Artificial Agency and the Ethics-Law Divide in AI Governance: A Philosophical Reflection

1st Alexander Richard Kriebitz
Ludwig-Maximilian University
Technical University of Munich
Munich, Germany
kriebitzalexander@gmail.com
0000-0001-7959-5980

2nd Ali Hessami
City University London
United Kingdom
hessami@vegaglobalsystems.com
0000-0001-6693-0436

3rdNell Watson
European Responsible Artificial Intelligence Office
United Kingdom
nell@nellwatson.com
0000-0002-4306-7577

4th Amanda Horyzk
University of Edinburgh
United Kingdom
amandahorzyk@outlook.com
0009-0006-6657-3305

5th Patricia Shaw
Beyond Reach Consulting Limited
United Kingdom
trish@beyondreach.uk.com
0000-0003-1477-3090

***Abstract*—AI governance commonly treats ethics and law as complementary normative instruments, often framed through the distinction between "soft" and "hard" law. This paper challenges that view by arguing that ethics and law constitute a fragile normative equilibrium destabilized when AI encodes macro-level societal preferences into micro-level individual moral reasoning spaces, with adverse consequences for personality rights. As AI systems increasingly embed societal logics of standardization, optimization, and behavioral steering into spaces commonly associated with individual moral reasoning, the functional differentiation between ethics and law becomes blurred, placing pressure on the traditional anthropological and meta-ethical assumptions underlying existing normative traditions. The paper concludes that the emergence of artificial moral agency reinforces the need for an anthropologically grounded framework of normativity that preserves the functional distinction between ethics and law, in order to uphold individual autonomy.**

***Index Terms*—ethics, moral philosophy, artificial intelligence, human rights, artificial moral agency**

I. Introduction

Artificial intelligence (AI) systems increasingly shape the conditions under which human decisions, social interactions, and collective practices take place [1]. Far from merely supporting human choice, contemporary AI systems structure options, prioritize values, and mediate judgments across a wide range of domains, including language processing, mobility, care, content governance, and public security [2]. For example, recommender systems curate moral and political information, autonomous vehicles encode risk trade-offs, and conversational agents provide guidance in ethically sensitive situations. In doing so, they contribute to the evolution and emergence of norms that influence individual and institutional behavior.

Contemporary approaches to AI governance have largely addressed this issue through two instruments: ethics and legal regulation [3]. Ethical frameworks, guidelines, and 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 [4]. These approaches are often implicitly framed as complementary, with ethics functioning as "soft" or voluntary guidance and law as an enforceable constraint [5].

This paper rejects the commonplace view of ethics as 'soft' guidance and law as 'hard' enforcement. Instead, it makes the case that ethics and law are structurally distinct yet interdependent normative domains, which are increasingly destabilized by the functional agency of AI [6]. Across heterogeneous use cases including recommender systems, automated driving, large language models, and robotics, AI systems increasingly generate outcomes that shape human behavior and social norms, yet without necessarily possessing moral subjectivity or intentionality. This leads to increasingly blurred lines between ethics and law, with an increasing shift towards integrating societal preferences within spaces which were previously reserved for individual moral agency.

The paper seeks to make the following contributions:

1) It develops a conceptual model that frames ethics and law as functionally differentiated yet mutually equilibrating normative domains in the context of AI governance.
2) It provides a systematic diagnosis of how AI-mediated expressions of moral agency destabilizes this differentiation by embedding collective, macro-level norms into contexts of individual ethical reflection.
3) It proposes preliminary operational guidance for AI governance and system design by identifying when AI systems risk displacing individual moral agency through the embedding of macro-level normative logics.

The remainder of the paper proceeds as follows: First, it reconstructs the philosophical foundations of normativity 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.

II. ETHICS AND LEGAL REGULATION AS SEPARATE YET INTERDEPENDENT EXPRESSIONS OF NORMATIVITY

Both ethics and legal regulation are primarily concerned with norms. 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 meta-ethical questions, considering the existence of universal moral truths or epistemic justification of normative claims and judgments through means such as reason, common sense, conscience, or sense of justice [7].

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 [8]. These assumptions shape not only the articulation of norms but also lay the foundation of attributive principles, such as responsibility, accountability, complicity, or culpability, which tie human agency to norm realization, or failure to do so [9].

Despite these parallels, ethics and law are different dimensions of normativity. Ethics is primarily non-institutionalized. Ethical norms typically emerge from cultural traditions, personal convictions, religious practices, or philosophical reflection rather than from formal political or legislative processes. Accordingly, adherence to ethical norms is motivated by persuasion and moral guidance rather than legal force. Moreover, ethics encompasses a wide plurality of normative traditions such as deontological, utilitarian, virtue-ethical, and faith-based approaches found in Western and Eastern Christianity, Islam, Buddhism, or Hinduism, some of which have historically been hesitant toward codification [10]. Ethical theories on human agency are therefore not only more deeply connected to fundamental beliefs but also tend to prescribe a wider area of desirable human conduct, filling spaces left by the law. The normative space of ethics is, thus, expansive, as it offers a holistic normative theory of human agency [11].

A deeper divergence concerns the very construction of norms themselves. Ethics often emphasizes the individual dimension: the alignment of action with one's own values and beliefs, grounded in autonomy rather than heteronomy. Ethical normativity, even in faith-based traditions, is typically anchored "from within", explaining why intentionality traditionally plays a crucial role within normative ethics [11]. By contrast, legal norm construction is fundamentally collective and institutional, shaped through political processes and oriented towards social coordination. Law must explicitly consider incentives, causal chains, and predictable effects, and expect compliance with codified and incentivized norms, focusing on whether concrete acts align with legal norms. Certain ethical traditions, such as virtue ethics, Kantian ethics, or religious moral frameworks in Mithraism, Christianity or Buddhism de-emphasize consequentialist considerations and incentives alike. These approaches prioritize aligning the intentions of the acting individual with moral duties, rights, or character virtues over external rewards [12]. In practice, however, ethics and law exert mutual influence. Individuals and groups shape moral sentiments, which in turn inform legal developments, and vice versa. Hence, recent years have witnessed the emergence of mixed types, for instance professional ethics or bioethics, which have been formalized, yet not incentivized or enforced. The concept of rights provides a particularly strong point of interaction: historically, individual moral convictions at the individual level are often grounded in religious or ethical traditions that have been translated into collective norms, for example, in matters of sexual morality or fundamental social values. Meta-ethical shifts, particularly in Western societies, have played a key role in the transition from monolithic notions of the "common good" towards pluralistic understandings of individual moral outlooks, alongside an increasing philosophical acknowledgement of normative uncertainty, aligning with a higher regard for personal choices and meta-norms such as human rights that seek to guarantee fundamental freedoms and rights [13], [14].

The idea to conceptually separate ethics and legal regulation manifests in protections like the right to resistance, civil disobedience, personality rights, and the "forum internum" as the inviolable inner sphere of conscience shielded from external authority. These principles formalize individual ethical autonomy through institutionalized protections over freedom of conscience and belief, academic freedom, procedural rights, and, ultimately, the prohibition of torture [15]. Yet at the same time, this normative equilibrium is inherently fragile, given strong societal preferences that invariably collide with rights-protected individual spaces.

III. ARTIFICIAL MORAL AGENCY

Ethics and law have long presupposed human moral agents [16]. Artificial intelligence (AI) encompasses a family of computational technologies characterized by statistical modelling, data-driven learning, algorithmic opacity, and automated decision-making [17]. 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. Most formal definitions of AI, therefore, focus on replicating, optimizing or even extending selected human or biological capacities through statistical modeling, and technological mimicry across domains and generations of AI developments.

Importantly, the problem of artificial moral agency does not begin with recent developments in agentic AI; rather, it visibly occurs in earlier forms of algorithmic systems insofar as they shaped morally relevant decisions, structured options, and mediated normative judgments without themselves being

fully autonomous agents. Nevertheless, AI is not autonomous in the biological sense. It is heteronomous, operating within the conditions defined by human developers [18]. Unlike traditional tools, which primarily extend human bodily functions, AI extends cognitive functions and, thus, the agency of human beings. This agency is functional: the capacity of a system to produce effects in the world, independent of conscious intention, yet derived entirely from human-directed processes. An AI system trained to detect cancer in medical images, for instance, substitutes the diagnostic role of the radiologist, by performing visual pattern recognition but it does not reflect other human faculties.

The functional agency of AI draws on clinicians labeling images, hospitals contributing diverse patient datasets, engineers writing code, and data scientists curating and preprocessing the inputs. Over time, outputs generated or assisted by AI modules emerge from a statistical combination of these multiple human agencies, they reflect a distributed form of functional agency that is irreducible to any single individual and are shaped by moral views and legal frameworks that were implicitly or explicitly embedded in system design [19]. The agency associated with AI systems is not only functional but also exposes patterns in past human decisions, and respectively, their underlying rationales, intentions, and deviations from these. It challenges traditional human-centered, atomized notions of moral agency – its expressions and consequences – as non-human components interfere with analogue processes relying on forms of moral deliberation and decision-making.

A major challenge lies in embedding norms within this process [20]. Computational models of established moral theories, such as Kantian deontology or consequentialism, have been proposed within value-based engineering and ethics-by-design approaches [21]. Yet human moral intuitions and contextual ethical judgments often defy formalization, making flexible, context-aware ethical reasoning difficult to implement. Implementation is further complicated by human bias, limited representativity of data, and uncertainty in the interpretation of legal norms, which limits machine learning approaches of reconstructing artificial moral behavior through the observation of actual human behavior [22]. As a result, the realization of artificial morality tends to gravitate towards pre-programmed rule-based and more legalistic approaches to normative questions. While this is compatible, for instance, with the rule-based approach of the Kantian tradition, it would not only collide with the observed diversity within ethics but also challenge existing assumptions regarding the division of labor between law and ethics [23].

This is amplified by recent research on human-machine interaction, which illustrates how human interactions with AI shape perception, trust, and normative expectations. Empirical studies show that humans tend to ascribe agency and even moral responsibility to AI, particularly when systems appear anthropomorphic or socially interactive [24]. These interactions can foster parasocial relationships, one-sided emotional bonds that influence individual decision-making, from formulations of apology letters to complicated questions of vocational ethics [25]. Besides, individuals may blame or seek justification from conversational AIs after following its advice on sensitive matters, or professionals may consult algorithmic systems for high-stakes decisions, underestimating the influence of AI on their own reasoning and the relatively weak epistemic status of such moral judgments. Hence, the implications of human machine interactions are in themselves morally relevant.

Ultimately, the human-AI interaction illustrates the tangible consequences of embedding norms within AI, or choosing not to do so. Against this background, the growing introduction of AI systems raises a host of questions:

- Whose agency does AI replace or augment?
- 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 interactions between humans and AI shape normative perceptions and responsibilities?

IV. PRACTICAL EXAMPLES IN APPLIED AI

To investigate this further, we explore four distinct cross-domain applications of AI: recommender systems in social media (A), autonomous driving technologies (B), large language models (C), and care robots (D). Through these examples, we aim to trace how the introduction of artificial intelligence affects the equilibrium between ethics and law as distinct yet intertwined pillars of normativity.

A. Recommender Systems in Social Media

Recommender systems on social media exemplify how AI operationalizes functional agency and influences human decision-making [25]. 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: that of users, content contributors, platform content moderators or the executed algorithm its developers and designers.

While such systems affect individual agency for example through pre-selecting videos or news their design is typically oriented toward commercial objectives, such as maximizing engagement, rather than directly advancing users' interests [26], [52]. Recommender systems are therefore not neutral tools of delegation but are embedded within broader socio-economic and technical architectures that shape the conditions under which individual judgment is exercised. Their normative impact manifests at both individual and societal levels. At the individual level, algorithmic curation may reinforce biases, foster echo chambers, or shape moral and emotional responses. At the societal level, it interacts with democratic values, public discourse, and freedom of expression, embedding corporate and technical norms into the informational environment [26]. In this respect, recommender systems do not merely organize information; they contribute to structuring the moral and political horizons within which users orient themselves.

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.

Recommender systems thus function as distributed normative agents, in which agency emerges from the interaction of developers, data, algorithms, and users. Due to the prevalent engagement-based incentive structures, recommenders optimize for systemic or collective preferences over preserving or merely extending delegated forms of individual perspectives. Consequently, they tend by design to embody legal or societal frameworks rather than to model or extend individual moral agency. This dynamic reinforces debates about transparency and opt-out mechanisms that seek to enable individual human judgment. Transparency specifically may enable users to better understand the normative logic shaping not only their behavior but also their emotional well-being and sense of agency [27].

B. 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 not only by traffic law but also by conventions, tacit expectations, and context-sensitive judgments, all of which become more visible as machines progressively assume tasks once performed by human drivers [28].

AVs must navigate distributions of risk, including situations involving unavoidable harm, which necessitates the preprogramming 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 [29].

Comparable forms of discretionary judgment are also evident in professional roles, such as a bus driver waiting briefly for a child to board safely. Such practices are often tolerated, and even morally expected, despite departing from strict procedural consistency. Once translated into machine behavior, however, these forms of flexibility become difficult to justify, as every exception risks being treated as a programmable standard. In this respect, AI design tends to gravitate toward the logic of legal consistency rather than the contextual elasticity often associated with human ethical judgment [30].

These design choices reveal a shift in agency: human drivers' judgment is replaced by decisions embedded in software, concentrated in engineers, companies, and regulators. Ethical and legal defaults, once distributed amongst countless individuals, are now standardized, producing predictable outcomes but narrowing the scope for individual moral discretion. In this way, societal preferences, such as the prioritization of safety, are embedded into systems of everyday use. The German Ethics Commission on Automated Driving exemplifies this tendency by translating abstract ethical principles into concrete programming requirements prior to binding legislation [31].

Similar tendencies can be observed in the growing reliance on risk-based and utilitarian ethical approaches [29], [32]. 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 about whose agency is actually being expanded. For instance, does an automated school bus primarily enhance the agency of the school (by ensuring punctual attendance), that of the children or their parents (by enhancing safety), or rather that of collective administrative and societal interests?

Autonomous driving thus brings the tension between individual and societal normativity into particularly sharp focus. Decisions about issues such as trajectory planning or risk thresholds do not merely reflect technical optimization but encode judgments about acceptable conduct, permissible trade-offs, and the distribution of risk between individuals and the public [32]. AVs therefore exemplify how AI does not merely automate action but redistributes the authority to define normatively acceptable behavior in morally and legally significant contexts.

C. Large Language Models

Large language models (LLMs) such as ChatGPT or Co-Pilot illustrate a particularly expansive form of AI-mediated functional agency: they influence human decision-making, self-expression, and normative orientation through language itself [33]. Unlike task-specific AI, LLMs respond to human prompts across a wide range of contexts, producing statistically plausible outputs without independent moral reflection. 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 expression—for instance, when 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 [34]. Compared to other AI applications, LLMs operate in unusual proximity to the protected space of individual moral reasoning (*the forum internum*), potentially interfering with users' intimate, reflective, and relational expressions. This heightened proximity intensifies ethical concerns surrounding trust, dependency, and the internalization of normative influence.

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 [35]. While LLMs can provide coherent responses, they lack reflective moral reasoning and any form of meaningful interpretable logics of its data processing. Personalization may create "ethical comfort zones," reinforcing preferences or biases, and fostering emotional attachment, mirroring the parasocial dynamics raised previously [36]. Rather than presuming political neutrality, it is more precise to treat many deployed systems as presenting neutrality while operationalizing implicit normative commitments through data priors, safety policies, and institutional risk constraints.

From an individual ethics perspective, LLMs are often implicitly treated as confidants, giving rise to expectations of confidentiality and loyalty. Yet these expectations may conflict with legal, institutional, or public-interest logics embedded in system design, particularly in contexts involving self-harm, criminal intent, or potential self-incrimination. In such cases, the tension between micro-level ethical expectations and macro-level normative commitments becomes more visible.

At the macro level, LLMs shape societal norms by standardizing language, framing discourse, and implicitly promoting certain normative or epistemic patterns [37]. Defaults in system design, such as moderation policies or prompt responses, operate as normative anchors, and can subtly influence collective behavior, public discourse, and ethical expectations. In this respect, LLMs do not merely mediate communication but increasingly participate in shaping the normative and interpretive environment within which communication takes place. Therefore, LLMs exemplify with particular intensity the dual role of interactional AI as both an individual companion and an agent realizing societal preferences that are closely aligned with a macro-level normative order [37].

D. 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 [38]. Through their embodied presence including manifested in movement, 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 [24]. 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.

In healthcare, elder care, and domestic settings, are expected not merely to perform routine tasks but also to simulate empathy, attentiveness, and context-sensitive responsiveness, while respecting norms rooted in human rights, professional care obligations, and cultural expectations. 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 [39].

This introduces a particularly sharp tension between individual and societal perspectives on normativity. 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. However, personalization as well as anthropomorphization can amplify perceptions of intimacy and trust, meanwhile risking the creation of "comfort zones," which can subtly shape individual behavior, thought and even expectations around machine-mediated moral engagement [40].

This tension becomes especially visible where confidentiality, safety, and intervention duties intersect. Human care professionals often exercise situational discretion in such cases, weighing vulnerability, therapeutic trust, and context rather than mechanically reporting every deviation. A care robot, by contrast, may be designed to register and escalate such conduct automatically, for instance by notifying staff that a patient failed to take prescribed medication, ignored dietary restrictions, or disclosed suicidal ideation. This reveals a central normative design challenge: whether the robot should function primarily as a trustworthy relational presence within a care relationship, or as a monitoring and compliance-enforcement device acting on behalf of the institution. In practice, such tensions are unlikely to be resolved through absolute rules alone but instead require carefully defined thresholds, limited exception structures, and proportionate forms of oversight.

Care robots thus operate as distributed normative agents, where functional agencies emerge from designers, programmers, and embedded ethical rules, while human perception projects moral weight onto their actions [41]. This duality prompts broader questions: Should robots be allowed to shape human behavior? 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 make visible the ethical and regulatory challenge of embedding morality into AI systems deployed in spaces of dependency, vulnerability, and trust. In this respect, they illustrate with particular clarity how the artificial reconstruction of agency can destabilize the balance between individual ethical expectations and socially institutionalized forms of normativity.

V. PRELIMINARY IMPLICATIONS FOR AI GOVERNANCE

The case studies reveal a recurring pattern: artificial moral agency incorporates collectively shaped normative logics into spaces historically associated with individual judgment, discretion, and moral reflection. In doing so, it destabilizes the fragile equilibrium between ethics and law. What varies across domains is not the existence of this pattern but the intimacy of the context, the stakes of the intervention, and the extent to which AI systems are perceived as trustworthy interlocutors, decision-making partners, or substitutes for human agency.

AI operates along a spectrum, from narrowly programmed tools to semi-autonomous systems, interacting across multiple

domains with varying ethical, legal, and societal stakes and implications. Even at lower levels of agency, AI's lack of subjectivity and reliance on statistical correlations limits its ability to replicate nuanced human moral reasoning, making it ill-suited to fully anticipate the ethical diversity of situations encountered in applications such as autonomous vehicles, LLMs, or care robots. Particularly, developments in robotics and AI systems deployed in intimate contexts raise distinct ethical considerations that demand closer attention to individual perspectives and values.

A central implication is the ambiguity of agency. AI rarely replaces human agency entirely; rather, it mediates, redistributes, and aggregates multiple human inputs into statistically derived decisions, which forces us to query how agency is meaningfully enhanced, if at all, by these systems. Functional agency is therefore distributed and algorithmically mediated: clinicians, developers, regulators, deployers, and users all contribute indirectly to outcomes that remain mostly opaque and often cannot be reduced to any single human intention. This challenges the simplistic opposition between "human" and "machine" agency and instead suggests a hybrid, co-constructed understanding of agency. The key question is therefore not only whether AI acts but whose agency is amplified, constrained, or displaced through artificial agency.

This ambiguity becomes especially salient when AI systems are deployed in intimate and normatively complex environments: Should care robots, autonomous vehicles, or large language models be permitted to serve as evidence-gathering tools for law enforcement and judicial proceedings? To what extent should content moderation systems incorporate active deradicalization strategies beyond mere content removal, and when do these strategies potentially infringe upon fundamental rights, such as freedom of expression? Given AI's agentic characteristics, these technologies enable the implementation of societal preferences within relatively intimate environments, amplifying concerns about privacy and autonomy.

AI's normative influence manifests both directly and indirectly. Directly, AI systems shape individual behavior and moral choices; for instance, LLMs providing guidance on personal dilemmas, or care robots influencing daily routines in intimate settings. Indirectly, AI affects societal norms through regulation, public discourse, and infrastructure design. This dual influence underscores the need to distinguishing between:

- Micro-level norms, grounded in individual ethical reflection, context-specific reasoning, and personal autonomy, and
- Macro-level norms, codified in law, institutional protocols, and collective societal expectations.

Navigating the tension between these levels of norms is especially salient in "moral grey zones," where legal and ethical guidance is incomplete or context-dependent. AI must often balance competing values and prioritize which normative layer dominates depending on the application, consequences, potential harms, and societal expectations. For example, medical diagnostic AI prioritizes micro-level medical confidentiality, while autonomous vehicles emphasize macro-level safety standards and liability frameworks. In other words, assuming the Hippocratic oath is treated as a baseline for a personalized approach to ethics in, for instance, a care robot, this would imply that the robot's black box nature is legitimate and that human oversight would be unable to access conversations between robots and patients, even in the event of a patient's suicide, due to the personality rights of the patient. This implies a certain level of tolerance for inefficiencies in the design of preventive measures, while treating individual rights such as freedom of conscience as constraints on realizing pro-social preferences normative consideration at stake.

A human rights–centered approach offers a critical bridge to reconcile these tensions. By grounding AI governance in recognized rights, such as freedom of conscience, expression, and mental integrity, developers and regulators can diffuse the tension between ethics and law. Human rights frameworks provide pre-existing meta-ethical assumptions such as freedom of will, and pluralistic understanding of morality that inform the conceptual distinction between micro-level autonomy and macro-level regulation. These assumptions are particularly salient in relation to rights protecting the *forum internum*, such as Article 18 of the International Covenant on Civil and Political Rights and Article 8 of the African Charter on Human and Peoples' Rights [42]. Its application to AI requires, therefore, doctrinal extension, particularly owing to the growing impact of AI on human agency, particularly considering its behavioral impact but also the integration of societal logic in the design of such systems. Hence, integrating these underlying anthropological and meta-ethical considerations into ethics-by-design approaches strengthens the alignment between AI functional agency and socially accepted normative boundaries, but more importantly, the articulation of redlines within the development of AI and specifically agentic AI.

Four operational factors emerge to embed this perspective into AI governance:

1) Graduated Autonomy: AI systems should progress through clearly defined levels of autonomy, calibrated to domain-specific risks and ethical stakes (e.g., higher standards for medical diagnostics than entertainment recommendations).
2) Normative Transparency: Developers and deployers of AI systems should disclose whether they optimize for individual autonomy or societal preference, and whose values are encoded.
3) Domains Reserved for Individual Ethics: Certain contexts (end-of-life, political belief, therapeutic, justice) warrant protection from AI-mediated societal preference injection, analogous to attorney-client privilege or the relationship between care-givers and patients.
4) Reflection Modes: Systems in sensitive contexts should offer modes that suspend persuasive optimisation and present options neutrally.

In summation, the factors described might be reinforcing each other. A tool which operates with high level of autonomy, is opaque in terms of the embedding of values in its design,

takes place in the context of the justice system, and does not offer interpretable reasoning, is likely to constitute a major risk for human rights, while also shifting by design to more mainstream societal preferences. The factors described might be highly relevant for risk assessments of AI systems.

Even outside of such high risk cases, the reconstruction of morality in AI increasingly forces us to balance norms emerging from individual ethics and societal legal considerations. In the absence of this, the further evolution of AI might result in a further shift within normative reasoning at the expense of individual ethics with strong repercussions on rights meant to protect individual ethical reasoning.

Consequently, the observation of a shift within normativity towards a more legalistic and rule-based understanding aligned with societal preferences has strong repercussions on the interpretation of existing AI ethics frameworks. Contemporary approaches, exemplified by the High-Level Expert Group on AI (HLEG) guidelines and the AI4People initiative, often privilege societal-level considerations such as beneficence and collective welfare [43]. However, this macro-ethical orientation necessitates critical examination from multiple theoretical perspectives. First, theories addressing the reconciliation of competing ethical principles within AI system design require reconstruction through the lens of agency: specifically, whose agency is augmented or diminished through AI deployment, particularly when analyzed from role-based perspectives? Second, scholars must investigate how AI systems influence human moral judgment and decision-making processes. Third, the normative logics themselves (the fundamental frameworks of ethical reasoning) embedded within AI ethics discourse warrant scrutiny. These distinctions matter because individual-level ethical protections operate according to different normative principles than aggregate utilitarian calculations or deontological principles aimed at maximizing societal benefit.

VI. CONCLUSION

This paper has made the case that the common framing of ethics as "soft" normativity and law as its "hard" counterpart is insufficient for understanding the challenges posed by artificial moral agency. Rather than differing merely in enforceability, ethics and law represent distinct but interdependent modes of norm construction: ethics remains more closely tied to individual moral reflection, conscience, and situational judgment, whereas law institutionalizes collectively binding norms through procedural and coercive structures [1], [9], [15]. The increasing deployment of AI, and particularly of more agentic systems, places this relationship under pressure by introducing functional forms of agency into domains historically structured around human judgment and responsibility [18], [19].

Across the case studies, a recurring pattern emerges: AI systems do not simply automate tasks but increasingly mediate, redistribute, and embed normative expectations within individual contexts of action, reflection, and dependence. Recommender systems shape attention and preference formation [25], [26]; autonomous vehicles transform discretionary judgment into preconfigured decision architectures [29], [30]; large language models enter spaces of reflection, interpretation, and advice [26], [32], [37]; and care robots mediate trust, vulnerability, and intervention in embodied relational settings [38], [40], [41]. In each case, AI introduces collectively shaped, institutionally structured, or commercially optimized normative logics into spaces that have traditionally remained more closely associated with individual ethical self-determination.

This has significant implications for AI governance but also on the understanding of human rights protecting these areas from external interference. The central challenge is not only how to align AI systems with values, legal requirements, or social preferences, but also how to preserve the normative distinction between domains that should remain open to personal moral judgment and those in which collective coordination and institutionalization are justified. In this respect, governance needs not only address the values are embedded into AI, but where those values originate, whose interests they reflect, and under what conditions their implementation risks displacing ethical plurality, moral discretion, or protected forms of inward freedom [43].

A human rights–centered approach is especially important here, not merely as an external constraint on technical systems, but as a normative framework capable of identifying protection zones for autonomy, conscience, privacy, and personality. Human rights, in particularly personality rights but also freedom of thought are relevant as they aim to protect the *forum internum*, including intimate reflection, belief formation, therapeutic interaction, and morally sensitive decision-making from societal pressure [42]. In such contexts, safeguarding autonomy requires more than transparency, explainability, or generic human oversight but also the articulation of clearly defined red-lines that are embedded in AI systems operating in these fields. It requires preserving the conditions under which persons remain capable of forming, revising, and acting upon their own judgments without undue normative pre-structuring by technical systems. These considerations can have important implications for architectural choices and socio-technical measures to build safeguards around the *forum internum* while monitoring the impact of algorithms on moral agency through e.g., audit trails, edge computing, regulatory oversight as well as privacy-enhancement, user-empowerment, contestation and critical AI literacy.

The challenge of artificial moral agency is thus not merely whether systems comply with specific ethical or legal standards but whether they preserve the normative conditions of meaningful human moral agency. In this respect, the future of human centred AI governance depends not only on better alignment with ethical values but on the inclusion of human rights and anthropological considerations when navigating areas characterized by tensions between ethics and law.

REFERENCES

[1] OECD, "Explanatory memorandum on the updated OECD definition of an AI system," Mar. 5, 2024. [Online]. Available: https://www.oecd.org/en/publications/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_623da898-en.html

[2] A. Bandi, B. Kongari, R. Naguru, S. Pasnoor, and S. V. Vilipala, "The rise of agentic AI: A review of definitions, frameworks, architectures, applications, evaluation metrics, and challenges," *Future Internet*, vol. 17, no. 9, Art. no. 404, 2025.

[3] M. R. Carrillo, "Artificial intelligence: From ethics to law," *Telecommunications Policy*, vol. 44, no. 6, Art. no. 101937, 2020.

[4] M. Khamassi, M. Nahon, and R. Chatila, "Strong and weak alignment of large language models with human values," *Scientific Reports*, vol. 14, no. 1, Art. no. 19399, 2024.

[5] L. Floridi and M. Taddeo, "Moral vs legal norms: Soft and hard ethics," in *A Companion to Digital Ethics*, 2025, pp. 11–23.

[6] W. B. Wendel, "Legal ethics and the separation of law and morals," *Cornell Law Review*, vol. 91, pp. 67–120, 2005.

[7] F. Kissling, "The place for individual conscience," *Journal of Medical Ethics*, vol. 27, suppl. 2, pp. ii24–ii27, 2001.

[8] B. Williams, *Morality: An Introduction to Ethics*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[9] F. Rowe, M. Jeanneret Medina, B. Journé, E. Coëtard, and M. Myers, "Understanding responsibility under uncertainty: A critical and scoping review of autonomous driving systems," *Journal of Information Technology*, vol. 39, no. 3, pp. 587–615, 2024.

[10] A. Brysk, "Engaged Buddhism as human rights ethos: The constructivist quest for cosmopolitanism," *Human Rights Review*, vol. 21, no. 1, pp. 1–20, 2020.

[11] P. Cane, "Morality, law and conflicting reasons for action," *The Cambridge Law Journal*, vol. 71, no. 1, pp. 59–85, 2012, doi: 10.1017/S0008197312000207.

[12] Y. Bathaee, "The artificial intelligence black box and the failure of intent and causation," *Harvard Journal of Law & Technology*, vol. 31, pp. 889–938, 2017.

[13] D. Ulansey, *The Origins of the Mithraic Mysteries: Cosmology and Salvation in the Ancient World*. Oxford, U.K.: Oxford Univ. Press, 1991.

[14] J. R. Rowan, "Grounding hypernorms: Toward a contractarian theory of business ethics," *Economics & Philosophy*, vol. 13, no. 1, pp. 107–112, 1997.

[15] G. Hughes, "The concept of dignity in the Universal Declaration of Human Rights," *Journal of Religious Ethics*, vol. 39, no. 1, pp. 1–24, 2011.

[16] S. A. Fasoro, "Kant on human dignity: Autonomy, humanity, and human rights," *Kantian Journal*, vol. 38, no. 1, 2019.

[17] R. I. Mukhamediev et al., "Review of artificial intelligence and machine learning technologies: Classification, restrictions, opportunities and challenges," *Mathematics*, vol. 10, no. 15, Art. no. 2552, 2022.

[18] J. Burrell, "How the machine 'thinks': Understanding opacity in machine learning algorithms," *Big Data & Society*, vol. 3, no. 1, 2016.

[19] L. Floridi and J. W. Sanders, "On the morality of artificial agents," *Minds and Machines*, vol. 14, no. 3, pp. 349–379, 2004.

[20] J. H. Moor, "Why we need better ethics for emerging technologies," *Ethics and Information Technology*, vol. 7, no. 3, pp. 111–119, 2005.

[21] F. Berreby, G. Bourgne, and J.-G. Ganascia, "Modelling moral reasoning and ethical responsibility with logic programming," in *Logic for Programming, Artificial Intelligence, and Reasoning*. Berlin, Germany: Springer, 2015, pp. 532–548.

[22] C. Allen, I. Smit, and W. Wallach, "Artificial morality: Top-down, bottom-up, and hybrid approaches," *Ethics and Information Technology*, vol. 7, no. 3, pp. 149–155, 2005.

[23] J. Graham, J. Haidt, and B. A. Nosek, "Liberals and conservatives rely on different sets of moral foundations," *Journal of Personality and Social Psychology*, vol. 96, no. 5, pp. 1029–1046, 2009.

[24] C. Akbulut et al., "All too human? Mapping and mitigating the risk from anthropomorphic AI," in *Proc. AAAI/ACM Conf. on AI, Ethics, and Society*, vol. 7, no. 1, Oct. 2024, pp. 13–26.

[25] S. Milano, M. Taddeo, and L. Floridi, "Recommender systems and their ethical challenges," *AI & Society*, vol. 35, no. 4, pp. 957–967, 2020.

[26] Y. Benkler, "Don't let industry write the rules for AI," *Nature*, vol. 569, no. 7754, pp. 161–162, 2019.

[27] S. Bonicalzi, M. De Caro, and B. Giovanola, "Artificial intelligence and autonomy: On the ethical dimension of recommender systems," *Topoi*, vol. 42, no. 3, pp. 819–832, 2023.

[28] C. Luetge, "The German ethics code for automated and connected driving," *Philosophy & Technology*, vol. 30, no. 4, pp. 547–558, 2017.

[29] M. Geisslinger, F. Poszler, and M. Lienkamp, "An ethical trajectory planning algorithm for autonomous vehicles," *Nature Machine Intelligence*, vol. 5, no. 2, pp. 137–144, 2023.

[30] A. Bordum, "Immanuel Kant, Jürgen Habermas and the categorical imperative," *Philosophy & Social Criticism*, vol. 31, no. 7, pp. 851–874, 2005.

[31] J. Gogoll et al., "Ethics in the software development process: From codes of conduct to ethical deliberation," *Philosophy & Technology*, vol. 34, no. 4, pp. 1085–1108, 2021.

[32] E. Awad et al., "The moral machine experiment," *Nature*, vol. 563, no. 7729, pp. 59–64, 2018.

[33] B. Padiu, R. Iacob, T. Rebedea, and M. Dascalu, "To what extent have LLMs reshaped the legal domain so far? A scoping literature review," *Information*, vol. 15, no. 11, Art. no. 662, 2024.

[34] M. Baggot, "The quest for connection in AI companions," *Journal of Ethics and Emerging Technologies*, vol. 35, no. 1, pp. 1–20, 2025, doi: 10.55613/jeet.v35i1.202.

[35] Z. Zhang et al., "Secret use of large language models (LLMs)," *Proceedings of the ACM on Human-Computer Interaction*, vol. 9, no. 2, pp. 1–26, 2025.

[36] N. Noor, S. Rao Hill, and I. Troshani, "Artificial intelligence service agents: Role of parasocial relationship," *Journal of Computer Information Systems*, vol. 62, no. 5, pp. 1009–1023, 2022.

[37] S. Kruegel, A. Ostermaier, and M. Uhl, "Zombies in the loop? Humans trust untrustworthy AI-advisors for ethical decisions," *Philosophy & Technology*, vol. 35, Art. no. 17, 2022, doi: 10.1007/s13347-022-00511-9.

[38] A. van Wynsberghe, "Service robots, care ethics, and design," *Ethics and Information Technology*, vol. 18, no. 4, pp. 311–321, 2016.

[39] A. van Wynsberghe, "Designing robots for care: Care-centered value-sensitive design," *Science and Engineering Ethics*, vol. 19, no. 2, pp. 407–433, 2013.

[40] L. D. Riek et al., "How anthropomorphism affects empathy toward robots," in *Proc. ACM/IEEE Int. Conf. on Human-Robot Interaction*, Mar. 2009, pp. 245–246.

[41] A. M. Rosenthal-von der Puetten et al., "An experimental study on emotional reactions towards a robot," *International Journal of Social Robotics*, vol. 5, no. 1, pp. 17–34, 2013.

[42] T. Kaime and S. Chirwa, "Freedom of thought, conscience and religion," Pretoria, South Africa: Pretoria University Law Press (PULP), 2024, pp. 1–16.

[43] L. Floridi et al., "AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations," *Minds and Machines*, vol. 28, no. 4, pp. 689–707, 2018.

[44] J. Finnis, *Natural Law and Natural Rights*. Oxford, U.K.: Oxford University Press, 2011.

[45] L. L. Fuller, *The Morality of Law*. New Haven, CT, USA: Yale University Press, 1964.

[46] B. Friedman, P. H. Kahn Jr., A. Borning, and A. Huldtgren, "Value sensitive design and information systems," in *Early Engagement and New Technologies*. Dordrecht, Netherlands: Springer, 2013, pp. 55–95.

[47] M. I. Cornejo-Plaza, R. Cippitani, and V. Pasquino, "Chilean Supreme Court ruling on the protection of brain activity: Neurorights, personal data protection, and neurodata," *Frontiers in Psychology*, vol. 15, Art. no. 1330439, 2024.

[48] A. G. Hessami et al., "Artificial intelligence for the benefit of everyone," *Computer*, vol. 57, no. 9, pp. 68–79, 2024.

[49] A. Mantelero, "AI and Big Data: A blueprint for a human rights, social and ethical impact assessment," *Computer Law & Security Review*, vol. 34, no. 4, pp. 754–772, 2018.

[50] E. Dritsas, M. Trigka, and P. Mylonas, "A survey on privacy-enhancing techniques in the era of artificial intelligence," in *Proc. Novel & Intelligent Digital Systems Conf.*, Cham, Switzerland: Springer Nature, Sep. 2024, pp. 385–392.

[51] J. L. Guerrero Quiñones, "Using artificial intelligence to enhance patient autonomy in healthcare decision-making," *AI & Society*, vol. 40, no. 3, pp. 1917–192

[52] A. Horzyk, "Data protection and privacy: Risks and solutions in the contentious era of AI-driven ad tech," in *Neural Information Processing (ICONIP 2023)*, B. Luo, L. Cheng, Z. G. Wu, H. Li, and C. Li, Eds., vol. 1968, *Communications in Computer and Information Science*. Singapore: Springer, 2024, ch. 27, doi: 10.1007/978-981-99-8181-6_27.