Article | Received 28 February 2026; Revised 18 April 2026; Accepted 11 May 2026; Published 10 June 2026
https://doi.org/10.55092/let20260007

Artificial agency and the ethics-law divide in AI governance: a call for a paradigm shift in reconstructing artificial moral agency

Alexander Kriebitz[1,2,*], Ali Hessami[3], Nell Watson[4], Amanda Horzyk[5] and Patricia Shaw[6]

[1] Institute for Orthodox Technology, Ludwig Maximilian University of Munich, Munich, Germany
[2] Institute for Ethics in Artificial Intelligence, Technical University of Munich, Munich, Germany
[3] Vega Systems, London, UK
[4] European Responsible Artificial Intelligence Office (EURAIO), Leuven, Belgium
[5] School of Informatics, University of Edinburgh, Edinburgh, UK
[6] Beyond Reach Consulting Limited, Rotherham, UK

* Correspondence author; E-mail: a.kriebitz@lmu.de.

Highlights:

- Ethics and law constitute different normative orders with different functions. Ethics is concerned with individual moral reasoning, whereas law serves to coordinate human behaviour in institutional and societal contexts. Human rights traditionally mediate between these normative domains by protecting spheres of individual moral autonomy from societal interventions.
- The increasing introduction of AI in society risks collapsing the distinction between ethical and legal reasoning. This is particularly evident, where AI systems substitute human decisions, influence choices made by individuals, and introduce societal aims in areas traditionally reserved for individual deliberation.
- To address this issue, the paper proposes an approach to preserve individual moral judgments in AI by calling for a refined understanding of ethics and law as different normative orders, but also by introducing specific safeguards such as the creation of safe spaces for individual moral deliberation in morally critical settings.

Abstract: Contemporary AI governance typically treats ethics and law as complementary normative domains with different levels of enforcement. This paper argues that this distinction alone is insufficient given the evolving role of AI systems as artificial moral agents. Conceptualizing ethics and law as functionally distinct yet interdependent domains, the authors argue that the notion of functional agency to describe how AI systems generate normatively relevant outcomes through decision substitution, the embedding of societal norms in the design of AI systems, and behavioural steering. Drawing on case studies of recommender systems, large language models, autonomous driving, and care robotics, the paper demonstrates the systematic displacement of micro-level ethical reasoning by standardized, societal logics that tend to prioritize macro-level considerations such as beneficence over individual

autonomy. This shift poses a structural challenge to human rights, which explicitly protect individual moral deliberation. The paper therefore calls for a paradigm shift in AI governance from embedding societal norms in AI systems to governing the construction of normativity itself grounded with an emphasis on the protection of individual ethical reasoning and stronger acknowledgement of moral pluralism.

Keywords: AI ethics; moral philosophy; artificial intelligence; human rights; artificial moral agency

1. Introduction

Artificial intelligence (AI) systems increasingly shape the conditions under which human decisions, social interactions, and collective practices take place [1]. Rather than merely supporting human choice, contemporary AI systems structure options, prioritize values, and mediate judgments across domains such as mobility, care, content governance, and public security [2]. These developments mark an important shift from AI as an informational tool to AI as an "artificial moral agent" [3,4]. As AI systems increasingly affect domains traditionally associated with individual moral choices, they raise questions related to the preservation of genuine moral autonomy. In particular, the growing influence of AI in high-stakes decisions such as lethal autonomous weapons systems or the use of AI in recruiting raises the question, whether and how to embed and artificially reconstruct moral reasoning in situations when artificial agents make morally relevant decisions—a phenomenon commonly referred to as "artificial moral agency [3]."

Contemporary approaches to AI governance have largely addressed the increasing influence of AI on normative questions through two instruments: ethics and legal regulation [5]. Ethical design principles are widely invoked to guide responsible AI development, while parallel efforts seek to establish legally binding rules to align AI with societal values [6]. These approaches are often implicitly framed as complementary, with ethics functioning as "soft" or voluntary guidance and law as an enforceable framework [7]. However, this distinction remains conceptually underdeveloped. Even influential initiatives, such as the United Nations Educational, Scientific and Cultural Organization's (UNESCO) Recommendation on the Ethics of Artificial Intelligence or the EU Ethics Guidelines for Trustworthy AI, rely on implicit rather than systematically articulated accounts of the relationship between ethical and legal normativity [8–10].

This paper argues that (pre-)agentic and conversational AI systematically collapses the historically established conventional differentiation between individual ethics and collective legal normativity. AI systems that provide guidance on intimate relationships, existential questions, or personal life choices illustrate a broader development: across heterogeneous use cases including recommender systems, automated driving, large language models, and robotics, AI systems increasingly generate outcomes that shapethe emergence of social norms and relatedly actual human behaviour. This development leads to increasingly blurred lines between ethics and law, with an increasing shift towards integrating broader societal preferences and expectations in socio-technical designs, some of which operate in contexts that pertain to the formation of human values, beliefs and thoughts.

Given the stakes involved in this conversation, the paper makes the case that ethics and law are structurally distinct yet interdependent normative domains, which are increasingly destabilized by the functional and delegated agency of AI [1,2]. The paper develops this argument along the following lines:

- Ethics and law are functionally differentiated yet mutually equilibrating normative domains in the context of AI governance. Ethics mainly seeks to address the question of right and wrong centered on the purpose of life or human flourishing and thus a more personalized perspective on normativity. Law, on the contrary, seeks to coordinate and incentivize human behavior from a societal (macro-normative) perspective.
- Metanormative theories have sought to reconcile both pillars of normativity by developing individual rights as protections for societal overreach. Particularly, the right of freedom of thought, personality rights and freedom of conscience serve as structural limits on the expansion of systemic or collective logics into spheres of individual moral deliberation.
- Artificial agents unsettle this differentiation by integrating collective, macro-level norms into contexts of individual ethical reflection. This has strong implications for individual moral autonomy, as AI agents increasingly make autonomous decisions, but also as they already influence behavior, particularly in the contexts of conversational and interactive AI.
- The increasing use of AI systems in personal domains demands a paradigm shift in reconstructing artificial moral agency in such spaces and requires measures to preserve individuals' moral agency. Such a shift is necessary to prevent the gradual normalization of practices that could undermine constitutionally protected rights. It is also important to prevent wider precedent effects on the interpretation of human rights as such, particularly when it comes to the derogation between societal benefit and individual autonomy.

The remainder of the paper proceeds as follows: First, it reconstructs the philosophical foundations of normative orders underlying the distinction between ethics and law. Subsequently, it analyses how artificial 'agency' disrupts this structure through a series of case studies that illustrate a common normative pattern across different technological domains. Finally, it reflects on the implications of this disruption for AI governance, arguing for an anthropologically grounded and human rights-based framework that preserves the functional differentiation between ethics and law whilst accounting for AI-mediated forms of co-agency.

2. Ethics and law as distinct normative domains

The relationship between law and ethics is essential to the argument that AI disrupts or at least challenges their historically grounded functional differentiation. To expose the gravity of this shift, the following section elaborates on the shared anthropological and epistemic foundations of ethics and law, while also elaborating on their divergent logics of norm creation, justification, and authority. Further, the section discusses the implications of the balance between ethics and law from the perspective of human rights.

2.1. Common normative foundations

Both ethics and legal regulation are fundamentally concerned with norms, setting them apart from empirical inquiry, which focuses on what is rather than what ought to be [11]. Their vocabulary overlaps extensively, employing shared distinctions such as should, ought, and must, permitted and prohibited, desirable and undesirable, right and wrong. They also share underlying metaethical questions, in respect to the existence of universal moral truths or the epistemic justification of normative claims and judgments through means such as reason, common sense, conscience, or sense of justice [12,13].

In addition, both domains are concerned with human agency. Normative theories generally presuppose individuals who can act responsibly, can be held accountable for their actions, and are able to engage in moral deliberation [13]. These assumptions shape not only the articulation of norms, but also lay the foundation of mediating principles, such as responsibility, accountability, complicity, or culpability, which tie human agency to the realization of norms, or the failure to do so [13]. In a similar vein, ethics and law share a common goal: fostering mutual obligations and cooperative behavior within communities, ensuring that individuals' actions are oriented toward the common good. In this way, both systems contain a set of shared norms, for example, the prohibition of murder or the positive regard of solidarity, that guide individual behavior.

Structurally, this reveals deep commonalities in their underlying anthropological assumptions such as the assumption of free will and ontological questions, for instance whether universal moral truths exist independently of the observer [14].

2.2. Divergent logics and norm embedding

Despite these parallels, ethics and law represent different normative orders. The structural differences between them arise from how they encourage or enforce certain behaviours, how they justify and construct norms, and reconcile tensions between norms within their internal structure [15].

Ethics is primarily grounded in individual moral deliberation rather than external incentivization [7]. Ethical norms typically emerge from cultural traditions, personal convictions, religious practices, or philosophical reflection, and adherence is motivated by intrinsic reasons rather than formal sanctions. By contrast, law relies on institutionalized enforcement mechanisms. Legal frameworks are explicitly designed to shape behavior through external incentives, including sanctions and rewards, in alignment with societal objectives [15].

Ethics encompasses a wide plurality of normative traditions such as deontological, utilitarian, virtue-ethical, and faith-based approaches, some of which have historically been hesitant toward codification [15–18]. Ethical theories on human agency are therefore not only more deeply connected to fundamental beliefs, on the very structure of the world but also tend to prescribe a wider area of desirable human conduct, as they offer a holistic theory of human agency [17]. Thus, ethics as a domain of normative reasoning is pluralistic and, in practice, highly personal. By contrast, legal systems aim for a holistic normative theory defined by internal consistency and are primarily concerned with observable behavior. This aspiration towards consistency is reflected in key legal doctrines such as proportionality, which provides a structured method for balancing competing rights and principles while preserving the integrity of the legal system. Through such doctrinal tools, law seeks to mediate normative conflicts in a way that maintains coherence rather than pluralism and, in the case of rights, by circumscribing individual discretion within defined limits. Accordingly, a key distinction between ethics and law lies in the heterogeneity of ethical normativity as opposed to the comparatively homogeneous character of law as an internally coherent system.

A further distinction lies in norm construction and justification. Ethics is primarily internal and agent-centered, focusing on the alignment of agency with internalized moral convictions [16,18]. This explains why many ethical traditions, such as virtue ethics, Kantian ethics, or religious moral frameworks in Mithraism, Christianity, or Buddhism de-emphasize consequentialist considerations and incentives alike [19,20]. By contrast, legal norm construction is fundamentally collective and

institutional, shaped through political processes and oriented towards social coordination and consensus. Law must explicitly consider incentives, causal chains, and predictable effects, and expect compliance with norms, focusing on whether concrete acts align with legal norms. In this sense, ethics can be characterized as autonomous, whereas law is a heteronomous process of norm formation. Thus, the different ways of norm construction result in slightly differentiated sets of norms. Norms emerging from individual contexts (micro-level) are often relational, for instance personal trust, friendship reciprocity, confidentiality or loyalty, whereas systemic norms (macro-level) revolve around institutional stability such as strict compliance with legal norms or societal goals such as poverty reduction or sustainability [11].

Taken together, the main conceptual differences between ethics and law manifest in their nature of operational logic (deliberation *vs.* incentivization), their internal structures (heterogeneity *vs.* homogeneity) and their respective points of departure (individual *vs.* collective).

2.3. Integrative and meta-normative theories

Ethics and law exert mutual influence on one another, as both are concerned with human agency [11]. Individuals and groups shape moral sentiments, which in turn inform legal developments, and vice versa. While the described differences between ethics and law cannot always be sharply delineated in practice, their distinction is normatively relevant and the result of a historical development which positions inalienable rights as a means to protect individual moral reasoning from societal pressure [21–24].

Conflicts between societal reasoning and individual moral reasoning become salient in cases where general legal rules produce outcomes that appear unjust in individual circumstances, or when legal obligations conflict with personal ethical values. A prominent example is the tension between mandatory military service and pacifist convictions, where individuals may regard compliance with the law as morally impermissible. The appearance of conflicts between personal ethics and societal norms explains the relevance of integrative normative "meta-theories".

Historically, the alignment of ethics and law served the collective "common good," but the rise of modern, heterogeneous societies necessitated a shift toward protecting personal choice from collective interference (summum bonum) [25,26].

This separation is codified through theoretical approaches that establish a "barrier against excessive" societal interference in personal moral decisions [26]. A paradigmatic example is John Stuart Mill's harm principle, whereby one individual's liberty ends where it infringes upon another's rights [27]. Such normative theories articulate strong boundaries of the use of legal incentive structures and have been directed against excessive societal interventions into domains shaping individual personality development [20,26,27].

Within this context, human rights serve as the most commonly accepted framework, as they impose obligations on states under international law. In particular, rights to freedom of thought, belief, and conscience protect an inviolable inner sphere of deliberation from external authority, commonly referred to as the forum internum [27]. These rights are closely related to so-called epistemic rights—such as academic freedom, the right to education, and the right to unrestricted access to information—which are essential for the meaningful exercise of informed autonomy [28]. They also intersect with rights to protest, civil disobedience, and personality rights [29,30]. Comparable protected spheres can further be found in procedural and relational safeguards, such as the privilege against self-incrimination, the

right to remain silent, and the legal protection of confidential relationships, notably between physicians and patients or lawyers and their clients [29,30].

Normative approaches thus emphasize meta-principles such as freedom of the press, religious tolerance, and political neutrality in the design of societal institutions and arrangements [27]. At their core lies a protective function that sharply distinguishes liberal systems from totalitarian ones, in which universal truths are ideologically predefined and imposed across both moral and legal domains [31].

Yet by the same token, this normative equilibrium is inherently fragile, given strong societal pressures that invariably collide with rights-protected individual spaces. Crises such as wars, epidemics, or other states of emergency can prompt restrictions on individual rights in favour of collective considerations [32]. In many instances, technologies, particularly surveillance systems, have played a central role in enabling intrusive forms of political control, as illustrated by the extensive monitoring practices in the former German Democratic Republic [31,33]. Such experiences have given rise to the doctrinal expansion of human rights, including the emergence of the right to data protection or the emergence of personality rights that seek to reinforce the protection of the forum internum in the face of new technologies [29].

Thus, the structural separation of ethics and law as separate normative orders is a fundamental necessity derived from human rights and the underlying assumptions on human agency, metaphysical freedom and moral pluralism.

3. Artificial moral agency as functional agency

Historical experience demonstrates that technologies have long influenced the balance between ethics and law, for instance by enabling stronger enforcement of collective norms through surveillance or the use of polygraphs [33]. The emergence of artificial moral agency constitutes a novel challenge to the conventional equilibrium between both normative orders. The following section addresses how this new phenomenon fundamentally challenges the existing separation between law and ethics.

3.1. The concept of artificial moral agency

Ethics and law have long presupposed free will, personhood, and consciousness as preconditions for moral agency [34]. The Universal Declaration of Human Rights itself reinforces this view, as it positionsfreedom, equality, reason and conscience as inherent human qualities [14]. This human-centered paradigm is increasingly challenged by technologies that operationalize, simulate, and in some cases, mimic human agency.

At its core, AI aims to replicate certain human cognitive faculties, performing tasks historically reserved for human agents: interpreting complex data, making predictive judgments, navigating physical environments, executing financial transactions, and mediating social interactions [35–37].

Recent developments in advanced AI systems further accelerate the transfer of decision-making processes from humans to machines [36]. As a result, AI systems increasingly operate in domains that are highly relevant to the relationship between ethics and law, as decisions made or influenced by AI have normative implications. This phenomenon, in which AI systems act as decision-making entities in morally relevant contexts, is commonly defined as artificial moral agency [37,38].

3.2. Artificial moral agency as functional agency and its implications for ethics and law

Functional agency as used in this paper refers to the capacity of AI systems to produce normatively relevant effects in the world without possessing moral subjectivity or intentionality [37]. While such systems operate without real-time human control, their behaviour remains structured by prior human decisions, including design choices, training data, and deployment contexts [28]. What distinguishes functional agency from traditional tools is the system's ability to shape outcomes, structure decisions, and mediate human behaviour in ways that carry normative significance [37].

This form of agency has three analytically distinct but interrelated implications for the relationship between ethics and law: the heteronomization, homogenization, and incentivization of moral reasoning:

- The 'heteronomization' of morality: AI systems are already assuming roles that traditionally required human moral discernment, such as medical diagnostics or autonomous driving. However, these systems do not replace human judgment with a singular "digital mind"; instead, they replace it with distributed agency. The resulting outputs reflect a statistically mediated aggregation of multiple human inputs, forming a type of agency that cannot be attributed to any single individual, but rather a complicated interplay of different co-agencies [33,38]. By mediating human interaction through statistical models, AI extends the causal chain across time and individuals—and in many instances across different cultural areas and value system. A developer's explicit or implict normative assumptions, for instance on the interpretation of medical confidentiality encoded in an AI system, or behavioral assumptions on human behaviour, for instance in job interviews, can create downstream effects on the performance of a system years later [38]. This temporal and spatial extension of human agency complicates traditional accountability models, as it replaces human agency through statistically mediated forms of collective agency, but creates situations where individuals make normative decisions on behalf of others [38].
- The homogenization of morality: AI systems embed norms directly into their design architectures through content moderation rules, safety constraints, and ethical guardrails. Approaches such as value-based engineering and ethics-by-design attempt to formalize moral theories such as deontology or consequentialism within computational models [38]. Computational models of established moral theories, such as Kantian deontology or consequentialism, have been proposed within value-based engineering and ethics-by-design approaches [39]. Yet human moral intuitions and contextual ethical judgments often defy formalization, making flexible, context-aware ethical reasoning difficult to implement, particularly in cases of societal moral dissent [28,40,41]. Implementation is further complicated by human bias, limited representativity of data, and uncertainty in the interpretation of legal norms and moral values, which limits machine learning approaches such as reinforcement learning from human feedback for reconstructing artificial moral behavior [40]. As a result, the realization of artificial morality likely gravitates towards pre-programmed rule-based and, thus, more homogeneous approaches to normative questions. This conflicts with the cultural dimensions of morality, which are highly contextual [28].
- The incentivization of morality: AI systems exert indirect and sometimes even unintended influence over human conduct through algorithmic curation, nudging, and architecture of human machine interfaces. Moreover, the psychological dimension of human–machine interaction often

leads users to attribute agency and moral authority to AI, especially when systems exhibit anthropomorphic or socially interactive traits [42]. Users may seek justification from AI systems after following their advice, while professionals may rely on algorithmic outputs in high-stakes situations, potentially underestimating their epistemic limitations. Consequently, AI systems do not merely produce morally relevant outcomes but can also reshape the moral decision-making processes of individuals [42].

Together, these developments illustrate a profound shift toward forms of normative reasoning that increasingly approximate legal rationality, thereby placing structural pressure on the distinction between ethics and law and progressively displacing individual moral deliberation. This implies that by design artificial moral agents reproduce moral reasoning which is closer to legal normative reasoning in its function of coordinating human behavior in society [11].

3.3. The human rights relevance of functional moral agency

In addition to its impact on the distinction between law and ethics, the functional agency of AI has far-reaching implications for human rights, particularly insofar as these rights protect domains of individual moral reflection and shield them from undue and biased external influences. Rather than directly violating rights through coercion, for instance through traditional means of censorship, AI systems tend to reconfigure the conditions under which rights such as autonomy, conscience, and responsibility can be meaningfully exercised.

- Displacement of individual decisions and the right to conscience: As AI systems shift the locus of ethical deliberation from the individual to pre-programmed collective standards, the use of AI systems in situations with high moral stakes can undermine the right to exercise discretion and the "right to conscience", particularly if decision-making is fully automated [42]. The traditional conversation in AI ethics focuses on those subjected to AI-driven decisions and the lack of meaningful contestation due to algorithmic opacity and procedural abstraction [43,44]. What is at stake is not merely the outcome of a decision and its effect on the "moral patient", but the capacity of the moral agent to refuse, hesitate, or deviate on moral grounds [45,46]. The result is a form of "responsibility without agency", particularly acute in high-stakes and irreversible contexts for instance in military settings or other high-stakes situations [47,48]. This implies that human operators are increasingly put in positions where they might witness morally relevant decisions but are effectively unable to intervene in operations with morally questionable outcomes [48].
- Artificial norm construction and collective preferences: AI systems often encode rigid decision structures that resemble legal norms rather than ethical guidance, which leads to the homogenization of morality [36]. As a consequence, private technological infrastructures increasingly function as de facto normative regimes, exercising regulatory power in domains that have historically been governed by personal judgment or informal social norms [42]. This is particularly evident in machine learning systems, where normative patterns are inferred from empirical behaviour. Here, morality is not explicitly codified but statistically reconstructed, rendering the underlying value trade-offs opaque and resistant to scrutiny [44]. In both rule-based and data-driven approaches, normativity is detached from the moral preferences of those

directly affected by the system, but it is also standardized across different use cases [40,41]. A particular example here is the embedding of moral preferences in autonomous driving.

- Behavioral steering and the erosion of human autonomy: A further challenge arises from the capacity of AI systems to shape human behavior indirectly [49]. Through mechanisms such as nudging, default settings, personalization, and engagement optimization, AI influences the cognitive and attentional conditions under which decisions are made [42]. Unlike overt coercion, these forms of influence operate below the threshold of explicit constraint and therefore often evade traditional legal safeguards [49]. However, informed agency requires the preservation of the internal conditions of deliberation, including access to unbiased and full information. This explains why epistemic rights, such as the right to education or access to information, are enabling conditions for freedom of thought and conscience [28]. Behavioral steering technologies can systematically erode these preconditions by pre-structuring the horizon of available options, weighting, and visibility, thereby challenging the very foundations of informed decision-making, for instance in the case of large language models (LLMs) or recommender systems [50–52].

The cumulative effect of these developments suggests a systematic narrowing of the normative space reserved for individual ethical agency. Due to its functional character, AI introduces both a temporal and a personal disentanglement between those who design or indirectly influence the pathways of artificial moral agency and those whose agency is replaced. This depersonalization of moral agency raises three sets of questions, particularly if these systems operate in areas that are traditionally part of individual moral reasoning:

- Whose agency does AI replace? Who ultimately determines the course of AI decisions?
- Which set of norms—macro-level societal standards or micro-level individual judgments—are embedded in AI systems and how are they balanced in case of conflict?
- How do human-machine interactions shape normative perceptions and decisions?

Taken together, the functional agency of AI characterized by decision substitution, behavioural steering, and the encoding of societal norms exerts continuous structural pressure on the distinction between ethics and law. In doing so, the very concept of artificial moral agency actively undermines the human rights designed to shield individual moral deliberation from collective intervention.

4. Case studies: how AI reconfigures the ethics–law divide

To investigate these dynamics in greater depth, we examine four use cases of AI: recommender systems in social media, autonomous driving technologies, LLMs and care robots. These cases allow us to trace how artificial intelligence reconfigures the equilibrium between ethics and law and provide insight into how existing AI governance frameworks attempt or do not attempt to respond to these transformations.

Each case reflects at least one of the factors that contribute to the depersonalization of moral agency. Recommender systems (4.1) illustrate behavioral steering at a relative distance from the forum internum, shaping the informational conditions of moral reasoning. LLMs (4.2) mediate norms within intimate deliberative contexts and strongly influence human behavior. Autonomous driving (4.3) exemplifies decision substitution in physically embedded but comparatively external contexts, where ethical trade-offs and legal norms are pre-defined in systems' design. Finally, care robots (4.4), combine these forms of functional agency in embodied and emotionally salient settings, placing them in particular proximity to morally critical decisions.

4.1. Recommender systems in social media

Even pre-agentic AI systems, such as recommender algorithms on social media, exemplify how AI operationalizes functional agency and influences human decision-making [49,51]. By curating content based on user behavior and inferred interests, they influence attention, time allocation, and opinion formation. In doing so, they partially replace human judgment in selecting and prioritizing information, raising questions about whose agency is exercised, for instance the agency of users, or that of developers, relevant platform operators and data contributors [49,51,53].

While such systems affect individual agency to a certain extent, for example through pre-selecting videos or news articles, their design is typically oriented toward commercial objectives, such as maximizing engagement, rather than directly advancing users' interests [53,54]. Efforts to introduce ethical safeguards such as steering vulnerable users away from harmful content highlight the difficulty of implementing normative principles in practice. While users may primarily expect reliability, pro-social interventions require balancing privacy, political neutrality, and cultural diversity, with each design choice inevitably shaping users' moral and informational horizons [49]. In order to realize pro-social interventions, someone needs to collect data, which is by definition personal.

Besides, the normative impact of recommender systems manifests at individual and societal levels. Nolens volens, recommender systems shape informational environments. Individually, users are nudged towards specific content, potentially reinforcing biases, fostering echo chambers, or shaping emotional and moral responses [51]. Societally, algorithmic curation interacts with democratic ideals, public discourse, and freedom of expression, effectively embedding corporate and technical norms into the informational environment [49]. Even if such designs happen satisfy the criterion of beneficence, they have an impact on the very process of individual moral reasoning.

This dynamic reinforces debates about transparency and opt-out mechanisms that seek to enable individual human judgment [49]. At the same time, the prevalence of incentive structures challenges the view of AI as a mere extension of delegated human autonomy, suggesting instead that AI ethics often prioritizes the optimization of systemic or collective preferences over the expansion of individual autonomy through AI. In this sense, recommender systems function as decentralized channels of normative influence, where agency emerges from the interaction of developers, data, algorithms, and users, but tends to reflect aggregated societal or commercial preferences rather than the genuine preferences of the individuals involved. Consequently, they tend by design to embody and replicate legal or societal frameworks rather than to model or extend individual agency, which in turn reinforces heteronomous norm construction in fields of opinion formation.

This dynamic has prompted regulatory responses focused on transparency and user control. The Digital Services Act (DSA), for example, introduces obligations for very large online platforms to disclose key parameters of recommender systems and to provide at least one option not based on profiling (Article 38 DSA) [55]. Similar requirements can be found in China's 2022 Provisions on Algorithmic Recommendations [56]. While meant to prevent harm, these frameworks represent important steps directed at the mitigation of behavioural steering in the context of personalized recommender systems. Nevertheless, these changes apply only to individuals that choose to opt out, and not as a default practice. Moreover, it raises questions about the absence of measures to involve affected

individuals within the decision-making process on moral values and norms embedded in recommender systems, particularly as they shape the preconditions for reflected moral judgment [57].

4.2. Large language models

LLMs such as ChatGPT or Co-Pilot illustrate AI's capacity to influence human decision-making and societal norms through language [58]. Unlike task-specific AI, LLMs respond to human prompts across a wide range of contexts, producing statistically plausible responses based on next-token prediction algorithms. Yet despite this versatility, they do not engage in independent moral reflection or possess normative understanding; their responses remain the product of probabilistic pattern recognition rather than ethical deliberation. Their functional agency emerges from massive datasets, developer choices, and user interactions, raising questions about whose agency is exercised and how normative guidance is embedded. As general-purpose models, LLMs can substitute for or mediate multiple forms of human agency—for instance, in drafting texts, maintaining diaries, or providing individualized advice. The roles they assume may range from conversational partner or confidant to legal or medical advisor, or even surrogate patient [59]. Compared to other AI applications, LLMs operate closer to the forum internum, the protected space of individual moral reasoning, potentially interfering with users' intimate, reflective, and relational expressions. This heightened proximity intensifies ethical concerns surrounding trust, dependency, and the internalization of normative influence, and also affects the conditions for the formation of independent thoughts and truly autonomous moral decision-making.

Within this, the differences within norm construction between ethics and legal systems are particularly salient: At the micro level, individuals increasingly rely on LLMs for advice in ethically sensitive or personally consequential situations, such as relationships, career choices, or moral dilemmas [60,61]. While LLMs can provide coherent responses, they lack reflective moral reasoning that takes into account the structural differences in the conceptualization of norms and contextualization in highly specific settings [28]. Personalization may create "ethical comfort zones", reinforcing preferences or biases, and fostering emotional attachment, mirroring the parasocial dynamics raised previously [62]. Users may treat an LLM like a trusted confident and expect loyalty in kind [62].

At the macro level, LLMs shape societal norms by standardizing language, framing discourse, and implicitly promoting certain moral or epistemic patterns [63]. Defaults in system design, such as moderation policies or prompt responses, can therefore operate as normative anchors, and can subtly influence average human behavior, the formation of public discourse, as well as the moral self-perception of individuals, including feelings of guilt, remorse or confidence.

Empirical research increasingly substantiates these concerns. A 2024 Harvard Business Review analysis of real-world ChatGPT usage revealed that a significant proportion of queries involve ethically sensitive domains including relationship advice, career decisions, and personal moral dilemmas, which are situated within highly personal moral decisions [62]. Studies on "secret use" of LLMs demonstrate that users often consult these systems on matters they would not disclose to human confidants, suggesting that LLMs have assumed a quasi-confessional role without the normative framework that traditionally accompanies such relationships [63]. As a result, highly anthropomorphic designs raise questions in respect to procedural rights: The conflict becomes particularly salient in cases involving potential self-incrimination, for instance when chat protocols are used for evidence collection. Furthermore, experimental evidence indicates that LLM-generated moral advice can shift users' ethical

judgments, with effects persisting beyond the immediate interaction [60]. These findings move the concern from theoretical speculation to empirically documented influence.

The behavioral dimension of LLMs has already emerged as a concern for regulators. Under the EU AI Act, providers and deployers of certain AI systems must ensure that individuals are clearly informed, in an obvious manner, when they are interacting directly with an AI system; with an accompanying code of practice for general purpose AI models (published in July 2025). However, neither framework explicitly addresses the norm-encoding function of LLMs in personal ethical contexts and also the limits of pro-social interventions in AI system designs [55].

The absence of preventive measures, protecting individual moral decision-making, raises important questions about responsibility, influence, and the appropriate integration of moral frameworks:

- How should LLMs balance guidance in ethically sensitive situations with respect for individual autonomy?
- Which moral principles should shape outputs, and how do they interact with diverse cultural and societal norms?
- How do personalized responses affect macro-level social commitments to pluralism and democratic deliberation?

From a human rights perspective, LLMs exemplify here the dual role of conversational AI systems as both individual companions that create user expectations in respect to confidentiality, honesty and loyalty, but also their role as functional agents of their deployers and partly legislators. This development highlights more general challenges of embedding ethics and law into systems that simultaneously mediate personal and collective decision-making [60,61]. Future developments such as Moltbook further accentuate this ambivalence, particularly as agentic AI systems are likely to become deeply integrated into LLM-based infrastructures, where they not only assist individual users but also structure communicative environments, coordinate interactions, and influence collective decision-making processes [64]. In this context, the combination of behavioral steering and norm embedding raises the stakes considerably, especially since machine learning-based approaches are based on heteronomous form of norm construction.

4.3. Autonomous driving

Autonomous vehicles (AVs) illustrate how AI moves from abstract decision-making into the physical world, directly interacting with humans and other agents. Driving is governed by norms such as traffic laws, conventions, and social expectations, which become especially visible as machines progressively assume tasks once performed by human drivers [65].

From the perspective of functional agency, AVs are particularly significant because they present a strong case for decision substitution in real-time, safety-critical contexts. AVs must navigate distributions of risk, including situations involving unavoidable harm, which necessitates the pre-programming of normative principles to guide action (e.g., trade-offs between prioritizing passenger versus pedestrian safety). At the same time, however, many aspects of driving have traditionally relied on individual moral agency. Examples include discretionary decisions such as choosing an appropriate speed on a highway without a speed limit, adapting to different driving styles, maintaining context-sensitive distances between vehicles, and engaging in countless micro-interactions. These include, for instance, yielding to other vehicles or allowing an elderly pedestrian to cross the road [66].

Autonomous driving systems replace these forms of human discretion by shifting decision-making into software architecture. In this sense, agency is redistributed from individual drivers to a network of developers, corporations, and regulators who define the permissible action space in advance. Particularly, the last years have seen a stronger shift of the realization of societal considerations in the design of AI ethics. The German Ethics Commission on Automated Driving exemplifies this tendency by translating abstract ethical principles into concrete programming requirements prior to the enactment of binding legislation [67]. The empirical literature on ethical trajectory planning provides further evidence of this normative pre-structuring. Geisslinger *et al*. [66] demonstrate that trajectory planning algorithms must operationalize ethical principles as quantifiable risk distributions, effectively translating abstract moral commitments into computational constraints. Their work reveals that even small variations in the weighting of competing values-such as passenger safety versus pedestrian protection-produce significantly different driving behaviours, underscoring that these are genuinely normative decisions embedded in code rather than mere technical optimizations. The regulatory landscape also reflects an awareness of this challenge: the United Nations Economic Commission for Europe (UNECE) Regulation No. 157 on Automated Lane Keeping Systems establishes binding safety requirements that implicitly encode ethical priorities [68]. Similar tendencies towards the embedding of societal optimization can be observed in the growing reliance on risk-based and utilitarian ethical approaches [67]. While such considerations, particularly those focused on harm reduction, are not inherently problematic, they nonetheless indicate a reallocation of moral agency from individual judgment to aggregated societal preferences. This shift reinforces the question of whose agency is actually being expanded, affected or augmented? Does an automated school bus primarily enhance the agency of the school (punctual attendance), that of children and parents (safety), or that of collective administrative and societal interests?

Autonomous driving brings the tension within the construction of norms into particularly sharp focus, including at the level of AI ethics frameworks themselves [69]. This tension is intensified by diverging moral preferences regarding the resolution of ethical dilemmas in the context of autonomous vehicles, with concrete implications for how risks are managed in system design [69]. Decisions about issues such as trajectory planning do not merely reflect technical optimization, but encode normative judgments about acceptable risk distribution between individuals, passengers, and the public [70]. AVs, therefore, function as hybrid normative systems that simultaneously encode ethics and law while reshaping the conditions under which human moral agency operates.

4.4. Care robots

Care robots exemplify a class of AI systems that directly inhabit physical and social spaces, in contrast to large language models, whose influence is primarily communicative, and in contrast to autonomous vehicles, which typically lack a comparable social dimension [70–72]. Through their embodied presence characterised bymovement patterns, gestures, voice, and anthropomorphic design features such systems trigger strong social and emotional engagement of users. These effects encourage users to attribute moral significance, intentionality, and even agency to machines [73,74]. This, in turn, raises fundamental questions about the legitimacy and limits of projecting human moral expectations onto robots, as well as about who is entitled to define and govern their operational norms.

As a hypothetical stress-test, consider a care robot that adopts a strong confidentiality default analogous to medical ethics: it withholds sensitive disclosures unless a narrow “public interest/serious

harm" threshold is met. Real-world professional and public-interest frameworks typically treat confidentiality as stringent but not absolute, permitting proportionate disclosure to prevent death or serious harm in exceptional cases. In system design terms, this tension can be reduced (not eliminated) via sealed audit trails, least-privilege disclosure, data minimization/on-device processing, and a consent-based "break-glass" pathway with post-hoc review [70].

In healthcare, elder care, and domestic settings, care robots are expected to perform tasks while simulating empathy, moral patiency and agency, and simultaneously respect norms rooted in human rights and cultural traditions [72]. Yet robots lack consciousness or genuine moral reasoning; their actions are guided by rules, data, and programming, approximating legal or procedural logic rather than reflective ethical judgment [31].

The artificial reconstruction of morality gives rise to a tension between individual and societal perspectives on normativity, particularly as care robots implicitly or explicitly encounter significant moral questions. At the micro level, individuals anticipate context-sensitive, morally nuanced responses, including but not limited to intervention in conflict, reminders about care routines, or guidance in ethical and personal situations. At the macro level, legal and societal frameworks require consistency, transparency, and accountability, limiting the scope for flexible moral discretion. As in the case of LLMs, highly personalized and anthropomorphic designs may enhance intimacy and trust, while simultaneously risking the emergence of "comfort zones" that shape inappropriate moral expectations of artificial moral agents, particularly when individuals come to treat them as confidants or share legally relevant information [73,74].

Care robots thus operate as distributed normative agents, where functional agency emerges from designers, programmers, and embedded ethical rules, while human perception projects moral weight onto their actions [74]. This duality prompts broader questions:

- Should robots be designed to shape human behavior, for instance to prevent suicides?
- How should intervention thresholds be set? And to what extent is attributing moral agency to machines compatible with human-centered law and ethics?

By embodying norms physically and socially, care robots highlight the ethical and regulatory challenges of embedding morality into AI. This issue is magnified in direct interactions with individuals in vulnerable conditions, such as older persons, medically fragile patients, or individuals with cognitive or social dependencies, illustrating the persistent tension between individual expectations, particularly in respect to confidentiality and societal standards that rather follow an incentivized logic, but also reinforcing the normative implications of artificially created agency, within different normative contexts [75].

Regulatory practice in the care robotics domain remains fragmented but evolving. The EU Medical Device Regulation (MDR 2017/745) provides a framework for robot-assisted care devices, yet its primary focus on safety and efficacy leaves normative design choices largely unaddressed [76]. Japan's Ministry of Economy, Trade and Industry has issued guidelines for robotic care devices that address physical safety but do not engage with the ethical dimensions of anthropomorphic norm projection [77]. While the aforementioned frameworks clarify many important questions surrounding safety, none of these frameworks explicitly addresses the distinctive challenge that care robots pose for the ethics-law divide: their combination of embodied presence, anthropomorphic design, and intimate deployment context means that all three forms of functional agency-decision substitution, behavioural shaping, and norm encoding-operate simultaneously and in direct proximity to the user's forum internum. This

convergence makes care robots a particularly acute test case for the adequacy of governance frameworks that fail to theorize the ethics-law boundary.

Across these four cases, a consistent structural pattern emerges, namely that relatively critical contexts for individual moral reasoning are increasingly affected by the functional agency of AI. Recommender systems demonstrate how behavioural shaping affects the informational preconditions of moral reasoning; large language models show how the combination between norm encoding and behavioral shaping affects highly personal moral decisions; autonomous driving reveals how decision substitution pre-programs ethical trade-offs that were previously the domain of individual judgment; and care robots combine all three forms of functional agency in embodied, emotionally charged settings. Taken together, these cases demonstrate that the erosion of the ethics-law boundary is not confined to a single technology or level of AI sophistication but presents a structural shift across heterogeneous domains. The regulatory responses examined-from the EU Digital Services Act to UNECE safety standards-address specific manifestations of this pattern but have not yet confronted the underlying structural challenge: the systematic displacement of individual moral autonomy by systems that replace or influence human agency based on collectively constructed norms.

5. Implications for global AI governance

The case studies demonstrate that the distinction between ethics and law is not merely theoretical but constitutes a practical challenge for human-centered AI governance, and specifically the alignment of artificial moral agency with human rights. Across diverse domains, a consistent pattern emerges: AI systems increasingly acts as artificial agents that replace individual moral deliberation with collectively aggregated logics. This transformation goes beyond the “enforceability” gap identified by scholars like Floridi [33] and points to a stronger integration of societal logics within areas designated in the forum internum. It further suggests that the substitution of human agency by artificial moral agents entails the introduction of external incentives, technical homogenization, and an increasing heteronomy in areas associated with protections to safeguard independent moral decision making.

The following section therefore examines the broader implication of this observation in respect to the governance of AI, the construction of norms in artificial agents and the interpretation of human rights in the context of artificial moral agency. Ultimately, the paper provides preliminary recommendations for a human rights-based alignment of artificial agents with societal norms and further recommendations for the conversation surrounding agency, norms, and decision making in AI ethics.

5.1. Functional agency and the process of norm construction in artificial moral agents

Taken together, the three mechanisms of functional agency, decision substitution, norm encoding, and behavioral steering reveal a common structural dynamic: the progressive displacement and reconfiguration of individual ethical agency through technologically mediated forms of normativity.

Across all examples, a common feature is the emergence of heteronomous normativity: artificial agents follow logics that differ from the reasoning of conventional human agents and which are shaped by a combination of regulatory frameworks, for instance legislated measures, and in unregulated contexts organizational and economic interests [78]. The distinction lies less in which normative preferences are actually encoded but in the heteronomization of moral decision-making in what were

previously exclusively personal spheres of moral discernment. As this relates to the integration of societal norms in individual moral reasoning, the use of AI reinforces existing conflicts between the protection of individual moral autonomy and the realization of positive societal outcomes, for instance in respect to education, sustainability or crime prevention [79].

This development becomes particularly problematic as these systems increasingly operate in domains traditionally shielded from societal reach. In this sense, the issue of artificial moral agency is not merely a policy vacuum [34] but rather the existence of societal and not just corporate pressures to optimize AI systems for different moral and immoral purposes: thus, AI systems actively thus introduce and often perpetuate opaque forms of norm diffusion.

Current legal approaches, such as those reflected in the EU AI Act or the Digital Services Act, primarily address risks of indirect manipulation, which addresses the issue of behavioural steering [55]. However, they do not adequately engage with the deeper issue of artificial moral agency and its implications for individual autonomy. This is particularly problematic in intimate domains where the substitution of personal ethical reflection with standardized algorithmic logic remains insufficiently addressed, threatening the very foundations of autonomous human agency [78]. To resolve this, AI governance must move beyond behavioural oversight toward a reconstruction of artificial agency that explicitly respects the functional divide between ethics and law.

5.2. Functional agency and the outcomes of norm construction

This structural shift reflects an increasing alignment of AI-mediated normativity with logics traditionally associated with legal systems: behavioural steering (including unintended forms), heteronomous preference formation, and a homogenization of moral reasoning. This development challenges individual moral autonomy as a foundational premise of normativity [80,81].

A useful lens for analysing this tension is the distinction between micro-level and macro-level norms. Micro-level norms arise from individual ethical reflection, relational judgment, and contextual sensitivity, whereas macro-level norms are stabilized through law, institutional design, and collective expectations. Ethics, particularly in care-ethical traditions, emphasizes situated responsiveness and interpersonal responsibility, while legal normativity operates in a systemic, generalizable, and often optimization-oriented manner [80]. The choice for one of these norms sets directly affects how broader principles such as beneficence or fairness are interpreted and operationalized.

This divergence becomes visible in concrete examples. An individual driver may stop for an elderly pedestrian despite minor traffic disruption, or a bus driver may wait for a late child, prioritizing relational responsibility over strict rule compliance. By contrast, macro-level systems such as traffic regulation or algorithmic control in autonomous vehicles prioritize consistency, efficiency, and collective safety. As AI systems require ex ante specification of behaviour, such context-sensitive judgments must be pre-encoded into system design, making these trade-offs normatively relevant prior their actual deployment and use.

Against this background, a central question emerges: is an artificial moral agent the prolonged agency of an individual (e.g., a person operating a vehicle, a student writing an essay, or a patient in a care setting), or the prolonged agency of society, as embedded in institutional, legal, and technical infrastructures? This ambiguity becomes particularly salient in systems that mediate everyday decision-making. Users tend to expect loyalty, reliability, and contextual sensitivity from AI systems functioning as extensions of their agency for example in large language models or care robots. Concurrently, these systems are shaped by

societal logics that prioritize aggregated goals such as safety, harm prevention, or compliance, which may conflict with individual preferences. The construction of normativity in AI systems thus emerges from the interaction between individual expectations and embedded design decisions across systems such as LLMs, recommender systems, and interactive agents.

These boundary questions remain largely underregulated and are often delegated to private actors that tend to optimize systems for commercial purposes [52]. Consequently, AI tends to increasingly embed macro-level norms in personal spaces, raising questions about the very construction of moral agency, particularly in the context of human-machine interactions. This problem is closely tied to emerging conversations on human-centric versus people-centric AI and to the interpretation of human rights as protecting either individual or collective normative spheres [28].

5.3. Human rights approach

Human rights considerations are central to addressing the tensions identified above. The problem contexts described are inherently embedded within a human rights framework, particularly where AI systems affect domains traditionally protected as part of the forum internum. Core rights relevant to this context derive from international instruments such as the International Covenant on Civil and Political Rights, which explicitly guarantee freedom of thought, conscience, and religion (Article 18(1)), as well as protection against arbitrary interference with privacy (Article 17). Comparable protections can be found in regional frameworks, including the African Charter on Human and Peoples' Rights and the European Convention on Human Rights, where Article 8 has been interpreted by the European Court of Human Rights as encompassing personal autonomy and self-determination, including the conditions under which individuals form beliefs and moral judgments [82]. Together, these frameworks reaffirm the protection of an inner sphere of autonomous norm formation.

Beyond their normative relevance, human rights frameworks offer a structural advantage in addressing the challenges posed by AI. They constitute historically evolved systems that have already grappled with the tension between individual ethical autonomy and collectively binding norms. As such, they provide established conceptual tools capable of mediating these tensions across different institutional and technological contexts, making them particularly suited to addressing the cross-domain implications of AI-mediated normativity.

At the same time, the applicability of human rights to AI has increasingly been recognized at both international and supranational levels. Soft law instruments such as the EU High-Level Expert Group guidelines and the UNESCO Recommendation on the Ethics of Artificial Intelligence, as well as binding initiatives like the Council of Europe's Convention on AI, the proposed UN Convention on AI, Data and Human Rights, and the EU AI Act, provide emerging governance structures [28,83,84]. These instruments, however, tend to remain relatively general, as they do not provide yet a fully articulated framework for addressing artificial moral agency yet [84].

A key challenge lies in the fact that the interpretation of these rights in the context of AI remains largely implicit. While adjacent debates exist—most notably around neurorights and cognitive liberty—they primarily focus on doctrinal expansion of data protection and personality rights, rather than directly addressing the normative implications of AI on individuals' ethical decision-making. Similarly, existing regulatory measures, such as the prohibition of social scoring systems or restrictions on deceptive practices, target specific risks, particularly those related to manipulation or political

influence [83]. These provisions constitute important precedents, yet they primarily address indirect forms of influence and do not fully engage with the broader question of how AI systems construct and mediate normativity.

Further guidance can be drawn from human rights frameworks in fields such as medical ethics, where instruments like the Oviedo Convention explicitly emphasize human dignity and patient autonomy [85–87]. These domains illustrate how strong protections of individual autonomy are maintained in contexts involving highly sensitive and irreversible decisions, including similar conversations of the doctrinal expansion of human rights in the context of human-machine interfaces [86,87]. Taken together, existing human rights instruments remain fragmented, but they provide a substantive foundation for a more explicit engagement with artificial moral agency in within the interplay between ethics and law.

From this perspective, a human rights-based approach is particularly valuable in identifying contexts where the reconstruction of societal optimization logics through AI undermines or at least challenges the preservation of an autonomous sphere of moral reasoning. This applies, for instance, to domains with a high premium on individual autonomy, such as medical decision-making or freedom of conscience in military contexts. These considerations suggest inherent limits to the acceptable scope of AI-mediated normativity, particularly in cases involving irreversible consequences or deeply personal ethical judgments

5.4. Design-level principles for meaningful action

The preceding analysis has shown that the very construction of functional agency constitutes a normative challenge, when it is situated in contexts that relate to the expression of opinions, beliefs, and views, but also their actualization in morally critical situations. This shift is driven by the core mechanisms of functional agency identified above—decision substitution, norm encoding, and behavioural steering—which contribute to the displacement of micro-level ethical reasoning by macro-level normative logics. Existing regulatory approaches, such as those reflected in the EU AI Act or the General Data Protection Regulation (GDPR), that tend to focus on indirect forms of behavioural steering [83], do not account for the more direct challenge of preserving individual moral autonomy in AI-mediated environments.

Each of the following principles responds directly to these mechanisms and aims to reintroduce conditions for autonomous, micro-level ethical reasoning within AI-mediated environments. In doing so, they translate the requirement to protect an inner sphere of individual moral deliberation into concrete design constraints.

(1) Graduated Autonomy

AI systems should operate according to differentiated levels of autonomy that reflect the normative weight of the domain in which they are deployed. As shown above, the displacement of ethical judgment has fundamentally different implications depending on whether systems operate in low-stakes environments (e.g., entertainment recommendations) or in contexts involving health, identity, or moral deliberation (e.g., care robots or LLM-based advice). Without such differentiation, AI systems risk applying macro-level optimization logics to contexts that require micro-level ethical judgment. This principle therefore requires a domain-sensitive calibration of autonomy, including stronger oversight, meaningful opt-out mechanisms, and stricter constraints in contexts characterized by irreversibility, vulnerability, or proximity to the forum internum.

(2) Normative Transparency

AI systems should disclose not only how they function technically, but which normative logic they implement—whether they prioritize collective optimization (macro-level norms) or the preservation of individual autonomy (micro-level ethics), and whose values are encoded. As the analysis has shown, many systems present themselves as neutral while embedding specific normative orientations aligned with regulatory or commercial incentives. Without such transparency, users cannot meaningfully distinguish between self-directed and externally imposed normativity, undermining the conditions for autonomous ethical judgment. At the same time, this principle must be balanced against risks of manipulation and proprietary constraints, suggesting the need for differentiated transparency regimes, including regulatory disclosure of normative design choices.

(3) Domains Reserved for Individual Ethics

Certain domains require categorical protection from AI-mediated injection of societal preferences. This principle operationalizes the protection of the forum internum in human rights law by translating it into design-level constraints. Contexts such as end-of-life decisions, political and religious belief formation, or therapeutic interactions are characterized by a primacy of relational and context-sensitive ethical reasoning, which cannot be adequately captured by generalized, system-level norms [61,62]. As illustrated in earlier examples, individual ethical reasoning may justify context-sensitive deviations from general rules such as prioritizing relational responsibility over efficiency whereas AI systems tend to enforce macro-level consistency. In such domains, systems should therefore default to enabling reflection rather than prescribing outcomes. While this may limit certain optimization potentials, such constraints are necessary to preserve the conditions of autonomous norm formation.

(4) Reflection Modes

AI systems operating in high-stakes or intimate contexts should include mechanisms that actively preserve the conditions for autonomous moral deliberation. Building on the analysis of behavioral steering, this principle addresses the often subtle ways in which AI systems shape preferences, attention, and judgment through nudging and optimization. Autonomy depends not only on the availability of options, but on the conditions under which decisions are made. Reflection modes therefore temporarily suspend persuasive or engagement-maximizing functions, creating a protected space for deliberation that functionally mirrors the autonomy guarantees of the forum internum. While such mechanisms may introduce frictions in user experience, these trade-offs reflect a necessary rebalancing between micro-level ethical reasoning and macro-level optimization pressures embedded in AI systems.

Taken together, these principles shift the focus of AI governance from ex post risk mitigation to the ex ante design of normativity itself. They operationalize the insight that AI systems do not merely apply norms but participate in their construction—often at the intersection of individual ethical reasoning and collectively imposed frameworks. Embedding these principles into development processes is therefore essential not only for protecting individual autonomy, but also for ensuring the legitimacy of AI-mediated normativity across domains [65].

The practical realization of this paradigm shift faces significant hurdles, as the inherent requirement to establish “normative red lines” often clashes with broad societal expectations of technological utility. The drive for societal optimization against individual autonomy remains a major area of contention. Furthermore, the doctrinal expansion of personality and self-determination rights to counter the encroachments of artificial moral agency requires not only stronger constitutional safeguards but also a

degree of societal self-restraint. Society must decide whether it is willing to limit the full potential of algorithmic optimization in favor of preserving the "inefficient" but vital spaces of individual moral deliberation [32].

5.5. Implications for AI ethics

Ultimately, the impact of AI on the relationship between ethics and law extends beyond governance questions to the foundations of AI ethics frameworks themselves. If left unaddressed, the continued evolution and development of AI risks reinforcing a structural shift in normative reasoning toward macro-level, rule-based logics, potentially at the expense of individual ethical agency and the rights designed to protect it.

This shift has significant implications for how existing AI ethics frameworks are interpreted and applied. Contemporary approaches, such as the guidelines of the High-Level Expert Group on AI (HLEG) and the AI4People initiative, failed to establish normative theories aligned with human rights frameworks to reconcile conflicts between different AI ethics principles, most importantly between autonomy and beneficence [88].

The differing constructions of agency from individual and societal levels raise the question, whose agency is actually augmented or diminished through AI deployment, particularly when analyzed from role-based perspectives? These distinctions are significant because ethical protections at the individual level operate according to different normative principles than aggregate utilitarian calculations or deontological constraints at the societal level. As a result, strong alignment with collectively constructed norms can generate asymmetries in the protection of the forum internum across both digital and non-digital contexts, while also setting precedents for the interpretation of such norms in other domains of technological application, including law enforcement—potentially at the expense of structurally related rights, such as the right to remain silent.

6. Conclusion

Governing AI systems, particularly AI agents, demands rethinking the relationship between ethics and law. While often framed as complementary, these domains represent distinct yet interdependent normative orders, sharing anthropological foundations like free will and epistemic uncertainty but diverging in purpose and value construction [14,89,90]. Ethics and law therefore approach normative questions according to different logics, with ethics itself being characterized by a high degree of internal heterogeneity.

AI agents unsettle this fragile—yet human-rights-grounded—equilibrium by introducing forms of *functional agency*: they generate adaptive, socially consequential outcomes, while lacking moral subjectivity and legally legitimated agency. In doing so, they reveal a deeper tension within the common space of normativity, namely competing constructions of norms from the individual and from the society's perspective, including different perspectives on the underlying purpose of norms as such.

Value-based approaches often assume consensus on what counts as a relevant value, yet they rarely ask where these values come from. AI systems increasingly mediate between different normative layers, from individual expectations for loyalty and trust as opposed to societal preferences operationalizing values in ways that may privilege one level over the other without explicit justification. This implicit

prioritization risks transforming ethical pluralism into algorithmic uniformity, but also further raises the question of whose agency AI truly seeks to expand and who gets to meaningfully participate in these processes.

In such contexts, safeguarding autonomy requires more than transparency or consent; it demands clearly articulated protection zones for individuals interacting with AI systems that are situated within spaces protected by personality rights.

The purpose of this paper has been to expose the gap in current normative debates on AI regulation, namely the insufficient consideration of the conceptual relationship between law and ethics. This paper argues that AI, particularly in the form of artificial agents, systematically collapse the historically established functional differentiation between individual ethics and collective legal normativity, thereby threatening the human-rights-based architecture that protects private moral autonomy. A central problem is that this relationship is usually treated implicitly. For the reasons outlined above, however, it is essential to make the law–ethics relationship explicit. This paper advances the debate by reframing ethics–law differentiation through an anthropologically grounded, human-rights–based lens. It argues for a paradigm shift in which the questions of whose actions are being replaced and how such replacement is situated within a normative tradition are jointly and systematically addressed.

While many of the points raised in the paper are preliminary in their nature, the conceptual distinction between ethics and law as separate yet interdependent expressions of normativity has significant implications across AI governance, ethics-by-design methodologies, neurorights, technical standardization, and the interpretation of legal definitions—particularly regarding prohibited practices and impact assessments [86,91–93]. In addition, the distinction between law and ethics aligns also with existing trends in applied AI ethics to design AI systems that seek to empower human agency, for instance in the contexts of health and privacy [94,95]

Coherent AI governance, therefore, requires sustained attention to: (1) the provenance of values embedded in AI systems; (2) the anthropological assumptions underpinning ethical and legal norms; and (3) the protection of individual normative spaces in contexts of intensified human–machine interaction.

More broadly, there is a pressing need for normative frameworks that not only address the ethics–law bipolarity but also critically examine the implications of this distinction for artificial moral agency. Developing a coherent framework and a more conceptually integrated connection between ethics and law is therefore essential—without further formalizing ethics in a way that undermines its plural and context-sensitive character. A key component of such foundational work lies in clarifying the difficulties and limits of ethical approaches that are primarily grounded in individual perspectives, as well as in demonstrating the necessity of an interdisciplinary methodology that brings together legal scholarship and ethics.

Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors used generative AI tools only to improve language and readability. Specifically, the authors used ChatGPT for language polishing in substantial portions. The authors take full responsibility for the content of the manuscript.

Authors' contribution

Conceptualization, Alexander Kriebitz; methodology, Ali Hessami, Nell Watson, Amanda Horzyk and Patricia Shaw; writing—original draft preparation, Alexander Kriebitz; writing—review and editing, Ali Hessami, Nell Watson, Amanda Horzyk and Patricia Shaw; visualization, Ali Hessami, Nell Watson; supervision, Patricia Shaw; project administration, Amanda Horzyk; funding acquisition, Ali Hessami. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] OECD. Explanatory memorandum on the updated OECD definition of an AI system. 2024. Available: https://www.oecd.org/en/publications/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_623da898-en.html (accessed on 25 May 2026).

[2] Bandi A, Kongari B, Naguru R, Pasnoor S, Vilipala SV. The rise of agentic AI: a review of definitions, frameworks, architectures, applications, evaluation metrics, and challenges. *Future Internet* 2025, 17(9):404.

[3] Behdadi D, Munthe C. A normative approach to artificial moral agency. *Minds Mach.* 2020, 30(2):195–218.

[4] Schmelzer R. Rentahuman.ai turns humans into on-demand labor for AI agents. 2026. Available: https://www.forbes.com/sites/ronschmelzer/2026/02/05/when-ai-agents-start-hiring-humans-rentahumanai-turns-the-tables/ (accessed on 25 May 2026).

[5] Carrillo MR. Artificial intelligence: from ethics to law. *Telecommun. Policy* 2020, 44(6):101937.

[6] Khamassi M, Nahon M, Chatila R. Strong and weak alignment of large language models with human values. *Sci. Rep.* 2024, 14(1):19399.

[7] Floridi L, Taddeo M. Moral *vs* legal norms: soft and hard ethics. In *A Companion to Digital Ethics*. Hoboken: Wiley-Blackwell, 2025. pp. 11–23.

[8] Ganbaatar U. Do ethics in AI still matter? A review of the 2021 UNESCO recommendation on the Ethics of AI. *Rev. Faith Int. Aff.* 2025, 23(3):26–33.

[9] Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, *et al*. AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Mines Mach.* 2018, 28(4):689–707.

[10] Smuha NA. The work of the high-level expert group on AI as the precursor of the AI Act. *SSRN* 2025, SSRN 5012626.

[11] Wendel WB. Legal ethics and the separation of law and morals. *Cornell Law Rev.* 2005, 91:67–128.

[12] MacIntyre AC. Hume on 'is' and 'ought'. In *The Is-Ought Question*. London: Palgrave Macmillan, 1969. pp. 35–50.

[13] Rowe F, Jeanneret Medina M, Journé B, Coätard E, Myers M. Understanding responsibility under uncertainty: a critical and scoping review of autonomous driving systems. *J. Inf. Technol.* 2024, 39(3):587–615.

[14] Hughes G. The concept of dignity in the universal declaration of human rights. *J. Relig. Ethics* 2011, 39(1):1–24.
[15] Williams B. *Morality: An Introduction to Ethics.* Cambridge: Cambridge University Press, 2012.
[16] Brysk A. Engaged Buddhism as human rights ethos: the constructivist quest for cosmopolitanism. *Hum. Rights Rev.* 2020, 21(1):1–20.
[17] Florovsky G. The ethos of the Orthodox Church. *Ecu. Rev.* 1960, 12(2):183–198.
[18] Kissling F. The place for individual conscience. *J. Med. Ethics* 2001, 27(2):ii24–ii27.
[19] Ulansey D. *The Origins of the Mithraic Mysteries: Cosmology and Salvation in the Ancient World.* Oxford: Oxford University Press, 1991.
[20] Fasoro SA. Kant on human dignity: autonomy, humanity, and human rights. *Kant. J.* 2019, 38(1):81–98.
[21] Pavone IR. The role of soft law in bioethics. In *International Biolaw and Shared Ethical Principles*, 1st ed. Abingdon: Routledge, 2018. pp. 99–118.
[22] Mares R. *The UN Guiding Principles on Business and Human Rights: Foundations and Implementation*. Leiden: Martinus Nijhoff Publishers, 2012.
[23] Rowan JR. Grounding hypernorms: toward a contractarian theory of business ethics. *Econ. Philos.* 1997, 13(1):107–112.
[24] North DC. Institutions, ideology, and economic performance. In *The Revolution in Development Economics*. Oxford: Oxford University Press, 1998. pp. 113–128.
[25] Buksiński T. Metagoods, metavalues and metanorms in politics. *Dial. Univ.* 2017, 2:129–140.
[26] Black D. *Moral Time*, 1st ed. Oxford: Oxford University Press, 2011.
[27] Mill JS. *On Liberty*. London: Longman, Roberts & Green, 1859.
[28] Kriebitz A, Corrigan C, Pevkur A, Ferro AS, Horzyk A, *et al*. Cultural rights and the rights to development in the age of AI: implications for global human rights governance. *arXiv* 2025, arXiv:2512.15786.
[29] Decker DC, Fresa L. The status of conscientious objection under Article 4 of the European Convention on Human Rights. *N. Y. U. J. Int. Law Polit.* 2000, 33:379.
[30] Valero MJ. Freedom of conscience of healthcare professionals and conscientious objection in the European Court of Human Rights. *Religions* 2022, 13(6):558.
[31] Arendt H. *The Origins of Totalitarianism*. Boston: Houghton Mifflin Harcourt, 1973.
[32] Pettit P. The instability of freedom as noninterference: the case of Isaiah Berlin. *Ethics* 2011, 121(4):693–716.
[33] Pfaff S. The limits of coercive surveillance: Social and penal control in the German Democratic Republic. *Punish. Soc.* 2001, 3(3):381–407.
[34] Clement G. Animals and moral agency: the recent debate and its implications. *J. Anim. Ethics* 2013, 3(1):1–14.
[35] Mukhamediev RI, Popova Y, Kuchin Y, Zaitseva E, Kalimoldayev A, *et al*. Review of artificial intelligence and machine learning technologies. *Mathematics* 2022, 10(15):2552.
[36] Burrell J. How the machine ‘thinks’: understanding opacity in machine learning algorithms. *Big Data Soc.* 2016, 3(1):2053951715622512.
[37] Jho H, Park C, Ahn D. Towards a philosophy of ensemble cognition: reconceptualising agency and mind in AI-mediated educational environments. *Educ. Philos. Theory* 2026, pp. 1–20.

[38] Moor JH. Why we need better ethics for emerging technologies. *Ethics Inf. Technol.* 2005, 7(3):111–119.
[39] Berreby F, Bourgne G, Ganascia JG. Modelling moral reasoning and ethical responsibility. In *Logic for Programming, Artificial Intelligence, and Reasoning*. Berlin: Springer, 2015. pp. 532–548.
[40] Allen C, Smit I, Wallach W. Artificial morality: top-down, bottom-up, and hybrid approaches. *Ethics Inf. Technol.* 2005, 7(3):149–155.
[41] Graham J, Haidt J, Nosek BA. Liberals and conservatives rely on different sets of moral foundations. *J. Pers. Soc. Psychol.* 2009, 96(5):1029–1046.
[42] Akbulut C, Weidinger L, Manzini A, Gabriel I, Rieser V. All too human? Mapping and mitigating the risk from anthropomorphic AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, San Jose, USA, October 21–23, 2024, pp. 13–26.
[43] Koeszegi ST. Automated decision systems: Why human autonomy is at stake. In *Collective Decisions: Theory, Algorithms and Decision Support Systems*. Switzerland: Springer, 2021. pp. 155–169.
[44] Bathaee Y. The artificial intelligence black box and the failure of intent and causation. *Harv. J. Law Technol.* 2017, 31:889–938.
[45] MF Major. Conscientious objection and international law: a human right. *Case W. Res. J. Int'l Law* 1992, 24(2):349.
[46] Berber A. Automated decision-making and the problem of evil. *AI Soc.* 2025, 40(2):1049–1058.
[47] Kriebitz A, Lütge C. Artificial intelligence and human rights: a business ethical assessment. *Bus. Hum. Rights J.* 2020, 5(1):84–104.
[48] Baudisch I. Germany v. N. Decision No. 2 WD 12.04. *Am. J. Int'l Law* 2006, 100(4):911–917.
[49] Milano S, Taddeo M, Floridi L. Recommender systems and their ethical challenges. *AI Soc.* 2020, 35(4):957–967.
[50] Meissner G. Artificial intelligence: consciousness and conscience. *AI Soc.* 2020, 35(1):225–235.
[51] Catena E. AI and human autonomy: a literature review. *AI Ethics* 2026, 6(1):126.
[52] Benkler Y. Don't let industry write the rules for AI. *Nature* 2019, 569(7754):161–162.
[53] Horzyk AM. Data protection and privacy: risks and solutions in the contentious era of AI-driven ad tech. In *Proceedings of the 30th International Conference on Neural Information Processing*, Changsha, China, November 20–23, 2023, pp. 352–363.
[54] Bonicalzi S, De Caro M, Giovanola B. Artificial intelligence and autonomy: on the ethical dimension of recommender systems. *Topoi* 2023, 42(3):819–832.
[55] Naudts L, Helberger N, Veale M, Sax M. A right to constructive optimization: a public interest approach to recommender systems in the Digital Services Act. *J. Consum. Policy* 2025, 48(3):269–296.
[56] Yang F, Yao Y. A new regulatory framework for algorithm-powered recommendation services in China. *Nat. Mach. Intell.* 2022, 4(10):802–803.
[57] Djeffal C, Hitrova C, Magrani E. Recommender systems and autonomy: a role for regulation of design, rights, and transparency. *Indian J. Law Technol.* 2021, 17(1):3.
[58] Padiu B, Iacob R, Rebedea T, Dascalu M. To what extent have LLMs reshaped the legal domain so far? A scoping literature review. *Information* 2024, 15(11):662.
[59] Baggot M. The quest for connection in AI companions. *J. Ethics Emerg. Technol.* 2025, 35(1):1–20.
[60] Zhang Z, Shen C, Yao B, Wang D, Li T. Secret use of large language model (LLM). *Proc. ACM. Hum. Comput. Interact.* 2025, 9(2):1–26.

[61] Noor N, Rao Hill S, Troshani I. Artificial intelligence service agents: role of parasocial relationship. *J. Comput. Inf. Syst.* 2022, 62(5):1009–1023.
[62] Zao-Sanders M. How people are really using GenAI. 2024. Available: https://hbr.org/2024/03/how-people-are-really-using-genai (accessed on 25 May 2026).
[63] Krügel S, Ostermaier A, Uhl M. Zombies in the loop? Humans trust untrustworthy AI-advisors for ethical decisions. *Philos. Technol.* 2022, 35:17.
[64] Brohi S, Mastoi Q, Jhanjhi NZ, Pillai TR. A research landscape of agentic ai and large language models: applications, challenges and future directions. *Algorithms* 2025, 18(8):499.
[65] Luetge C. The German ethics code for automated driving. *Philos. Technol.* 2017, 30(4):547–558.
[66] Geisslinger M, Poszler F, Lienkamp M. An ethical trajectory planning algorithm for autonomous vehicles. *Nat. Mach. Intell.* 2023, 5(2):137–144.
[67] Bordum A. Immanuel Kant, Jürgen Habermas and the categorical imperative. *Philos. Soc. Critic.* 2005, 31(7):851–874.
[68] United Nations Economic Commission for Europe. UN Regulation No. 157. 2021. Available: https://unece.org/sites/default/files/2023-12/R157e.pdf (accessed on 25 May 2026).
[69] Gogoll J, Zuber N, Kacianka S, Greger T, Pretschner A, *et al.* Ethics in the software development process: from codes of conduct to ethical deliberation. *Philos. Technol.* 2021, 34(4):1085–1108.
[70] Awad E, Dsouza S, Kim R, Schulz J, Henrich J, *et al.* The moral machine experiment. *Nature* 2018, 563(7729):59–64.
[71] Van Wynsberghe A. Service robots, care ethics, and design. *Ethics Inf. Technol.* 2016, 18(4):311–321.
[72] Vallor S. *The AI Mirror*. Oxford: Oxford University Press, 2024.
[73] Van Wynsberghe A. Designing robots for care: care centered value-sensitive design. *Sci. Eng. Ethics* 2013, 19(2):407–433.
[74] Riek LD, Rabinowitch TC, Chakrabarti B, Robinson P. How anthropomorphism affects empathy toward robots. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, San Diego, USA, March 11–13, 2009, pp. 245–246.
[75] Rosenthal-von der Pütten AM, Krämer NC, Hoffmann L, Sobieraj S, Eimler SC. An experimental study on emotional reactions towards a robot. *Int. J. Soc. Robot.* 2013, 5(1):17–34.
[76] Ebers M. AI robotics in healthcare between the EU Medical Device Regulation and the Artificial Intelligence Act. *Oslo Law Rev.* 2024, 11(1):1–12.
[77] Suwa S, Tsujimura M, Ide H, Kodate N, Ishimaru M, *et al.* Home-care professionals' ethical perceptions of the development and use of home-care robots for older adults in Japan. *Int. J. Hum.-Comput. Interact.* 2020, 36(14):1295–1303.
[78] Kaime T, Chirwa S. Freedom of thought, conscience and religion. In *The African Charter on the Rights and Welfare of the Child: A Commentary*. Pretoria: Pretoria University Law Press, 2024.
[79] Kistakis Y. Protection of the forum internum. In *European Convention of Human Rights*. Brussels: Bruylant, 2011.
[80] Cohen S. Beneficence and autonomy. *Med. Health Care Philos.* 2019, 22(2):297–304.
[81] Zhang H, Wang Y, Zhang Z, Guan F, Zhang H, *et al.* Artificial intelligence, social media, and suicide prevention: principle of beneficence besides respect for autonomy. *Am. J. Bioeth.* 2021, 21(7):43–45.

[82] Reguart-Segarra N, Camarero-Suárez V. Camarero-Suárez, Freedom of thought, conscience and religion under the European Convention on human rights: new approaches. In *Protection and Promotion of Freedom of Religions and Beliefs in the European Context*. Cham: Springer, 2023.
[83] Neuwirth RJ. Prohibited artificial intelligence practices in the proposed EU artificial intelligence act (AIA). *Comput. Law Secur. Rev.* 2023, 48:105798.
[84] Lebret A. The Council of Europe Convention on Artificial Intelligence and Human Rights: a primarily procedural step towards safeguarding health rights in the digital age. *J. Glob. Health Law* 2025, 2(1):93–113.
[85] Seatzu F, Fanni S. The experience of the European Court of Human Rights with the European Convention on Human Rights and Biomedicine. *Utrecht J. Int. Eur. Law* 2015, 31(81):5.
[86] Ienca M. On neurorights. *Front. Hum. Neurosci.* 2021, 15:701258.
[87] Cornejo-Plaza MI, Cippitani R, Pasquino V. Chilean supreme court ruling on the protection of brain activity: neurorights, personal data protection, and neurodata. *Front. Psychol.* 2024, 15:1330439.
[88] Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, *et al.* AI4People—an ethical framework for a Good AI Society: opportunities, risks, principles, and recommendations. *Mines Mach.* 2018, 28(4):689–707.
[89] Guerrero Quiñones JL. Using artificial intelligence to enhance patient autonomy in healthcare decision-making. *AI Soc.* 2024, 40(3):1917–1926.
[90] Friedman B, Kahn Jr PH, Borning A, Huldtgren A. Value sensitive design and information systems. In *Early Engagement and New Technologies: Opening up the Laboratory*. Dordrecht: Springer, 2013. pp. 55–59.
[91] Finnis J. *Natural Law and Natural Rights*. Oxford: Oxford University Press, 2011.
[92] Fuller LL. *The Morality of Law.* New Haven: Yale University Press, 1965.
[93] Hessami AG, Kriebitz A, Weger G, Watson EN, Shaw P. Artificial intelligence for the benefit of everyone. *Computer* 2024, 57(9):68–79.
[94] Mantelero A. AI and big data: a blueprint for a human rights, social and ethical impact assessment. *Comput. Law Secur. Rev.* 2018, 34(4):754–772.
[95] Dritsas E, Trigka M, Mylonas P. A survey on privacy-enhancing techniques in the era of artificial intelligence. In *Proceedings of the 4th International Conference on Novel and Intelligent Digital Systems*, Athens, Greece, September 25–27, 2024, pp. 385–392.