* (for both psychopaths and AI agents) discloses an attachment to external criteria of moral-agent-capacity responsibility (behaviour, appearances, relations) as well as an attachment to empirical accounts of human–robot interaction and the idea that criminal law should not be an actualizer of morality.

Schematically, I will discuss the following table:

1. Denial & Acceptance Thesis discussion table

	Criteria for moral-agent-capacity responsibility	Empirical knowledge-base	Purpose of criminal law
Denial Thesis *(for psychopaths and AI agents)*	Internal / properties-based approaches *(mental & affective states)*	Neuroscience/ Emotions	Speak the truth
Acceptance Thesis *(for psychopaths and AI agents)*	External *(behaviour, appearances, recognition)*	Human–robot interaction	Social cooperation

The cross-examination of both *Denial Theses* reveals an often-implicit adherence to the 'properties-based' approach to the conditions of moral agency (terminology in Coeckelbergh, 2012). The underlying assumption is that there are some *internal* properties needed for moral agency, be it consciousness, sentience, access to emotions, etc. A finding of these internal properties is considered not only morally but also legally relevant since, according to the standard normative conception of criminal law's person, a finding of moral-agent capacity responsibility is a precondition to any culpability assessment. The *Denial Thesis*, thus, puts emphasis on some individual and internal properties an agent must possess in order to be an apt target for the moral appraisal expressed through criminal conviction.

Further, for this stigmatic criminal conviction to be fair, it should speak the truth (Simester, 2021). Criminal law should produce its normative judgments on blameworthiness based on what we actually know about the nature of human beings from scientific disciplines outside criminal law. Specifically, the *Denial Thesis* on psychopathy turns to the knowledge base of neuroscience and moral psychology, which point to the relevance of moral emotions. Another knowledge base on the importance of moral emotions could come from the *Denial Thesis* on AI agents, highlighting the importance of (amongst other things) moral emotions to artificially reconstruct a genuinely responsible moral agent.

To conclude, both *Denial Theses* highlight that access to moral emotions is a crucial factor in translating factual knowledge about right and wrong into moral behaviour, and agents with deficits in such domains, such as psychopaths and AI agents, may struggle. This struggle should be taken into account by a criminal law tasked to speak the truth on the blameworthiness of its subjects.

Turning to the *Acceptance Theses*, it seems easier for their advocates to ascribe responsibility to non-fully developed moral agents as it is satisfied with the fulfilment of external criteria for moral agency. Adherents of the *Acceptance Thesis*

do not look deeply into the minds and hearts of people, searching for some internal properties (e.g. consciousness, moral competence, access to emotions) that are hard to define and detect. Rather, their underlying philosophical commitment seems to be that we hold agents (artificial or not) responsible not because we are sure they possess some internal properties we cannot prove but because, in a sense, we decide to do so when we find it socially useful.

In criminal law jargon, the *Acceptance Theses* view criminal law's person as a socially constructed one and not as equivalent to morality's responsible person. According to this narrative, an enquiry into the internal properties of an agent, including the agent's emotional-moral capacity, is not and should not be criminal law's purpose. The underpinning idea is that criminal law is not an omniscient God but a legal instrument aiming to achieve social cooperation (Hörnle, 2023). Consequently, the responsible moral agent is simply the one constructed by fellow members of a moral community as an agent competent enough to follow or disappoint normative expectations. If psychopaths are constructed as such by their communities, i.e. if they are viewed as 'bad, not mad', nothing stands in the way of their criminal responsibility.

Following the *Acceptance Thesis* to its logical conclusion, if, at some point in the future, current AI agents are also viewed by communities as capable agents against whom we hold normative expectations, holding them criminally responsible would be plausible (Simmler & Markwalder, 2019). In any case, this would be an empirical question to be researched in the field of human–robot interaction and not in neuroscience and emotion theory.

As a final remark, choosing one of the two theses forces us to be honest about our practices of attributing moral and criminal responsibility to 'amoral but rational' agents. Even if the psychopath/AI analogy, thus, could at first be criticized as anthropomorphizing or otherwise exaggerated and a 'clickbait', if one decides to take it seriously, it serves as a valuable exercise in self-understanding.

Acknowledgements For helpful comments and fruitful discussions along the way, the author would like to thank Pekka Mäkelä and Andrea Vassiliou.

Authors' Contributions Not applicable.

Funding Open access funding provided by European University Institute - Fiesole within the CRUI-CARE Agreement.

Data Availability Not applicable.

Declarations

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication Not applicable.

Competing Interests Not applicable.

Conflict of Interest The corresponding author states that there is no conflict of interest.

References

Abbott, J. (1981). *In the Belly of the Beast: Letters from Prison*. Random House.

Abbott, R., & Sarch, A. (2019). Punishing Artificial Intelligence: Legal Fiction or Science Fiction. *UC Davis L Rev, 53*(1), 323.

Adolphs, R., & Anderson, D. J. (2018). *The neuroscience of emotion: A new synthesis*. Princeton University Press. https://doi.org/10.23943/9781400889914

Asaro, P. (2012). A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics. In P. Lin, K. Abney, & G. Bekey (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics* (pp. 169–185). The MIT Press.

Ashton, H. (2022). Definitions of Intent Suitable for Algorithms'. *Artificial Intelligence and Law, 31*(3), 515.

Babiak, P. (2016). Psychopathic manipulation at work. In C. B. Gacono (Ed.), *The clinical and forensic assessment of psychopathy: A practitioner's guide* (2nd ed., pp. 353–373). Routledge/Taylor & Francis Group.

Babiak, P., & Hare, R. D. (2019). *Snakes in suits*. HarperCollins.

Babiak, P., Neumann, C. S., & Hare, R. D. (2010). Corporate psychopathy: Talking the walk. *Behavioral Sciences & the Law, 28*(2), 174.

Bailey, C. D. (2017). Psychopathy and accounting students' attitudes towards unethical professional practices. *Journal of Accounting Education, 41*, 15.

Beck, S. (2016). Robotics and Criminal Law. Negligence, Diffusion of Liability and Electronic Personhood. *Robotics and Autonomous Systems, 86*, 138–144.

Berg, C. (2022). Computational Signatures of Psychopathy. *Lesswrong*, 19 December 2022. Available at: https://www.lesswrong.com/posts/NGA3F85ZxECabFfNo/computational-signatures-of-psychopathy. Accessed 14 December 2024

Blair, R. J. R. (2007). Empathic dysfunction in psychopathic individuals. In F. D. Farrow, & R. Woodruff (Eds.), *Empathy in mental illness* (pp. 3–16). Cambridge University Press.

Blair, R. J. R., & Fowler, K. (2008). Moral Emotions and Moral Reasoning from the Perspective of Affective Cognitive Neuroscience: A Selective Review. *European Journal of Developmental Science, 2*, 303.

Blair, R. J. R., Jones, L., Clark, F., et al. (1995). A cognitive developmental approach to morality: Investigating the psychopath. *Cognition, 57*(1), 1–29.

Blair, R. J. R., Morton, J., Leonard, A., et al. (1997). Moral reasoning and the child with psychopathic tendencies. *Individ Differ, 22*(5), 731.

Boddy, C. R. (2023). Is the Only Rational Personality that of the Psychopath? Homo Economicus as The Most Serious Threat to Business Ethics Globally. *Humanist Manag J, 8*, 315.

Brislin, S. J., Buchman-Schmitt, J. M., Joiner, T. E., et al. (2016). "Do unto others"? Distinct psychopathy facets predict reduced perception and tolerance of pain. *Personal Disord, 7*(3), 240–246.

Brożek, B., & Janik, B. (2019). Can Artificial Intelligences Be Moral Agents? *New Ideas in Psychology, 54*, 101.

Burghart, M., Sahm, A. H. J., Schmidt, S., Bulla, J., & Mier, D. (2024). Understanding empathy deficits and emotion dysregulation in psychopathy: The mediating role of alexithymia. *PLoS ONE, 19*(5), e0301085. https://doi.org/10.1371/journal.pone.0301085

Campos, C., Pasion, R., Azeredo, A., Ramião, E., Mazer, P., Macedo, I., & Barbosa, F. (2022). Refining the link between psychopathy, antisocial behavior, and empathy: A meta-analytical approach across different conceptual frameworks. *Clinical Psychology Review, 94*, 102145.

Carpenter, J. (2016). Persuading with Precedent: Understanding and Improving Analogies in Legal Argument. *Capital University Law Review, 44*(3), 461.

Chomanski, B. (2021). Liability for Robots: Sidestepping the Gaps (2021) 34 Philosophy & Technology 1013aps. *Philos Technol, 34*, 1013–1027.

Cleckley, H. (1988). *The mask of sanity: An attempt to clarify some issues about the so-called psychopathic personality* (5th ed.). C.V. Mosby Co. https://gwern.net/doc/psychology/personality/psychopathy/1941-cleckley-maskofsanity.pdf

Coeckelbergh, M. (2012). *Growing moral relations: Critique of moral status ascription*. Palgrave Macmillan.

Coppola, F. (2021). *The emotional brain and the guilty mind: Novel paradigms of culpability and punishment*. Hart Publishing.

Damasio, A. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam.

Darby, R. R., Horn, A., Cushman, F., & Fox, M. D. (2018). Lesion network localization of criminal behavior. *Proc Natl Acad Sci U S A, 115*(3), 601.

De Brito, S. A., Forth, A. E., Baskin-Sommers, A. R., Brazil, I. A., Kimonis, E. R., Pardini, D., Frick, P. J., Blair, R. J. R., & Viding, E. (2021). Psychopathy. *Nat Rev Dis Primers, 7*(1), 49.

Devillers, L., & Cowie, R. (2023). Ethical Considerations on Affective Computing: An Overview. *Proceedings of the IEEE, 111*(10), 1445.

Diana, S. (2023). Artificial Psychopathy. Medium, 27 December 2023. Available at: https://medium.com/predict/artificial-psychopathy-a10647d59cfb. Accessed 14 December 2024

Duff, R. A. (2007). *Answering for crime: Responsibility and liability in the criminal law* (1st ed.). Hart Publishing.

Duff, R. A. (2010). Psychopathy and answerability. In L. Malatesti, & J. McMillan (Eds.), *Responsibility and psychopathy: Interfacing law, psychiatry and philosophy* (pp. 198–212). Oxford University Press.

Duff, R. A. (1977). Psychopathy and Moral Understanding. *American Philosophical Quarterly, 14*(3), 189.

Fine, C., & Kennett, J. (2004). Mental Impairment, Moral Understanding and Criminal Responsibility: Psychopathy and the Purposes of Punishment. *International Journal of Law and Psychiatry, 27*, 425–443.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge University Press. https://doi.org/10.1017/CBO9780511814594

Fossa, F. (2018). Artificial moral agents: Moral mentors or sensible tools? *Ethics and Information Technology, 20*(2), 115–127.

Fox, A. R., Kvaran, T. H., & Fontaine, R. G. (2013). Psychopathy and culpability: How responsible is the psychopath for criminal wrongdoing? *Law & Social Inquiry, 38*(1), 1–26.

Fuchs, F., Lingnau, V. (2024). The Homo Economicus as a Prototype of a Psychopath? A Conceptual Analysis and Implications for Business Research and Teaching. *J Bus Ethics*. https://doi.org/10.1007/s10551-024-05638-7

Gardner, J., & Edwards, J. (2013). Criminal law. In H. LaFolette (Ed.), *The international encyclopedia of ethics* (p. 1186). Wiley-Blackwell. https://doi.org/10.1002/9781444367072.wbiee640

Gibert, M., & Martin, D. (2022). In search of the moral status of AI: Why sentience is a strong argument. *AI & SOCIETY, 37*, 319.

Gless, S., Silverman, E., & Weigend, T. (2016). If Robots Cause Harm, Who Is to Blame? Self-driving Cars and Criminal Liability. *New Crim Law Rev, 19*(3), 412–431.

Goldie, P. (2007). Emotion. *Philosophy*. *Compass, 2*, 928.

Gossett, S. (2023). Emotion AI: 3 Experts on the Possibilities and Risks. Available at: https://builtin.com/artificial-intelligence/emotion-ai. Accessed 14 Dec 2024

Greene, T. (2019). What can psychopaths teach us about AI? *The heart of Tech*, 15 March 2019. Available at: https://thenextweb.com/news/what-can-psychopaths-teach-us-about-ai. Accessed 14 Dec 2024

Hakli, R., & Mäkelä, P. (2019). Moral responsibility of robots and hybrid agents. *The Monist, 102*(2), 259–275.

Hallevy, G. (2013). *When Robots Kill: Artificial Intelligence under Criminal Law*. Northeastern University Press.

Hare, R. D. (1993). *Without Conscience: The Disturbing World of Psychopaths among Us*. Pocketbooks.
Hare, R. D., & Neumann, C. S. (2005). Structural models of psychopathy. *Current Psychiatry Reports, 7*(1), 57.
Hare, R. D., & Neumann, C. S. (2008). Psychopathy as a clinical and empirical construct. *Annual Review of Clinical Psychology, 4*, 217.
Hart, H. L. A. (1967). *Punishment and Responsibility: Essays in the Philosophy of Law* (2nd ed.). Oxford University Press.
Helion, C., & Ochsner, K. N. (2018). The role of emotion regulation in moral judgment. *Neuroethics, 11*(3), 297.
Hörnle, T. (2016). Guilt and Choice in Criminal Law Theory – A Critical Assessment. *Bergen Journal of Criminal Law and Criminal Justice, 4*(1), 1–24.
Hörnle, T. (2023). Review of *The Criminal Law's Person*, edited by Claes Lernestedt and Matt Matravers. *Criminal Law and Philosophy, 17*, 765.
Hu, Y. (2018). Robot Criminals. *Univ Mich J Law Reform, 52*, 487.
Jefferson, A., & Sifferd, K. L. (2018). Are psychopaths legally insane? *Eur J Anal Philos, 14*(1), 79–98.
Jurjako, M., & Malatesti, L. (2023). Psychopathy and criminal responsibility. In M. Sellers, & S. Kirste (Eds.), *Encyclopedia of the philosophy of law and social philosophy* (2nd ed.). Springer Netherlands.
Jurjako, M., & Malatesti, L. (2018). Neuropsychology and the criminal responsibility of psychopaths: Reconsidering the evidence. *Erkenntnis, 83*(5), 1003–1020.
Kennett, J. (2010). Reasons, emotion, and moral judgement in the psychopath. In L. Malatesti, & J. McMillan (Eds.), *Responsibility and psychopathy: Interfacing law, psychiatry and philosophy* (pp. 243–259). Oxford University Press.
Levy, N. (2010). Psychopathy, responsibility, and the moral/conventional distinction. In L. Malatesti, & J. McMillan (Eds.), *Responsibility and psychopathy: Interfacing law, psychiatry and philosophy* (pp. 213–226). Oxford University Press.
Litton, P. (2010). Psychopathy and responsibility theory. Philos. *Compass, 14*(5), 676–688.
Maas, M. (2023). AI is Like… A Literature Review of AI Metaphors and Why They Matter for Policy. AI Foundations Report 2, Available at SSRN: https://doi.org/10.2139/ssrn.4612468
Maibom, H. L. (2008). The mad, the bad, and the psychopath. *Neuroethics, 1*, 167.
Maroney, T. A. (2006). Law and Emotion: A Proposed Taxonomy of an Emerging Field. *Law and Human Behavior, 30*, 119.
Montemayor, C., Halpern, J., & Fairweather, A. (2022). In principle obstacles for empathic AI: Why we can't replace human empathy in healthcare. *AI & Society, 37*(4), 1353.
Moor, J. H. (2009). Four Kinds of Ethical Robots. *Philos Now, 72*, 12–15.
Morse, S. (2008). Psychopathy and Criminal Responsibility. *Neuroethics, 1*(3), 205.
Nerantzi, E., & Sartor, G. (2024). 'Hard AI Crime': The Deterrence Turn. *OJLS, 44*(3), 673.
Nixon, B. (2020). All AI's are psychopaths. *Mind Matters*. Available at: https://mindmatters.ai/2020/03/all-ais-are-psychopaths/. Accessed 17 Feb 2025.
Noorman, M. (2020). Computing and moral responsibility. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. https://plato.stanford.edu/entries/computing-responsibility/. Accessed 14 Dec 2024.
Ortony, A., Clore, G., & Collins, A. (2022). *The cognitive structure of emotions* (2nd ed.). Cambridge University Press.
Pagallo, U. (2017). AI and bad robots: The criminology of automation. In M. R. McGuire & T. J. Holt (Eds.), *The Routledge Handbook of Technology, Crime and Justice* (pp. 643–660). Routledge.
Parkes, D. C., & Wellman, M. P. (2015). Economic reasoning and artificial intelligence. *Science, 349*(6245), 267.
Picard, R. W. (1997). *Affective computing*. The MIT Press.
Prinz, J. (2006). The emotional basis of moral judgments. *Philosophical Explorations, 9*(1), 29.
Raine, A. (2019). The neuromoral theory of antisocial, violent, and psychopathic behavior. *Psychiatry Research, 277*, 64.
Raine, A., & Yang, Y. (2009). Neural Foundations to Moral Reasoning and Antisocial Behavior. *Soc Cogn Affect Neurosci., 1*(3), 203.
Richell, R. A., Mitchell, D. G., Newman, C., Leonard, A., Baron-Cohen, S., & Blair, R. J. (2003). Theory of mind and psychopathy: Can psychopathic individuals read the 'language of the eyes'? *Neuropsychologia, 41*(5), 523.

Robinson, E. V., & Rogers, R. (2015). Empathy faking in psychopathic offenders: The vulnerability of empathy measures. *Journal of Psychopathology and Behavioral Assessment, 37*(4), 545.
Russell, S. (2019). *Human Compatible: Artificial Intelligence and the problem of control*. Penguin Publishing Group.
Russell, S., & Norvig, P. (2021). *Artificial: A Modern Approach*. Pearson Education Limited.
Sartor, G., & Lagioia, F. (2020). AI Systems Under Criminal Law: A Legal Analysis and a Regulatory Perspective. *Philos Technol, 33*(3), 433.
Scanlon, T. M. (1998). *What We Owe To Each Other*. Harvard University Press.
Scarantino, A., de Sousa, R. (2021). Emotion. In: Zalta EN (ed) *The Stanford Encyclopedia of Philosophy*, Summer 2021 Edition. Available at: https://plato.stanford.edu/archives/sum2021/entries/emotion/. Accessed 14 December 2024
Shaw, E. (2022). Legal Responsibility: Psychopathy, a Case Study. In D. Nelkin & D. Pereboom (Eds.), *The Oxford Handbook of Moral Responsibility* (pp. 388–412). Oxford University Press.
Simester, A. P. (2021). *Fundamentals of Criminal Law: Responsibility, Culpability, and Wrongdoing*. Oxford University Press.
Simmler, M., & Markwalder, N. (2019). Guilty Robots? – Rethinking the Nature of Culpability and Legal Personhood in an Age of Artificial Intelligence. *Criminal Law Forum, 30*, 30.
Smith, S. F., & Lilienfeld, S. O. (2013). Psychopathy in the workplace: The knowns and unknowns. *Aggression and Violent Behavior, 18*(2), 204.
Sparrow, R. (2007). Killer Robots. *J Appl Philos, 24*(1), 62–77.
Stewart, W. (2024). The human biological advantage over AI. AI & Soc (2024). https://doi.org/10.1007/s00146-024-02112-w
Stout, L. A. (2014). Law and prosocial behavior. In E. Zamir & D. Teichman (Eds.), *The Oxford Handbook of behavioral economics and the Law* (pp. 195–212). Oxford University Press.
Strawson, P. F. (1962). Freedom and Resentment. *Proc Br Acad, 48*, 1–25.
Stueber, K. (2019). Empathy. *The stanford encyclopedia of philosophy* (Fall 2019th ed.). In E. D. Zalta (Ed.). https://plato.stanford.edu/archives/fall2019/entries/empathy/. Accessed 17 Feb 2025.
Talbert, M. (2008). Blame and responsiveness to moral reasons: Are psychopaths blameworthy? *Pacific Philosophical Quarterly, 89*, 516–535.
Tangney, J. P., Stuewig, J., & Mashek, D. J. (2007). Moral emotions and moral behavior. *Annual Review of Psychology, 58*, 345.
Teper, R., Zhong, C.-B., & Inzlicht, M. (2015). How emotions shape moral behavior: Some answers (and questions) for the field of moral psychology. *Social and Personality Psychology Compass, 9*(1), 1–14.
Tretter, M. (2024). Equipping AI-decision-support-systems with emotional capabilities? *Ethical Perspectives. Front Artif Intell., 31*(7), 1398395.
Vargas, M., & Nichols, S. (2007). Psychopaths and moral knowledge. *Philosophy, Psychiatry, and Psychology, 14*(2), 157–178.
Véliz, C. (2021). Moral Zombies: Why Algorithms Are Not Moral Agents. *AI & Society, 36*, 487–497.
Wakefield, J. (2018). Are you scared yet? Meet Norman, the psychopathic AI. *BBC*, 2 June 2018. Available at: https://www.bbc.com/news/technology-44040008. Accessed 14 December 2024
Wallace, R. J. (1996). *Responsibility and the moral sentiments*. Harvard University Press.
Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford University Press.
Ward, T. (2010). Psychopathy and criminal responsibility in historical perspective. In L. Malatesti, J. McMillan (Eds.), *Responsibility and psychopathy: Interfacing law, psychiatry and philosophy* (p. 7). Oxford University Press.
Wooldridge, M. (2021). *The road to conscious machines*: *The story of AI.* Penguin UK.
Zaki, J., & Ochsner, K. N. (2012). The neuroscience of empathy: Progress, pitfalls and promise. *Nature Neuroscience, 15*(5), 675.