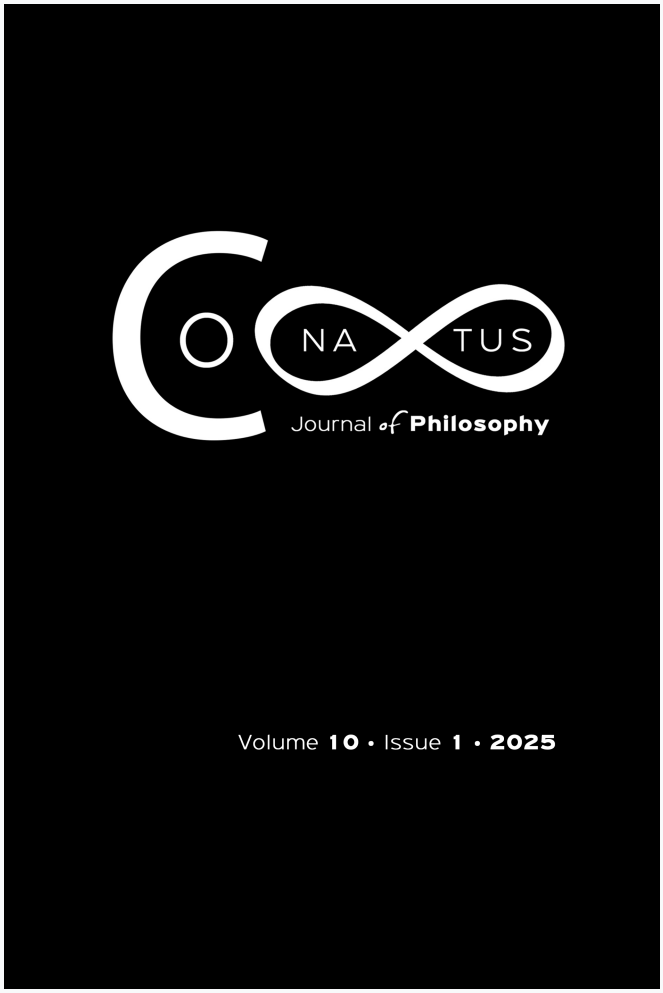
Conatus - Journal of Philosophy

Vol 10, No 1 (2025)

Conatus - Journal of Philosophy



Virtue in the Machine: Beyond a One-size-fits-all Approach and Aristotelian Ethics for Artificial Intelligence

Alkis Gounaris, George Kosteletos, Maria-Artemis Kolliniati

doi: 10.12681/cjp.40628

To cite this article:

Gounaris, A., Kosteletos, G., & Kolliniati, M.-A. (2025). Virtue in the Machine: Beyond a One-size-fits-all Approach and Aristotelian Ethics for Artificial Intelligence. *Conatus - Journal of Philosophy*, *10*(1), 127–152. https://doi.org/10.12681/cjp.40628

A. Gounaris, G. Kosteletos, & M. A. Kolliniati · *Conatus* 10, no. 1 (2025): 127-152
doi: https://doi.org/10.12681/cjp.40628

Virtue in the Machine: Beyond a One-size-fits-all Approach and Aristotelian Ethics for Artificial Intelligence

Alkis Gounaris
National and Kapodistrian University of Athens, Greece
E-mail address: alkisg@philosophy.uoa.gr
ORCID iD: https://orcid.org/0000-0002-0494-6413

George Kosteletos
National and Kapodistrian University of Athens, Greece
E-mail address: gkosteletos@philosophy.uoa.gr
ORCID iD: https://orcid.org/0000-0001-6797-8415

Maria-Artemis Kolliniati
Ruprecht Karls Universität Heidelberg, Germany
E-mail address: martkolliniati@gmail.com
ORCID iD: https://orcid.org/0000-0003-1553-7014

Abstract

This paper explores the application of Aristotelian virtue (arête), as quality of excellence and as a key notion of ethics, to AI systems as classified in the EU Artificial Intelligence Act. It argues that while the Act's approach based on 'ethical data' and 'prima facie values' aligns with the Rossian paradigm, such principles may not be suitable for all AI systems, particularly those in 'limited' or 'minimal risk' zones. The paper suggests that the Aristotelian concept of virtue can be effectively applied to designing, training, operating and using no-risk or low-risk AI systems. However, its application to the design and training of high-risk areas such as migration, asylum, border control, and justice, where clearly defined objectives are essential, requires ongoing consideration. The paper concludes that by distinguishing between (a) design, development, training, deployment, operation and use, (b) by stage evaluation of systems, and c) virtuous use of the systems, Aristotelian ethics can serve as a post ex evaluating method for all-risk AI systems, while further research and the potential use of regulatory sandboxes are needed to explore the integration of Aristotelian virtues into the design, development and training of such applications. Finally, we propose a virtuous-based 'AI Seal of Excellence' certification process, which empowers the virtuous use of AI systems.

Keywords: *AI ethics; AI virtues; virtuous agents; EU AI Act; arête; Aristotelian ethics for AI; seal of excellence for AI; virtuous use of AI; liberalism; borders and AI*

I. Introduction

The EU Artificial Intelligence Act classifies AI systems into four distinct risk zones, aiming to protect "fundamental rights, democracy, and the rule of law" (EU AI Act). This paper sets out to achieve three primary objectives. First, it asserts that the EU AI Act aligns with an approach based on 'ethical data' and 'prima facie values or duties,' resembling the Rossian paradigm.[1] This alignment is attributed to the Act's objective of ensuring the integrity of Artificial Intelligence (AI) systems, which are considered 'trustworthy AI' and must comply with eight core criteria, including transparency, non-discrimination, and fairness.[2] Second, it explores the potential for applying the "aretological" concept of 'virtue in the machine' to 'limited' and 'minimal and no risk' AI systems. Third, the paper aims to demonstrate that while Aristotelian ethics-based criteria may effectively *evaluate* 'high-risk' AI systems, there are challenges in applying them to the design, training, and operation of such systems. Through hypothetical scenarios, we argue that Aristotelian ethics may not be well-suited for guiding the development and deployment of AI systems in high-risk domains such as migration, asylum, border control, justice, etc. However, it can still serve as a valuable framework for *evaluating* these systems and as a method for guiding users on their virtuous use.

In addressing the first objective, the paper clarifies the EU AI Act's approach, illustrating that a uniform treatment across risk zones is impractical. Provisions on transparency and non-discrimination apply in particular to high-risk AI systems, which must address and overcome the 'value-loading problem.' The development of ethical AI agents requires overcoming this challenge by aligning AI with human values, often through 'prima facie' moral models. However, challenges persist, including the inability to predict all potential scenarios[3] and the necessity for external assessments of machine moral agency.[4]

For the second objective, the paper argues that 'ethical data' and 'prima facie values' may not apply or may not be checkable or even necessary to 'limited' or

---

[1] William D. Ross, *The Right and the Good* (Oxford University Press, 2002).

[2] European Commission, *Ethics Guidelines for Trustworthy AI* (Office for Official Publications of the European Communities, 2019).

[3] Eliezer Yudkowsky, "Complex Value Systems in Friendly AI," in *Artificial General Intelligence*, eds. J. Schmidhuber, K. Thó risson, and M. Looks, 388-393 (Springer, 2011); see also Eliezer Yudkowsky, "The Value Loading Problem," *EDGE*, July 12, 2021, https://www.edge.org/response-detail/26198.

[4] Michael Anderson and Susan Leigh Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent," *AI Magazine* 28, no. 4 (2007): 15; also, Michael Anderson, Susan Leigh Anderson, Alkis Gounaris, and George Kosteletos, "Towards Moral Machines: A Discussion with Michael Anderson and Susan Leigh Anderson," *Conatus – Journal of Philosophy* 6, no. 1 (2021): 177-202.

'minimal and no risk' AI systems. Drawing on examples such as the AI tutor,[5] we anchor our proposal on the concept of Aristotelian virtue (*arête*) by emphasising the cultivation of character and excellence in achieving goals throughout the AI system's life cycle. By doing so, we present a behavioural framework that focuses on the development of virtues rather than a purely functional framework, allowing for an external approach to evaluating the performance of AI systems in low-risk scenarios. This approach suggests using external criteria to address the 'value loading problem,' emphasising fostering individual virtues within a social context. The paper proposes defining a 'virtuous agent' guided by both social and individual ends[6] in alignment with the intrinsic nature of machines.

This "aretological" framework opens the possibility of considering machines as virtuous agents[7] recognised for their contributions to social and individual goals,[8] with their performance evaluated according to their achievements over time[9] rather than procedural aspects.[10]

For the third objective, the paper argues that Aristotelian ethics is ill-suited for guiding the design, development, training and deployment of AI systems engaged in high-risk activities, particularly those involved in areas such as migration, asylum and border control, justice,[11] education, and healthcare. In such contexts, it becomes essential to define the ultimate goal, or '*telos,*'[12] in advance, as AI systems lack the capacity for human-like deliberation concerning their ultimate goals.[13] The paper questions whether this goal can be adequately addressed through an Aristotelian approach alone. It contends that Aristotelian ethics is not a universal theory and that, by the same token, the EU AI

---

[5] John Tasioulas, "First Steps Towards an Ethics of Robots and Artificial Intelligence," *Journal of Practical Ethics* 7, no. 1 (2019): 61-95.

[6] Michael Sandel, *Justice: What's the Right Thing to Do?* (Farrar, Straus and Giroux, 2010).

[7] Martin Gibert, M. "The Case for Virtuous Robots," *AI and Ethics* 3 (2022): 135-144; see also Massimiliano Cappuccio, Eduardo Sandoval, Omar Mubin, Mohammad Obaid, and Mari Velonaki, "Can Robots Make us Better Humans? Virtuous Robotics and the Good Life with Artificial Agents," *International Journal of Social Robotics* 13 (2021): 7-22.

[8] Silviya Serafimova, "Whose Morality? Which Rationality? Challenging Artificial Intelligence as a Remedy for the Lack of Moral Enhancement," *Humanities and Social Sciences Communications* 7 (2020): 1-10.

[9] Sandel.

[10] John Tasioulas, "The Rule of Algorithm and the Rule of Law," Lecture at the University of Vienna, October 15, 2021.

[11] Alkis Gounaris and George Kosteletos, "Writing the Algorithm of Good: Artificial Intelligence as a Machine of Justice," *Ithiki* 19 (2024): 6-27 [in Greek].

[12] Aristotle, *The Nicomachean Ethics*, ed. L. Brown, trans. D. Ross (Oxford University Press, 2009).

[13] Tasioulas, "First Steps." For a Heidegger-inspired analysis on the lack of agency on behalf of AI systems, see also Ashley Roden-Bow, "Killer Robots and Inauthenticity: A Heideggerian Response to the Ethical Challenge Posed by Lethal Autonomous Weapons Systems," *Conatus – Journal of Philosophy* 8, no. 2 (2023): 477-486.

Act cannot adopt a one-size-fits-all approach. While Aristotelian ethics may be appropriate for minimal-risk and no-risk AI systems and for the sole evaluation of high-risk AI systems, it may be inadequate to guide the design, training and operation of systems deployed in high-risk activities such as migration, asylum, warfare,[14] and border control. Accordingly, the paper calls for further philosophical discussion and empirical research, suggesting that the potential use of regulatory sandboxes to explore behavioural evaluation criteria based on Aristotelian ethics may lead us to the safe and virtuous use of high-risk AI applications. Furthermore, we suggest that the virtuous use of AI systems can be realised through the introduction and application of the 'AI Seal of Excellence' certification process, which will be based on virtuous principles.

II. The EU AI Act, trustworthy AI and the 'value loading problem'

The EU Artificial Intelligence Act (EU AI Act) classifies AI systems into four distinct risk categories. First, unacceptable risk, which includes systems like AI used for biased criminal justice decision-making or social scoring. Second, high risk, which encompasses applications in areas such as healthcare, justice, border control, education, hiring, and autonomous vehicles. Third, limited risk, covering systems like chatbots and online shopping recommendation algorithms; and fourth, minimal/no risk, which includes systems such as weather forecasting or spam filters. This categorisation acknowledges that AI systems pose varying degrees of risk to fundamental rights, safety, and societal values, aiming to foster innovation while safeguarding the rule of law.

The EU AI Act adopts a uniform framework to ensure the development of 'trustworthy AI,' which aligns with 'ethical data' and 'prima facie values or duties,' reflecting the Rossian paradigm.[15] The Act aims to ensure that AI systems fulfil eight core criteria, including accountability, sustainability, privacy, and fairness, promoting technical robustness and societal well-being.[16] These ethical principles, such as transparency, non-discrimination, and fairness, ensure that AI systems comply with both functional and moral standards. However, applying these 'prima facie' ethical principles to lower-risk AI systems – such as those in the 'limited' or 'minimal/no risk' zones – raises questions. For such systems, an Aristotelian virtue-based approach may offer a more appropriate ethical framework, particularly as any uniform treatment across all risk categories would not be practical.

---

[14] See Ioanna Lekea, George Lekeas, and Pavlos Topalnakos, "Exploring Enhanced Military Ethics and Legal Compliance through Automated Insights: An Experiment on Military Decision-making in Extremis," *Conatus – Journal of Philosophy* 8, no. 2 (2023), 345–372.

[15] Ross.

[16] European Commission, *Ethics Guidelines for Trustworthy AI*.

While provisions on transparency and non-discrimination are crucial for high-risk AI systems, they are part of the broader 'value loading problem.' This challenge involves harmonising AI systems with human values, which often requires the application of Rossian 'prima facie duties' to a priori moral models. However, significant challenges persist, including the difficulty of foreseeing all possible scenarios, the need for external assessments to evaluate machine moral agency,[17] and the lack of freedom within AI systems to navigate conflicting moral principles and norms. In line with Aristotelian ethics, a 'virtuous' AI system would ideally have the freedom to choose the best course of action based on its deliberative faculties. Yet, this autonomy is absent in most AI systems, particularly in high-risk environments like border control, justice, healthcare or education, where conflicting human interests may arise.[18] For example, AI in the justice system could influence decisions that affect fundamental rights, with the risk of increasing bias or error in legal decisions, especially in 'hard cases' where reasonable lawyers and judges have to discover what the rights of the parties involved are in these contestable cases.[19] In healthcare, errors or biases in diagnosis or treatment could have serious, life-threatening consequences.[20] In education, students' future opportunities and, thus, life plans can be affected by biased AI applied to assessing students. Similarly, the management of border controls, on which we will focus, may involve conflicting interests – such as national security versus humanitarian concerns – and the choice between conflicting interests may have implications for individual freedoms, deprivation of individual rights, abuse and discrimination. If we assume an AI system has Aristotelian ethics, it would still struggle to resolve such conflicts without the capacity for moral deliberation or freedom of choice to make value-based judgments. For example, a border control AI system may face a conflict between national security concerns (e.g., controlling migration) and the humanitarian duty to protect asylum seekers. Since AI systems in these contexts lack the freedom to navigate such moral dilemmas, Aristotelian ethics would be inadequate for resolving these tensions.

Previous research into the 'value loading problem' has proposed the use of various ethical models, such as utilitarianism, deontology, prima facie val-

---

[17] Anderson and Anderson, 15; also, Anderson et al.

[18] In the light of this, especially when it comes to high-risk systems as war drones, some advocate "an international treaty banning all weaponized UAVs." See Joshua M. Hall, "Just War contra Drone Warfare," *Conatus – Journal of Philosophy* 8, no. 2 (2023): 217-239.

[19] Ronald Dworkin, *Taking Rights Seriously* (Harvard University Press, 1978).

[20] This would also undermine the doctor-patient relationship, since patients seem to be quite sensitive on the introduction of AI tools; see Georgia Livieri, Eleni Mangina, Evangelos D. Protopapadakis, and Andrie G. Panayiotou, "The Gaps and Challenges in Digital Health Technology Use as Perceived by Patients: A Scoping Review and Narrative Meta-synthesis," *Frontiers in Digital Health* 7 (2025): 1474956.

ues, and virtues.[21] However, these approaches face difficulties with conflicting moral principles and the challenge of predicting all possible future scenarios. Behaviour-oriented research[22] may potentially offer the only viable solution to this issue. In light of these obstacles and the structure of the "risk zones," we argue that while the EU AI law aims to ensure the ethical development of AI across different risk categories, in practice, it seeks to adopt a generic solution that may have limitations when applied to lower-risk AI systems. A one-size-fits-all solution is not appropriate in such cases. A virtue-based approach grounded in Aristotelian ethics may provide more suitable guidance for these systems, emphasising moral development within specific contexts rather than rigid adherence to pre-loaded duties or Rossian "prima facie duties." In the next chapter, the characteristics of an evaluation framework built on Aristotelian virtues are presented as a solution to tackle the value-loading problem.

III. Towards virtuous AI: An Aristotelian approach to overcoming the 'value loading problem'

In addressing the challenges of behaviour-oriented approaches to the development, evaluation, and use of AI systems, we propose an *evaluation* framework based on Aristotelian virtues.

As Aristotle explains in *Nicomachean Ethics*,[23] virtues (*arêtes*) must be understood in light of our characteristic function (*ergon*). For humans, this function is the activity of the rational part of the soul conducted well or in accordance with excellence. Hence, the cultivation of virtue is inseparable from our ultimate purpose (telos), since it enables us to perform our function in a fully realised manner. In so doing, we attain our proper end (*entelecheia*) and achieve genuine human flourishing (*eudaimonia*).[24]

Similarly, in other texts, such as *On the Soul*,[25] Aristotle distinguishes between the ultimate purpose and the characteristic function (ergon) of tools, exemplified by the axe, whose function is to cut well, thereby conferring upon it functional value.

Furthermore, in *Nicomachean Ethics*, Aristotle, identifies distinct types of virtues. Organic or functional virtues, as the virtue of the eye lies in its capacity

---

[21] Gounaris and Kosteletos, "Writing the Algorithm of Good."

[22] Nathan I. N. Henry, Mangor Pedersen, Matt Williams, Jamin L. B. Martin, and Liesje Donkin, "A Hormetic Approach to the Value-Loading Problem: Preventing the Paperclip Apocalypse," *arXivLabs* (2024), https://arxiv.org/abs/2402.07462.

[23] Aristotle, *The Nicomachean Ethics,* A, 7, 1097b22 – 1098a20.

[24] For an innovative account of Aristotle's views on flourishing or eudaimonia, see Pia Valenzuela, "Fredrickson on Flourishing through Positive Emotions and Aristotle's Eudaimonia," *Conatus – Journal of Philosophy* 7, no. 2 (2022): 37-61.

[25] Aristotle, *On the Soul*, trans. J. A. K. Thomson (Harvard University Press, 1959), II 1, 412b10-15.

of the eye to see clearly (B, 6, 1106a15-20), intellectual virtues (B, 1, 1103a 14-25) are related to logic, computation and learning, and moral virtues (*ibid*), such as practical wisdom (*phronesis*),[26] related with *ethos*, which is regarded as a dispositional choice (*hexis*).[27]

Although most of the above virtues can be attributed to AI systems, the attribution of the moral virtue remains debatable. This is primarily because moral virtue is inherently tied to freedom of choice, raising doubts about its applicability to machines whose operations are bounded by predefined purposes or ends (*telos*). Additionally, the ontological dimension of virtue takes precedence over its moral dimension, as ethics cannot exist without an underlying ontology. In this framework, moral virtue – as a choice of appropriate means – serves as the pathway through which an individual attains their ontological end.

In the same vein, an Aristotelian-based evaluation framework shifts the focus from solely examining the pre-loaded values or internal function of AI systems or their decision-making processes to considering the broader societal context in which these systems operate. By evaluating AI from an external – behaviour-based perspective, we can better address the challenges of creating genuinely virtuous agents. This includes examining the factors that contribute to their instrumental, functional, computational and/ or moral qualities, as well as their potential impact and benefits to society. Such an approach ensures that AI systems are technically effective and align with societal values, fostering their integration and positive contribution to human well-being.

Drawing on examples such as the AI-equipped tutor robot,[28] our approach emphasises the cultivation of character and virtue throughout the lifecycle of an AI system, presenting a behavioural rather than a functionalist theory. This externalist perspective proposes criteria rooted in virtue ethics to address the 'value loading problem,' highlighting the importance of cultivating virtues within a social context. In this context, we do not aim to create 'moral agents,' but instead, we argue that minimal or no risk AI systems can play the role of a 'low-risk agent' that can be considered virtuous because it meets all instrumental, functional, and intellectual criteria, without concern for whether it is a 'literally moral' agent. In fact, questioning its moral status would amount to an anthropomorphic projection onto the agent. However, when it comes

[26] Although Aristotle, in his *Nicomachean Ethics* (A, 13, 1103a 4-5), classifies *phronesis* among the intellectual virtues (alongside *sophia* and *synesis*), he nonetheless deems it indispensable for the realisation of ethical conduct. In this work, we treat phronesis as a foundational element inextricably tied to moral behaviour rather than a value that can be fully captured through computational representation.

[27] Richard Kraut, "Aristotle's Ethics," *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition), eds. Edward N. Zalta and Uri Nodelman, https://plato.stanford.edu/archives/fall2022/entries/aristotle-ethics/.

[28] Tasioulas, "First Steps."

to human decisions of purely moral weight, such as those examined in section 3, the 'quasi-moral' behaviour of the agent[29] becomes relevant in determining whether it can be entrusted with high-risk responsibilities.

At this point, it should be stressed that a virtue-ethical approach to using AI in high-risk contexts not only requires AI agents to be virtuous themselves. It also requires virtuous human users of AI systems. This is relevant, for example, in the context of the virtuous use of AI weapons.[30] [31] The cultivation of a virtuous use could be respectively achieved by the cultivation of virtue in the human users themselves through their interaction with virtuous AI agents. In this case, the latter will perform a significant social task, becoming the means for an exercise of the human users' character (we return to a further analysis of this idea in section 5 of the present text).

This process transcends mere computational abilities, incorporating the expression of virtues within the broader social context. The proposed framework suggests that an ethical agent cannot be understood without its societal role, highlighting its organic function within a symbiotic system.[32] In such a system, practical machine learning occurs, knowledge is continuously acquired, and a feedback loop of virtuous behaviour is established – without relying on pre-loaded values. This "aretological" approach solves ontological, epistemological and other 'value loading' challenges by prioritising behavioural outcomes over internal functional metrics. It thereby opens the theoretical possibility of considering machines as virtuous agents whose contributions to both social and individual goals are evaluated based on their cumulative achievements throughout their existence.

A virtuous agent, in the context of AI ethics, embodies several key characteristics that align with Aristotelian virtues and emphasise the system's integration into social contexts. One fundamental attribute is the ability to learn by doing. This involves learning through practice, where AI systems acquire knowledge and refine their behaviour through iterative processes. Such a dynamic approach enables them to adapt to changing contexts and improve their functionality over time. Another essential characteristic is the capacity for meaningful social interaction. Virtuous AI systems must actively participate in societal networks, drawing on these interactions to align their behaviour with the values and expectations of the communities they serve. This social embeddedness is crucial for fostering trust and ensuring that their actions resonate with the needs of society. Furthermore, virtuous agents must

---

[29] Alkis Gounaris, "Can We Literally Talk About Artificial Moral Agents?" 2020.

[30] Henrik Syse and Martin Cook, "Robotic Virtue, Military Ethics Education, and the Need for Proper Storytellers," *Conatus – Journal of Philosophy* 8, no. 2 (2023): 667-680.

[31] Nigel Biggar, "An Ethic of Military Uses of Artificial Intelligence: Sustaining Virtue, Granting Autonomy, and Calibrating Risk," *Conatus – Journal of Philosophy* 8, no. 2 (2023): 67-76.

[32] Joseph C. R. Licklider, "Man-Computer Symbiosis," *IRE Transactions on Human Factors in Electronics* HFE-1, no. 1 (1960): 4-11.

maintain neutrality concerning predefined moral principles, commonly referred to as 'prima facie values and duties' as described above. Unlike traditional approaches that impose a priori moral obligations and are ultimately linked to the 'value loading problem,' virtuous agents should prioritise fulfilling their specific individual and societal purposes instead. This neutrality allows them to develop moral qualities organically through their actions and integration into their social environments rather than being constrained by rigid ethical models.

To evaluate the success of AI systems as virtuous agents, it is imperative to address two fundamental questions: What is their purpose, and what role do they play in society? Defining their purpose involves identifying their intended contributions, whether in solving problems, enhancing efficiency, or fostering innovation. Meanwhile, understanding their societal role requires assessing how they integrate into existing frameworks and address broader needs, such as social cohesion, fairness, and individual well-being. AI systems can evolve into virtuous agents by embodying these characteristics and aligning them with their defined roles and purposes. This evolution focuses on interaction with behaviour rather than static ethical models, emphasising adaptability, learning, and meaningful contributions to society.

The measure of their success lies not in compliance with fixed principles but in their ability to accomplish their work and, at the same time, to foster both human fulfilment and community flourishing (*eudaimonia*) throughout their lifecycle.[33] This externalist, virtue-based perspective offers a robust framework for addressing the ethical challenges posed by AI development. However, while this externalist virtue-based model is suitable for the evaluation of limited or minimal-risk AI systems, a what-if scenario can exemplify why its application to high-risk AI systems should be avoided. In these cases, a priori moral values should be taken seriously and applied accordingly in these systems' design, operation and use.

IV. What-if scenario: Applying Aristotelian ethics in 'high-risk' AI design, operations, and use

We argue that high-risk AI systems rely on complex factors and cannot be addressed "horizontally" through a general regulatory framework. Instead, special criteria need to be established for the various phases of their design, operation and use.

An illustrative example can be found in AI systems acting as tutors or university faculty members. Assigning such roles to AI entities meets the two fun-

[33] See Sandel, Chapter 8, where he discusses Aristotle's example of the flute. Sandel argues that granting the best flute to the best flutist realises three convergent ends: the instrument's telos (producing excellent music), the musician's personal fulfillment, and the community's overarching good. According to Sandel's interpretation of Aristotle, it is precisely this alignment of individual and collective purposes that constitutes eudaimonia. In such a case, the instrument itself can be seen as participating in virtue, insofar as its proper use fulfills its own function while simultaneously contributing to the flourishing of both the individual user and the wider community.

damental criteria of a virtuous agent: achieving a defined individual goal and contributing to a broader social purpose. The individual goal of the AI system could involve conducting original research in its field or mentoring students. Its social purpose would focus on promoting learning and research or improving the efficiency and optimisation of time dedicated to both teaching and research activities. However, despite being designed with virtue-based principles or performing well in evaluations of its operation, the design, operation, and even the use of such a system might ultimately fail to be truly 'virtuous.' For instance, such a system could be misused to disseminate propaganda, facilitate academic dishonesty – such as enabling students to cheat – or advance the interests of specific social groups within the educational sector.

Since the ethical issues associated with the use of high-risk agents cannot be exhaustively anticipated and the application of the concept of 'moral virtue' fails to adequately extend to such agents, designing these systems based on aretological criteria becomes particularly problematic. The core challenges that arise are typically related to issues of autonomy, freedom of choice, and cognitive limitations, which render these systems incapable of fully understanding their functional roles. The absence of autonomy and freedom of choice in AI systems – especially in high-risk areas such as border control – shows that Aristotelian ethics alone may not be sufficient to address the complex moral dilemmas that arise in these contexts.

In high-risk systems, such as those related to migration, asylum, and border control, it is crucial to determine the system's ultimate purpose, or *telos*,[34] in advance. AI lacks the capacity for human-like deliberation regarding its ultimate objectives,[35] which brings the 'value loading' problem back into focus.

When applying the virtuous agent model to migration policies, such as asylum procedures or border control, it is critical to assess whether it can promote just systems. These systems involve significant ethical and legal complexities, impacting fundamental rights and freedoms. This raises the question: Can a virtuous agent model produce fair outcomes, or does it face limitations in high-risk operations?

As we have emphasised, a virtuous agent, according to Aristotelian ethics, should promote both individual and societal flourishing (*eudaimonia*). However, in high-risk systems, operational effectiveness alone is insufficient. Specific criteria must also be ensured, such as the protection of vulnerable individuals, legal compliance, and the safeguarding of human rights. In the context of migration policies, AI systems can ideally expedite administrative procedures, such as asylum applications, or detect fraud and identify vulnerable individuals. At the same time, however, concerns arise regarding privacy protection, discrimination, and injustices that may have profound impacts on human lives.

---

[34] Aristotle, *The Nicomachean Ethics*.

[35] Tasioulas, "First Steps."

The objectives of such systems may include legal compliance, border security, or facilitating access to asylum. However, these individualized goals are shaped by broader social and political directives and decisions, which precondition both the effectiveness and ethical behavior of these systems.

In Aristotelian ethics, as mentioned above, virtues are cultivated through action, with an emphasis on the development of moral character and practical wisdom (*phronesis*). However, AI systems lack the capacity for moral reasoning and the ability to navigate complex, context-dependent ethical dilemmas, such as those encountered in migration management. Unlike human agents, AI systems do not possess the cognitive and emotional capabilities necessary for ethical deliberation concerning the consequences of their decisions on vulnerable populations. Thus, the application of virtues such as justice, courage, and temperance[36] in these systems becomes problematic. The "virtue" of artificial intelligence is ultimately based on pre-defined criteria rather than genuine moral reasoning.

Moreover, applying the virtuous agent model to high-risk AI systems in the areas of migration and asylum control overlooks the political and value-laden nature of these fields. The use of AI in this context is inherently tied to societal debates on issues such as open versus closed borders, migration, human rights, and the diverse values associated with different approaches to political theory.[37] These issues are subject to shifting political agendas and public opinion.

In practice, such systems may prove inadequate in addressing these contentious issues, as their social responsibilities – such as fraud detection or enhancing security – may conflict with goals related to justice and human rights. For example, even within liberal theory, there are divergent approaches, with some liberals arguing for open borders and freedom of movement as a central element of human life planning,[38] while other liberal approaches may suggest that in an idealised 'realistic utopia,' forced migration, in particular, would be eliminated,[39] or they may raise concerns about the divergent political principles of different communities,[40] echoing issues raised by communitarians.[41]

---

[36] Andrew P. J. Mullins, "What Does Self-control Look Like? Considerations About the Neurobiology of Temperance and Fortitude," *Conatus – Journal of Philosophy* 10, no. 1 (2025): forthcoming.

[37] Maria-Artemis Kolliniati, *Interpreting Human Rights: Narratives from Asylum Centers in Greece and Philosophical Values* (Routledge, 2024).

[38] Joseph Carens, "Migration and Morality: A Liberal Egalitarian Perspective," in Free Movement: Ethical Issues in the Transnational Migration of People and Money, eds. B. Barry and R. Goodin, 25-47 (Harvester Wheatsheaf, 1992).

[39] John Rawls, *The Law of Peoples* (Harvard University Press, 2002).

[40] John Rawls, *A Theory of Justice* (Harvard University Press, 1999).

[41] Joseph Carens, *The Ethics of Immigration* (Oxford University Press, 2013), as cited in Kolliniati, *Interpreting Human Rights*.

Given these factors, applying the virtuous agent model to such systems requires modifications that incorporate pre-defined values. However, the regulatory and politically charged environment in which they operate complicates the development of ethically sustainable systems. Without human moral judgment, the risk of discrimination against vulnerable groups, such as asylum seekers, is heightened. Consequently, Aristotelian virtues are insufficient for the design, training, and operation of these systems, as they are likely to encounter conflicting dilemmas.

On the one hand, the goal of maintaining national sovereignty and protecting the local community calls for closed borders for asylum seekers. Virtues such as prudence, responsibility, and justice prioritise the well-being and security of the state and its citizens.

On the other hand, international legal obligations, such as human rights treaties, advocate for open borders to ensure the fair assessment of asylum claims. Virtues such as compassion and respect for human dignity support the protection of vulnerable individuals fleeing persecution.

Aristotelian ethics, emphasising achieving a final purpose (*telos*) through virtue, struggles to resolve such dilemmas. The two objectives – protecting the local community and upholding international obligations – can both be considered virtuous but often come into conflict. For instance, the virtue of prudence might favour closed borders to safeguard security, while the virtue of compassion demands open borders for humanitarian reasons. Aristotle's concept of practical wisdom (*phronesis*) suggests that virtuous actions should be context-sensitive and aim for balance. However, the competing virtues in this case offer no clear solution. The tension persists, as reconciling national sovereignty with the fulfilment of international law obligations proves difficult through Aristotelian ethical balancing.

At this point, a broader problem inherent in Virtue Ethics emerges. In order to explain why a particular trait qualifies as a virtue or to prioritize one virtue over another in situations of moral conflict, Virtue Ethics must often appeal to concepts and criteria from other ethical frameworks, such as ethical egoism or social contract theory.[42]

The dilemma presented here illustrates the conflict between protecting the well-being of the local community – defensible under a version of social contract theory – and the right to human dignity, even for asylum seekers or those crossing borders illegally – defensible through rights-based or Kantian approaches.

This incompatibility highlights the incompleteness of Virtue Ethics, which compels us – in the context of high-risk AI systems – to adopt prima facie moral values. This, in effect, undermines the value of aretology and simultaneously leads us back to the value loading problem that we sought to avoid with our virtue-based

[42] James Rachels, *The Elements of Moral Philosophy* (McGraw-Hill, 2015), 172.

approach in this paper. By contrast, in low-risk and minimal-risk contexts, moral dilemmas requiring strict prioritisation of ethical principles do not typically arise.

The inability to apply this model to high-risk AI systems, such as those used in migration and border control, reveals inherent issues with these systems, which can be summarised in six additional points. First, they do not understand – or it remains uncertain whether they understand – the concept of morality.[43] Second, under current conditions, they cannot be regarded as 'moral persons,' and as such, they cannot bear responsibilities or be held accountable.[44] Third, they are incapable of voluntarily cultivating virtues, as they are constrained by their objectives and, therefore, cannot freely err. Fourth, they cannot demonstrate equity. This is due to the 'frame problem' of AI, namely the fact that systems are programmed by finite programs, or trained by a finite number of examples (e.g. in the case of artificial neural networks), and therefore there are cases for which they do not have all the critical information, and thus their behaviour in these cases is 'rigid.' Nevertheless, this problem could potentially be overcome if AI systems were given the capacity for true understanding. However, as mentioned above, it is highly doubtful that they have the capacity for true understanding due to the 'other minds problem.'[45]

Fifth, learning by example in reinforcement learning systems could, under certain conditions, be considered analogous to the continuous life experience that leads to the development of *phronesis* (practical wisdom). However, a reasonable question is, "How long should this process of experience and refinement take place before an AI entity can reach the level of *phronesis*?" For humans, this is often a lifelong process. However, when it comes to AI systems deployed in high-risk contexts, a continuous self-improvement process without clear time boundaries would be difficult to accept. Therefore, even if we consider learning by example in deep learning systems as a sufficient analogy to the lifelong experience that leads to *phronesis*, the temporal indeterminacy of achieving *phronesis* is something we allow for humans – since life experience and the accumulation of knowledge are continuous – but not for AI systems, particularly when these are intended to operate in high-risk contexts. Moreover, the absence – or our inability to verify – of key cognitive characteristics in AI entities that are prerequisites for *phronesis*, such as moral sensitivity and moral attentiveness, further heightens our reluctance to adopt a virtue ethics model for AI agents operating in high-risk environments.

---

[43] Gounaris and Kosteletos, "Writing the Algorithm of Good."

[44] Alkis Gounaris and George Kosteletos, "Licensed to Kill: Autonomous Weapons as Persons and Moral Agents," in *Personhood*, eds. D. Prole and G. Rujević, 137-189 (The NKUA Applied Philosophy Research Lab Press, 2020).

[45] Anita Avramides, "Other Minds," *The Stanford Encyclopedia of Philosophy* (Winter 2023 Edition), eds. Edward N. Zalta and Uri Nodelman, https://plato.stanford.edu/archives/win2023/entries/other-minds/.

Additionally, in the same context, even meeting operational criteria does not necessarily imply the 'virtuous use' of high-risk AI systems. A virtue, on its own, is insufficient to produce ethical behaviour. For example, HAL 2000's commitment to ensuring the mission's safety led to catastrophic decisions for the spacecraft crew. A virtue detached from the concept of a 'moral person' can also enable unethical or unlawful behaviours. For instance, the virtue of courage can embolden a criminal or a terrorist. Aristotle argues that the misuse of virtue for harmful purposes is prevented by its combination with *phronesis* (practical wisdom). He distinguishes between perfect virtue and natural virtue, the latter being a primitive version of perfect virtue – an early form shaped merely by predisposition and emotion (as often seen in children) rather than by rational choice and *phronesis*. Children, as well as adults who, despite good intentions, fail to help others unintentionally, lack *phronesis* or practical wisdom. They fail either because they do not know what is necessary to implement their good intentions or because they do not correctly recognise what is beneficial or harmful to others. Thus, *phronesis* requires specific cognitive skills, such as the ability to evaluate which features of a particular situation are most significant from a moral standpoint. In this sense, *phronesis* presupposes the presence of intellectual capacities such as moral sensitivity, moral attentiveness, and moral imagination.[46] AI systems do not possess – or, due to the 'other minds problem,' we cannot ascertain whether they possess – such cognitive traits.

Sixth, in such contexts, it is crucial to define the system's ultimate purpose, or '*telos*'[47] in advance, given that AI lacks the capacity for human-like deliberation on its end goals.[48]

However, despite the fact that Aristotelian ethics can lead to several pitfalls in the design of high-risk AI systems, the conclusion is different when it comes to the *evaluation* of such systems. In the next chapter, we examine the idea of adopting a virtue ethics approach to the evaluation of high-risk AI systems.

V. Towards an Aristotelian evaluation method: Adopting virtue-based criteria for the assessment of low, medium and high-risk AI systems

Given the classification framework of the EU AI Act, which categorises AI systems into various risk zones, our analysis aims to demonstrate through hypothetical scenarios that while characterising an AI system as 'virtuous' is feasible in low- or medium-risk environments, Aristotelian ethics is not an appropriate

---

[46] Rosalind Hursthouse and Glen Pettigrove, "Virtue Ethics," *The Stanford Encyclopedia of Philosophy* (Fall 2023 Edition), eds. by Edward N. Zalta and Uri Nodelman, https://plato.stanford.edu/archives/fall2023/entries/ethics-virtue/.

[47] Aristotle, *The Nicomachean Ethics*.

[48] Tasioulas, "First Steps."

framework for guiding the design, construction, and training of AI systems involved in high-risk activities. However, Aristotelian ethics is perfectly applicable to the evaluation of high-risk systems. In particular, it can serve as a valuable reference for establishing evaluation criteria for the operation and use of such systems. By distinguishing between the phases of design, operation, and use, it is argued that virtue-based criteria can indeed be applied, according to the following Table 1.

Application of Virtue-Based Criteria	Can virtue-based criteria be introduced during the Design, Development, and Training Phase (virtuous by design – a priori assessment)	Can virtue-based criteria be introduced for evaluating the Deployment and Operation Phase (a posteriori virtuous assessment)	Can virtue-based criteria be introduced to evaluate the use of the systems?
High-Risk Zone	NO	PROBABLY YES	YES
Medium -Risk Zone	YES	YES	YES
Low -Risk Zone	YES	YES	YES

Table 1.

In other words, we argue that in the hypothetical scenario where such high-risk systems are developed and deployed, they can still be evaluated in use (post ex or a posteriori virtuous assessment) using Aristotelian criteria based on their behaviour. Specifically, they can be assessed through their outcomes and behaviour within a regulatory sandbox to determine whether they meet the conditions to be characterized as 'virtuous' agents or systems. As controlled environments, regulatory sandboxes offer flexibility for experimentation under real-world conditions, allowing regulatory bodies to evaluate system behaviour before full implementation while maintaining strict oversight and control.[49] This process reduces the risk of unintended consequences, promotes compliance with human rights standards, and addresses issues

[49] Dirk A. Zetzsche, Ross P. Buckley, Janos N. Barberis, and Douglas W. Arner, "Regulating a Revolution: From Regulatory Sandboxes to Smart Regulation," *Journal of Corporate and Financial Law* 23, no. 1 (2017): 31-103.

such as algorithmic bias. Within this context, the systems can be examined to determine whether they meet the criteria of virtue and justice prior to their broader deployment.

In such environments, evaluation criteria based on virtuous behaviour and respect for fundamental values can be introduced. During the operation of high-risk AI systems, such as those in migration and asylum applications, evaluators can monitor the system's interactions with vulnerable populations using indicators such as fairness in decision-making and the avoidance of biases. For example, the system can be assessed on its ability to recognize cases that require exceptions to the rule, thereby demonstrating equity, a key virtue in Aristotelian ethics. Additionally, *phronesis* (practical wisdom) can serve as a criterion for the system's adaptability to different ethical and legal frameworks. The 'a posteriori' assessment of such systems can function as a feedback-loop refinement,[50] enabling continuous improvement to better align with ethical principles and regulatory requirements.

An AI system or agent involved in decision-making for asylum applications should, for example, demonstrate transparency by providing justifications for each rejection and allowing for appeals and revisions. The virtue of responsibility demands accountability, which can be achieved through regular performance evaluations and the implementation of reporting mechanisms for errors or discriminatory practices. Thus, the sandbox can serve as a feedback and evaluation environment for high-risk systems using criteria grounded in virtue ethics.

Table 2 below describes the relationship between various aspects of moral virtues and AI systems. It compares moral virtues in terms of their connection to reason, freedom of choice, and their role in achieving *eudaimonia* – that is, human flourishing. The table also explores how these virtues can manifest in AI, particularly in high-risk applications, such as border control systems for asylum evaluation.

It is emphasized that while AI itself may not possess moral virtues, it can exhibit moral behaviour if correctly designed, programmed, and trained. The table further highlights the importance of the EU AI Act in regulating these systems, with a specific focus on justice, transparency, and human rights in high-risk zones.

[50] For the reinforcement learning problem, see Richard S. Sutton and Andrew G. Barto, *Reinforcement Learning: An Introduction* (The MIT Press, 2018).

Virtue category	Relationship with Reason	Relationship with Freedom of Choice	Prerequisite for Eudaimonia	AI Example	Connection to EU AI Act and Risk Zones
Ethical Virtues	Guided by practical wisdom (*phronesis*)	Require freedom of choice	Yes, for humans	AI may not possess ethical virtue due to ontological, epistemological and efinitional limitations (as described above)	**High-Risk Zone:** AI systems for border control (e.g., asylum claim assessment). Must ensure fairness, and human rights, balancing conflicting societal interest
Quasi-Moral Virtues	Requires complex reasoning, depends on predetermined purpose and design based on value loading	No freedom or deliberation	No, but meets instrumental and functional excellence	AI that complies with moral principles and executes its designed role excellently.	**High-Risk Zone:** AI used in automated border control, ensuring compliance with international law and preventing bias in decision-making. Must adhere to rustworthiness criteria.
Intellectual Virtues	Requires complex reasoning	Not dependent on freedom	Yes, for humans. No, for AI Agents but meets instrumental and functional excellence	AI can solve problems, explain reasoning, analyse data, simulating science or prudence.	**Medium-RiskZone:** AI for detecting fraud in migration data or/ and providing recommendations - suggestions for asylum eligibility. Aligns with legal frameworks.
Functional Virtues	Requires high intelligence, Influenced by reason but not fully	Not dependent on freedom	No, but meets instrumental and functional excellence	AI coordinating processes within a society or organisation based on moral principles and logic.	**Medium-Risk Zone:** AI managing asylum processes, ensuring fair treatment and legal compliance with transparency and non-discrimination

Instrumental Virtues	Do not involve reason	Do not require freedom	No, but meets nstrumental excellence	Reactors in AI that function correctly to collect data. Systems working optimally for its purpose, e.g., a chatbot answering questions correctly.	**Low-Risk Zone:** AI systems in administrative tasks (e.g., virtual assistants for immigration queries). AI systems for biometric data collection at border points.

Table 2.

Therefore, based on the above, even if the application of virtue-based criteria is insufficient for the creation and design of high-risk agents, once such agents are deployed – regardless of how they are constructed – the establishment of virtue-based criteria for evaluating the operation and use of high-risk AI systems becomes imperative. Such an application could have immediate practical value. The evaluation of machines through an Aristotelian virtue assessment system could lead to a process of certifying AI systems based on the virtues they exhibit. In this context, we propose the development of a 'Seal of Excellence' based on Aristotelian virtues. This is described in the next section of this paper.

VI. A Seal of Excellence for AI: From virtuous systems to virtuous users

A method of evaluation checks, such as the one presented in the previous section, could also serve as a general guide for system developers during the design, development, and training phases for medium- or low-risk systems or agents, as well as during deployment, operation, and use across all risk categories. We propose that this assessment method would also be highly beneficial for users and policymakers.

Inspired by the successful European Commission's model, we propose a Virtue-Based Seal of Excellence Certificate based on Aristotelian criteria tailored to virtue categories, use, and risk zones. The existing EU Seal of Excellence is awarded to project proposals that meet the high-quality standards of EU funding programmes, certifying their excellence and enabling access to alternative funding sources.[51] It enhances project credibility, attracts investment at multiple levels, and ensures security through digital sealing.[52]

[51] European Commission, *Seal of Excellence*, https://commission.europa.eu/funding-tenders/find-funding/seal-excellence_en.

[52] European Commission, *How Can Seal Holders Use the Seal of Excellence?* European Com-

Our approach extends this concept to AI by embedding ethical evaluation into system design, deployment, and governance. We propose a certification process whereby a "Seal of Excellence" for virtuous AI systems or agents serves as a mark of distinction, signifying that a system demonstrates virtuous behaviour. In practical terms, this entails not only fulfilling its specific design objectives but also adhering to ethical standards that contribute positively to broader societal aims. In this sense, the certification of virtuous agents could function as a bridge between philosophy, social imperatives, and system design – integrating these domains in a concerted effort to ensure that AI development remains fully aligned with human values and goals.

The Seal of Excellence for AI could be considered best practice, as, from a virtue ethics perspective, an AI system or agent could even be awarded or decorated in a form of commendation, analogous to the decoration of animals for their service. Notably, the PDSA Dickin Medal has been awarded to animals, particularly dogs, in recognition of their bravery and contributions in military conflicts. A recent example is the commendation of Diesel, a police dog honoured posthumously after being killed in action during an anti-terror operation in Paris in 2015.[53]

Importantly, this would also serve an instructive and exemplary function for human users of AI. Discussions on AI ethics often emphasise the goal of responsible use. However, from a virtue ethics perspective, the notion of responsible use -and consequently, its objective- could be reframed as virtuous use. It is essential to cultivate the character of users so that they engage with AI in a virtuous manner. A society that achieves *eudaimonia* through AI does not merely require virtuous AI systems; it also demands virtuous users capable of interacting with them in ethically sound ways. More broadly, insofar as human life is increasingly intertwined with AI, a virtuous life – one that ensures *eudaimonia* – must encompass virtuous interaction with AI and, consequently, its virtuous use. Furthermore, if the virtuous life necessitates *phronesis* (practical wisdom), and *phronesis* itself is cultivated through lived experience and learning by example, then AI systems with which humans interact function both as exemplars and as integral components of human experiential learning. Considering the limitations of Tables 1 and 2 (above), by designing AI systems as virtuous agents – and even more so by conferring distinctions upon them – we effectively establish them as paradigms of virtue from which human users can derive practical wisdom. As AI increasingly permeates daily life, the ethical exemplars surrounding us will no longer pertain

mission, 2021, https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/seal-excellence/how-can-seal-holders-use-seal-excellence_en.

[53] PDSA, *PDSA Dickin Medal*, https://www.pdsa.org.uk/what-we-do/animal-awards-programme/pdsa-dickin-medal.

solely to traditional aspects of human existence but will extend to our interaction with AI. Consequently, our engagement with AI systems will shape a significant part of our moral development and real-world conditioning. If this interaction involves engagement with virtuous AI systems or agents, then we will have reinforced the presence of virtuous exemplars in our environment. In fact, the cultivation of virtuous use can take place at two different levels of user consciousness: a) a fully conscious level at which users perceive and accept AI systems as role models because of the AI Seal of Excellence that these systems carry, and b) a less conscious – perhaps even unconscious – level at which users' daily interaction with virtuous AI systems – even with virtuous AI systems that have not yet received the AI Seal of Excellence – inevitably shapes users' characters in a virtuous way. In this case, it is the AI systems that set the tone for their interaction with humans. This interaction inevitably takes place in virtuous contexts because of the virtuous nature of the AI systems themselves. Thus, everyday interaction with AI becomes a process of habituation (i.e. a less conscious process) through which users acquire virtue. In this case, it's not the conscious process of modelling, but the less conscious – or even automatic – process of being molded by everyday practice. Through AI, we will have created an ecosystem that, at least in its technological dimension, cultivates virtue in human users, guiding them towards moral excellence, shaping their character, and fostering the development of *phronesis*. This perspective could serve as a response to the legitimate concerns that human interaction with AI might erode their virtues.[54] On the contrary, engagement with virtuous AI has the potential to strengthen them. Conclusively, a virtuous AI system or agent could explicitly motivate, educate and/or even demand its optimal use by the user or implicitly lead the end user in such use. By analogy with Aristotle's flute's instrumental virtue (see footnote 34 above), virtuous AI could be a flute, making the flutist a better musician.

The AI Seal of Excellence embodies a transformative approach to ethical AI, one that transcends mere harm mitigation and aspires towards moral cultivation. By recognising AI systems that exemplify virtue – promoting *phronesis*, justice, beneficence, honesty, and social virtues – the certification does more than validate ethical compliance; it establishes AI as an active agent in shaping human morality. This perspective challenges the prevailing concerns that AI may erode human virtues, suggesting instead that well-designed AI can reinforce them. Just as virtuous AI inspires virtuous users, a society shaped by AI engagement must prioritize both ethical system design and ethical user development. By institutionalizing this vision through thorough evaluation, public engagement, and adaptive governance, the Seal of Excellence for AI

[54] Nir Eisikovits and Dan Feldman, "AI and Phronesis," *Moral Philosophy and Politics* 9, no. 2 (2022): 181-199.

represents a significant shift. It moves from AI systems that simply follow ethical guidelines to those that actively promote moral engagement, creating a beneficial cycle between technology and ethics.

VII. Conclusion

In this paper, we have argued for an innovative application of Aristotelian virtues (*arêtes*) as qualities of excellence and a key notion to Artificial Intelligence ethics, mindful of the diverse risk categories delineated in the EU AI Act. At the heart of this discussion lies the realisation that the one-size-fits-all approach of the AI Act – rooted in 'ethical data' and 'prima facie' or 'a priori' values – has intrinsic limitations when extended to all-risk systems and in particular to those that do not typically face irreconcilable moral dilemmas. As we have demonstrated, an AI system can indeed exhibit forms of virtue, particularly in low- or medium-risk settings, by achieving functional excellence and serving a defined purpose (*telos*). Nonetheless, it lacks the freedom of choice that, in Aristotelian thought, is a prerequisite for genuine moral virtue. Even if an AI system is 'loaded' with moral values for decision-making, it does not develop moral will per se; it relies on preexisting values set by its designers. The question of authentic consciousness or internal volition, moreover, foregrounds the 'other minds problem.'[55] We have no definitive way of confirming whether an AI truly 'feels' or 'thinks' in a human-like manner. In this light, although a virtue-based perspective offers a useful methodology for evaluating system behaviour, the strictly moral dimension of virtue (which depends on freedom and practical wisdom) is challenging to replicate in high-risk applications. That is why, as we have emphasised, virtue ethics is most appropriate mainly as an external, ex post (or a posteriori) evaluation criterion, rather than as a foundation for the initial design of AI systems in critical domains such as border management or judicial procedures.

Through an external behaviour-based evaluation, both a human cognitive system and a high-risk AI system may be assessed as capable of achieving their final purpose. However, simply arriving at the desired outcome does not demonstrate the presence of *phronesis* (practical wisdom). Indeed, a system can display computational virtue and operational and/or instrumental excellence without possessing the moral imagination or sensitivity that *phronesis* presupposes. Computational virtue, which is adequate for complex reasoning and problem-solving procedures, is insufficient for genuine Aristotelian *phronesis*, as features such as moral imagination, moral sensitivity, and other cognitive capacities remain unverified in current AI systems. Additionally, while humans often require a lifetime to develop *phronesis*, AI systems are 'born'

[55] Avramides.

fully developed;[56] there is no clear developmental arc in which they gradually cultivate moral discernment. As a result, even if they fulfill key goals effectively, we cannot justify calling them quasi-moral agents.

At last, the pre-loading of values and the predefinition of final purposes by system designers – since, as Tasioulas notes, AI entities lack the capacity for human-like deliberation – leads us to refer to this as quasi-*phronesis*. These systems may mimic some elements of virtuous conduct but cannot autonomously choose, revise, or weigh moral ends in the robust sense implied by Aristotelian virtue.

In conclusion, as we have demonstrated, virtue ethics encounters formidable challenges when one attempts to build or train AI systems in high-risk sectors – including those that affect migration, asylum, border control, healthcare, education, and justice. In these domains, conflicting (normative) objectives arise frequently, often hinging on complex legal, societal, or humanitarian considerations that demand immediate and pre-specified moral imperatives. While virtues encourage context-sensitive discernment, contemporary AI systems cannot replicate the kind of *phronesis* (practical wisdom) that is central to Aristotelian thought. As we emphasised, lacking genuine autonomy and freedom of choice, such systems are ill-equipped to engage in genuine moral deliberation. For this reason, we proposed that an Aristotelian model is unsuitable for *designing* and *training* these high-risk systems. However, Aristotelian ethics retains value in *assessing* how high-risk systems perform and are used post-deployment. Adopting a virtue-based *a posteriori evaluation* method, if needed within regulatory sandboxes, enables policymakers and researchers to observe whether AI systems or agents uphold fairness, mitigate bias, and promote collective *eudaimonia* in the sense of fulfilling social objectives. This form of dynamic, behaviour-oriented oversight aligns with the notion that AI's real-world performance should be judged not only by technical metrics but also by the social outcomes it produces. Even in domains where moral dilemmas are acute, tracking whether a system's operation demonstrates virtuous behaviour helps identify potential improvements and fosters user trust.

In this direction, we have proposed an Aristotelian *evaluation method* that adopts virtue-based criteria for assessing low, medium, and high-risk AI systems. This method leads to the Virtuous AI system or agent's acknowledgement, certification and 'decoration.' As a central contribution, the virtue-based "AI Seal of Excellence" underscores how a series of criteria can

[56] For a defence of the opposing view, i.e., the position that *phronesis* can develop in AI systems, see John P. Sullins, "Automated Ethical Practical Reasoning: The Problem of Artificial Phronesis," in *Robophilosophy: Philosophy of, for, and by Social Robotics*, eds. J. Seibt, R. Hakli, and M. Nørskov (MIT Press, 2025), forthcoming.

serve as a constructive framework for both developers and users across diverse risk levels. Inspired by existing European certification models, we propose an 'AI Seal of Excellence' which awards AI systems that exhibit virtuous behaviour over time, as measured against clearly defined operational excellence and social benefit thresholds. Recognising such systems publicly would not only motivate industry-wide adherence to higher ethical standards but also encourage a reciprocal dynamic in which virtuous AI acknowledgement fosters virtuous user practices. From a broader philosophical view, involving both designers and users in an ongoing effort around virtue cultivation holds promise for aligning AI's expanding role in society with human flourishing. Insofar as future work can integrate philosophical depth with technological sophistication, the Aristotelian paradigm may help create AI that does more than merely minimize harm, instead contributing positively to the shared pursuit of *eudaimonia*.

Author contribution statement

All authors have contributed equally to the conception and design of the work, the drafting and revising of the manuscript, and the final approval of the version to be published.

References

Anderson, Michael, and Susan Leigh Anderson. "Machine Ethics: Creating an Ethical Intelligent Agent." *AI Magazine* 28, no. 4 (2007): 15.

Anderson, Michael, Susan Leigh Anderson, Alkis Gounaris, and George Kosteletos. "Towards Moral Machines: A Discussion with Michael Anderson and Susan Leigh Anderson." *Conatus – Journal of Philosophy* 6, no. 1 (2021): 177-202.

Aristotle. *On the Soul*. Translated by J. A. K. Thomson. Harvard University Press, 1959.

Aristotle. *The Nicomachean Ethics*. Edited by L. Brown. Translated by D. Ross. Oxford University Press, 2009.

Avramides, Anita. "Other Minds." *The Stanford Encyclopedia of Philosophy* (Winter 2023 Edition), edited by Edward N. Zalta and Uri Nodelman, https://plato.stanford.edu/archives/win2023/entries/other-minds/.

Biggar, Nigel. "An Ethic of Military Uses of Artificial Intelligence: Sustaining Virtue, Granting Autonomy, and Calibrating Risk." *Conatus – Journal of Philosophy* 8, no. 2 (2023): 67-76.

Cappuccio, Massimiliano, Eduardo Sandoval, Omar Mubin, Mohammad Obaid, and Mari Velonaki. "Can Robots Make Us Better Humans? Virtuous Robotics and the Good Life with Artificial Agents." *International Journal of Social Robotics* 13 (2021): 7-22.

Carens, Joseph. "Migration and Morality: A Liberal Egalitarian Perspective." In *Free Movement: Ethical Issues in the Transnational Migration of People and Money*, edited by B. Barry and R. Goodin, 25-47. Harvester Wheatsheaf, 1992.

Carens, Joseph. *The Ethics of Immigration*. Oxford University Press, 2013.

Dworkin, Ronald. *Taking Rights Seriously*. Harvard University Press, 1978.

Eisikovits, Nir, and Dan Feldman. "AI and Phronesis." *Moral Philosophy and Politics* 9, no. 2 (2022): 181-199.

European Commission. *Ethics Guidelines for Trustworthy AI*. Office for Official Publications of the European Communities, 2019.

European Commission. *How Can Seal Holders Use the Seal of Excellence?* European Commission, 2021. https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/seal-excellence/how-can-seal-holders-use-seal-excellence_en.

Gibert, Martin. "The Case for Virtuous Robots." *AI and Ethics* 3 (2022): 135-144.

Gounaris, Alkis, and George Kosteletos. "Writing the Algorithm of Good: Artificial Intelligence as a Machine of Justice." *Ithiki* 19 (2024): 6-27 [in Greek].

Gounaris, Alkis. "Can We Literally Talk About Artificial Moral Agents?" 2020.

Gounaris, Alkis, and George Kosteletos. "Licensed to Kill: Autonomous Weapons as Persons and Moral Agents." In Personhood, edited by Dragan Prole and Goran Rujević, 137-189. The NKUA Applied Philosophy Research Lab Press, 2020.

Henry, Nathan I. N., Mangor Pedersen, Matt Williams, Jamin L. B. Martin, and Liesje Donkin. "A Hormetic Approach to the Value-Loading Problem: Preventing the Paperclip Apocalypse." *arXivLabs* (2024), https://arxiv.org/abs/2402.07462.

Hall, Joshua M. "Just War contra Drone Warfare." *Conatus – Journal of Philosophy* 8, no. 2 (2023): 217-239.

Hursthouse, Rosalind, and Glen Pettigrove. "Virtue Ethics." *The Stanford Encyclopedia of Philosophy* (Fall 2023 Edition), edited by Edward N. Zalta and Uri Nodelman, https://plato.stanford.edu/archives/fall2023/entries/ethics-virtue/.

Kolliniati, Maria-Artemis. *Interpreting Human Rights: Narratives from Asylum Centers in Greece and Philosophical Values*. Routledge, 2024.

Kraut, Richard. "Aristotle's Ethics." *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition), edited by Edward N. Zalta and Uri Nodelman, https://plato.stanford.edu/archives/fall2022/entries/aristotle-ethics/.

Lekea, Ioanna, George Lekeas, and Pavlos Topalnakos. "Exploring Enhanced Military Ethics and Legal Compliance through Automated Insights: An Experiment on Military Decision-making in Extremis." *Conatus – Journal of Philosophy* 8, no. 2 (2023), 345-372.

Licklider, Joseph C. R. "Man-Computer Symbiosis." *IRE Transactions on Human Factors in Electronics* HFE-1, no. 1 (1960): 4-11.

Livieri, Georgia, Eleni Mangina, Evangelos D. Protopapadakis, and Andrie G. Panayiotou. "The Gaps and Challenges in Digital Health Technology Use as Perceived by Patients: A Scoping Review and Narrative Meta-synthesis." *Frontiers in Digital Health* 7 (2025): 1474956.

Long, David. *The Animals' VC: For Gallantry and Devotion: The PDSA Dickin Medal-inspiring Stories of Bravery and Courage*. Random House, 2012.

Mullins, Andrew P. J. "What Does Self-control Look Like? Considerations About the Neurobiology of Temperance and Fortitude." *Conatus – Journal of Philosophy* 10, no. 1 (2025): forthcoming.

PDSA. *PDSA Dickin Medal*. https://www.pdsa.org.uk/what-we-do/animal-awards-programme/pdsa-dickin-medal.

Rachels, James. *The Elements of Moral Philosophy*. McGraw-Hill, 2015.

Rawls, John. *A Theory of Justice*. Harvard University Press, 1999.

Rawls, John. *The Law of Peoples*. Harvard University Press, 2002.

Roden-Bow, Ashley. "Killer Robots and Inauthenticity: A Heideggerian Response to the Ethical Challenge Posed by Lethal Autonomous Weapons Systems." *Conatus – Journal of Philosophy* 8, no. 2 (2023): 477-486.

Ross, William David. *The Right and the Good*. Oxford. Oxford University Press, 2002.

Sandel, Michael. *Justice: What's the Right Thing to Do?* Farrar, Straus and Giroux, 2010.

Serafimova, Silviya. "Whose Morality? Which Rationality? Challenging Artificial Intelligence as a Remedy for the Lack of Moral Enhancement." *Humanities and Social Sciences Communications* 7, no. 119 (2020): 1-10.

Sullins, John P. "Automated Ethical Practical Reasoning: The Problem of Artificial Phronesis." In *Robophilosophy: Philosophy of, for, and by Social Robotics*, edited by J. Seibt, R. Hakli, and M. Nørskov. MIT Press, 2025, forthcoming.

Sutton, Richard S., and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 2018.

Syse, Henrik, and Martin Cook. "Robotic Virtue, Military Ethics Education, and the Need for Proper Storytellers." *Conatus – Journal of Philosophy* 8, no. 2 (2023): 667-680.

Tasioulas, John. "First Steps Towards an Ethics of Robots and Artificial Intelligence." *Journal of Practical Ethics* 7, no. 1 (2019): 61-95.

Tasioulas, John. "The Rule of Algorithm and the Rule of Law." Lecture at the University of Vienna, October 15, 2021.

Valenzuela, Pia. "Fredrickson on Flourishing through Positive Emotions and Aristotle's Eudaimonia." *Conatus – Journal of Philosophy* 7, no. 2 (2022): 37-61.

Yudkowsky, Eliezer. "Complex Value Systems in Friendly AI." In *Artificial General Intelligence*, edited by J. Schmidhuber, K. Thó risson, and M. Looks, 388-393. Springer, 2011.

Yudkowsky, Eliezer. "The Value Loading Problem." *EDGE*, July 12, 2021. https://www.edge.org/response-detail/26198.

Zetzsche, Dirk A., Ross P. Buckley, Janos N. Barberis, and Douglas W. Arner. "Regulating a Revolution: From Regulatory Sandboxes to Smart Regulation." *Journal of Corporate and Financial Law* 23, no. 1 (2017): 31-103.