International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

Humane democracy: Virtue ethics, artificial intelligence and merit-based governance

Nasip DEMİRKUŞ *

Department of Biology Education, Faculty of Education, Van Yüzüncü Yıl University, Tuşba, Van, Türkiye.

International Journal of Science and Research Archive, 2026, 19(03), 616-626

Publication history: Received on 05 May 2026; revised on 12 June 2026; accepted on 15 June 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.19.3.1338

Abstract

Addressing the two foundational crises of democratic governance — the crisis of representation and the crisis of morality — this study builds on the philosophical foundations of Prof. Dr. Nasip Demirkuş's (2026) Humane Democracy model to present an implementable sociotechnical governance architecture. It is not an empirical implementation report but a theoretical model proposal that combines normative political theory, comparative legal analysis, and sociotechnical system design. Its original contributions fall into three clusters: (i) a democracy and suffrage architecture — the Graduated Civil-Responsibility Weighted Vote (KSAOH), which grades the vote weight of individuals who gravely violate the social contract, strictly bound by judicial decision and the principle of temporariness; (ii) a moral-merit and citizenship architecture — the Moral-Maturity Assessment Framework (AODÇ) and the voluntary, cryptographically anonymised Citizen Moral-Development Portfolio (YAGP); and (iii) an AI and global-governance architecture — Constitutional AI integration, updated autonomous-system ethics laws, and a Federated Universal Governance Network. Synthesising Aristotelian phronesis, Constitutional AI (Bai et al., 2022), and the UNESCO (2021), OECD (2024), and EU AI Act (European Parliament & Council of the European Union, 2024) frameworks, the model proposes a normative yet implementable governance paradigm.

Keywords: Humane Democracy; AI Ethics; Virtue/Good Character; KSAOH; Phronesis; Constitutional AI; Synergistic Assembly; Autonomous-System Ethics; Federated Universal Governance Network

1. Introduction

In the twenty-first century, democratic systems grapple with two deeply interconnected crises: a crisis of representation and a crisis of morality. The rise of populism, deepening institutional distrust, and the manipulation of voters through disinformation channels expose the structural weaknesses of liberal democracy's purely numerical, majority-based conception of legitimacy (Foa & Mounk, 2016; Mounk, 2018). The liberal-democratic triumph that Fukuyama (1992) proclaimed with his "end of history" thesis has today given way to a profound anxiety over democratic backsliding.

At the centre of this crisis lies the question: what kinds of institutional and technological safeguards are required to bring intelligent, morally upright, and virtuous individuals into governance, and to make technology a part of humanity's moral maturation? While seeking to answer this, the Humane Democracy model (Demirkuş, 2026; Demirkuş, 2008) stresses that contemporary technology must likewise be equipped with the same universal moral principles (Génova et al., 2023; Gogoshin, 2021).

This study advances Demirkuş's (2026) prior work — which laid out the philosophical and political foundations of Humane Democracy (majoritarianism, merit, and the epistemic basis of democratic legitimacy) — through an original and independent technological architecture. Whereas that prior work focuses on normative political theory, the present study is an independent engineering-and-governance synthesis investigating how those foundations can be

* Corresponding author: Nasip DEMİRKUŞ

operationalised through Constitutional AI algorithms (Bai et al., 2022), the KSAOH system design, the anonymous YAGP data architecture (ZKP/DID), and global federated governance networks.

1.1. Research question

The study's central research question is: how can an ideal of virtue-ethics- and merit-based governance be designed through verifiable and accountable sociotechnical mechanisms without undermining one-person-one-vote, human rights, or cultural pluralism?

In response, the study's original contributions are as follows:

- To propose a temporary, proportionate vote-weighting mechanism (KSAOH) secured by DLT, made transparent through SHAP/counterfactual XAI, and safeguarded by a Human-in-the-Loop (HITL) judicial guarantee.
- To design a framework (AODÇ) that assesses moral maturity strictly through verifiable behaviour related to public office, together with a voluntary, anonymous portfolio (YAGP) structurally distinct from social scoring.
- To unify Constitutional AI, updated autonomous-system laws, and XAI within a single value-aligned architecture.
- To integrate all components under a Federated Universal Governance Network grounded in the Kantian federation tradition.

2. Materials and Methods

2.1. Research Design

The study adopts a mixed theoretical methodology composed of three complementary layers: (i) a comparative conceptual analysis of the notions of good character, merit, virtue, and value alignment against deep philosophical traditions (Aristotle, 2007; MacIntyre, 1984; Kahneman, 2011; Haidt, 2012) and contemporary technology-ethics standards (UNESCO, 2021; OECD, 2024; European Parliament & Council of the European Union, 2024); (ii) normative system modelling of the KSAOH, YAGP, and Constitutional AI governance architectures; and (iii) a comparative legal analysis evaluating the compatibility of the KSAOH mechanism with ECtHR case law (European Court of Human Rights, 2005), the EU AI Act (European Parliament & Council of the European Union, 2024), and OECD principles (OECD, 2024).

The study is primarily theoretical and architectural; it is not an empirical implementation report. Empirical validation — agent-based simulations of KSAOH weight distributions, formal verification of Constitutional AI constraints, privacy analysis of the ZKP-YAGP architecture, and pilot trials — is carried forward as the agenda of future research.

Falsifiability and testable propositions. Despite its theoretical nature, the model yields propositions testable in future work: (P1) in agent-based simulation, KSAOH coefficients should keep aggregate representational inequality below a predefined threshold (e.g., a Gini-like measure); (P2) inter-rater agreement among AODÇ assessors should exceed an acceptable reliability threshold (e.g., Cohen's κ or ICC); (P3) the ZKP-YAGP architecture should enable proof of participation without identity disclosure. Rejection of any proposition would require revision of the corresponding component.

2.2. Theoretical Framework

2.2.1. Science and Holistic Knowledge: Complementary Modalities

Every political and technological regime rests on an implicit or explicit conception of the human being and of knowledge. The Humane Democracy model (Demirkuş, 2026; Demirkuş, 2008) is grounded in the complementarity of holistic knowledge (İlim) and positive science (Bilim). These two dimensions should be understood not as mutually cancelling but as mutually complementary cognitive modalities. The engineering literature increasingly accepts that sociotechnical system design cannot be value-neutral (Génova et al., 2023).

2.2.2. Practical Wisdom (Phronesis) as a Multi-Objective Optimisation Weighting Function

This study adopts Aristotelian phronesis (practical wisdom) (Aristotle, 2007) as its core epistemic framework. From an engineering perspective, phronesis can be formulated as a weighting function within a Multi-Criteria Decision-Making (MCDM) process: an optimisation capacity that balances technical efficiency, ethical constraints, and social impact in context-dependent ways. Kahneman's (2011) System 1 (fast, intuitive, reactive) can be situated technologically by analogy with edge-device neural networks, and System 2 (slow, analytical, deliberative) with cloud-based large language models (LLMs). Haidt's (2012) Moral Foundations Theory likewise shows that moral decisions rest on multidimensional intuitive foundations — care, fairness, loyalty, authority, sanctity — which map directly onto MCDM criterion vectors.

2.2.3. Instrumental Rationality and Objective-Function Specification Drift

Weber's (1978) distinction between Zweckrationalität (instrumental rationality) and Wertrationalität (value rationality) maps precisely onto the problem of machine-learning models minimising a mathematical loss function while disregarding ethical side-effects — objective-function specification drift. Horkheimer and Adorno's (2002) critique of the domination of instrumental reason translates directly to the AI alignment problem: Wertrationalität means, in engineering terms, adding fairness metrics, human-rights constraints, and value-aligned reward modelling to the algorithmic optimisation loop.

2.3. Definition of Key Concepts

Virtue ethics is the ethical tradition that centres character and the virtues, defining the right action as the one a virtuous agent would perform in a given context (Aristotle, 2007; MacIntyre, 1984). **Good character (güzel ahlak)** in this study denotes the set of virtues that cannot be reduced to any single religious tradition and that are embodied in humanity's shared heritage of values — human dignity, justice, responsibility, and solidarity. **Moral maturity** is the capacity to apply these virtues, in a context-appropriate manner, through consistent moral reasoning (phronesis). Here these concepts are operationalised only at the level of verifiable behaviour related to public office; they do not encompass an individual's private beliefs, conscience, or lifestyle.

2.4. Critique of Existing Democratic Systems

From a Humane Democracy perspective, the greatest weakness of existing democracies is their numerical, majority-based conception of legitimacy (Demirkuş, 2008). Seizing power with 30–40% of the vote and then ruling over the whole of society concretises the problem of the tyranny of the majority debated since Tocqueville (2000) and Mill (1991), and brings with it the risk of permanent minority victimhood noted by Abizadeh (2021). The remedy lies in constitutional mechanisms, proportional representation, and a state-funded electoral model (Demirkuş, 2025a).

Brennan's (2016) defence of epistocracy diagnoses the problem of voter quality but carries its own dangers. Bell's (2015) Confucian-meritocracy model powerfully develops the ideal of virtuous governance; yet, as Sandel (2009) argues in his theory of justice, political philosophy cannot remain neutral on questions of the good life. Systems that make merit the sole criterion generate, in Sandel's (2020) apt phrase, a "tyranny of merit." The proposed model is differentiated from these approaches as summarised below:

Table 1 Comparison of the proposed model with major governance approaches

Approach	Core criterion	Principal weakness	Humane Democracy's difference
Liberal one-person-one-vote	Numerical majority	Majority tyranny; permanent minority	Weighting only for judicially established grave violations; temporary and proportionate
Epistocracy (Brennan, 2016)	Knowledge / competence	Rule by a knowledgeable elite; exclusion	Criterion is verifiable civil responsibility, not knowledge; HITL + judicial review
Confucian meritocracy (Bell, 2015)	Examination + virtue	Single-centred; legitimacy deficit	Virtue + citizens' assemblies + distributed veto for democratic legitimacy
Sortition/deliberative (Landemore, 2013; Fishkin, 2009)	Representative sample	Limits of continuity and expertise	Sortition assemblies integrated as a complementary component

3. Findings: The Proposed System Architecture

The principal output of this research is a multi-layered governance and AI-ethics architecture composed of five interoperable components: the KSAOH vote-weighting system, the AODÇ assessment framework, the YAGP civic-portfolio system, Constitutional AI integration for algorithmic systems, and the Updated Autonomous-System Laws. The components and layers of the architecture are shown in Figure 1.

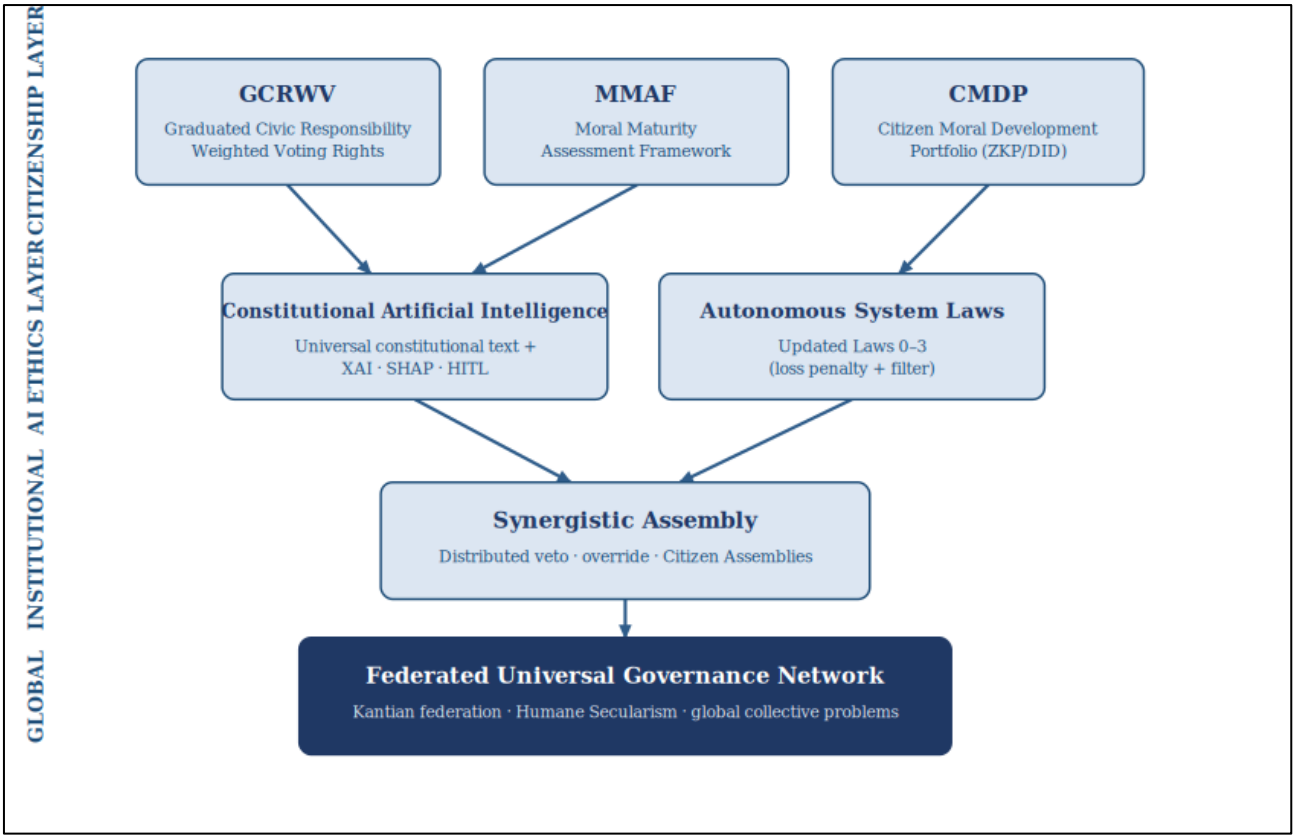

Figure 1 Humane Democracy sociotechnical system architecture: citizenship, AI-ethics, institutional, and global layers

3.1. Graduated Civil-Responsibility Weighted Vote (KSAOH)

3.1.1. Legal Nature and Rehabilitative Design

KSAOH is not punitive but a temporary, rehabilitative, civic arrangement grounded in social responsibility (Demirkuş, 2026). The theoretical tension it creates with one-person-one-vote is openly acknowledged. To meet this tension, KSAOH is framed as an **exceptional civil-responsibility mechanism bound solely to an independent judicial decision, proportionality, temporariness, the right of appeal, and human oversight**; under no circumstances is it an administrative or automatic sanction. In the European Court of Human Rights case Hirst v. United Kingdom (No. 2, 2005) the ECtHR held that general, automatic, and indiscriminate restrictions violate the Convention, whereas restrictions individualised according to the gravity of the offence and made under judicial review may be lawful.

3.1.2. Three Core Defence Strategies

- **Principle of proportionality:** The restriction is tied directly to the gravity of the offence.
- **Principle of temporariness:** All restrictions are lifted automatically once the term ends; the system rests on restorative justice.
- **Least-restrictive-means principle:** Rather than fully suspending the vote, KSAOH temporarily reduces its weight — the least intrusive instrument.

3.1.3. KSAOH Tiers

Table 2 KSAOH Tiers: Restrictions on Voting and Eligibility

Offence / Violation Category	Vote Restriction	Eligibility Restriction	Term	Recovery Condition
Intentional homicide, mass killing, crimes against humanity	Suspended (0)	Lifetime ban	Sentence term	Sentence + judiciary + jury decision
Aggravated sexual assault, child abuse	0.5	Lifetime ban	15 yrs (vote)	15 yrs + judicial review
Serious violent crime (armed robbery, grievous bodily harm)	0.7	10-yr restriction	Conviction + 5 yrs	Term + good-conduct record
Public-official corruption / bribery	0.8	Lifetime ban	10 yrs (vote)	10 yrs + judicial decision
Grave traffic offence (fatal crash + alcohol/speeding)	0.9	5-yr restriction	5 yrs	Term + traffic-ethics training
First-time minor traffic offence	None (1.0)	None	—	Recorded

Note: The values (0, 0.5, 0.7, 0.8, 0.9, 1.0) denote numerical weighting coefficients. These are not final statutory values but principled, debatable model values; in future work they should be calibrated through multi-stakeholder constitutional consensus meetings, agent-based simulations, and pilot trials. All restrictions rest on an independent court decision, cannot be applied automatically, and the right of appeal is guaranteed.

3.1.4. Distributed-Ledger Security and XAI Transparency

To compute the differing vote-weight coefficients of millions of citizens securely and protect them against tampering, KSAOH is proposed to run on a Permissioned Blockchain or Distributed Ledger Technology (DLT) infrastructure. Each weighted-vote transaction thereby forms an irreversible, auditable record, preventing single-point-of-failure risks and the manipulation of the weight database by state or private actors.

Every individual subject to KSAOH is automatically provided with an explainable summary of the algorithm that computes their vote weight, in line with the XAI principles set out by Arrieta et al. (2020). The system uses SHAP values to decompose the contribution of each input variable (offence category, sentence length, time elapsed) to the final coefficient, and Counterfactual Explanations to tell the individual what would need to change for the restriction to be lifted. Upon appeal, human-judge and citizen-jury review is engaged — the algorithm's computation is never binding on its own. The system preserves the balance between decision quality and democratic equality emphasised in Estlund's (2008) epistemic democracy and applies the Human-in-the-Loop (HITL) principle as the final judicial safeguard.

3.2. Moral-Maturity Assessment Framework (AODÇ)

The absence of an operational counterpart for the notion of good character exposes merit assessment to the risk of arbitrariness, a danger laid bare by Fricker's (2007) discussions of epistemic injustice. To bound this risk, AODÇ is proposed as a four-layered assessment architecture:

- **Multi-observer assessment:** A polyphonic process composed of peer, citizen-jury, and professional-body layers.
- **Ethical-reasoning test:** Operationalising Aristotle's (2007) framework to assess phronesis (MCDM) capacity through moral-dilemma scenarios.
- **Public-benefit action record:** Use of voluntary-service records accumulated in the YAGP as evidence.
- **Moral-Review Commission:** An independent body applied only to candidates for public office; members are drawn equally from members of the judiciary and citizens chosen by lot; every decision is open to judicial review; no decision may be based on political opinion, ethnicity, religion, or sex.

Scope limit and meta-safeguard ("who guards the guardians?"). AODÇ does not evaluate private belief, political opinion, ethnic identity, or lifestyle; it is confined to verifiable behaviour related to public office, ethical reasoning, public-benefit record, and legal responsibility. To prevent the commission itself from becoming a "moral bureaucracy": membership rotates and is term-limited; all reasons are transparently recorded and open to judicial review; and the commission assesses only candidates for public office, never ordinary citizens. For measurement validity, the construct

validity and inter-rater reliability of the AODÇ criteria (e.g., Cohen's κ / ICC) should be reported against predefined thresholds and audited regularly.

3.3. Universal Morality and the Citizen Moral-Development Portfolio (YAGP)

3.3.1. A Universal Source

The Humane Democracy model's conception of good character rests not on a specific religious tradition but on humanity's shared heritage of values (Demirkuş, 2026; Demirkuş, 2008). As supported by the work of Soroush (2000), Abou El Fadl (2004), Parray (2024), and Casanova (1994), different faith traditions — Islam, Christianity, Buddhism, Confucianism, secular humanism — are treated as cultural sources making a shared contribution to this framework. The state's reference points are the principles of human dignity, justice, and responsibility in the UN Universal Declaration of Human Rights.

3.3.2. YAGP Architecture and Legal Safeguards

YAGP is not a scoring system that ranks, punishes, or determines access to rights. It is solely a voluntary, anonymous, verifiable, consent-based public-benefit portfolio, and it is structurally distinct from the social-scoring practices prohibited under Article 5 of the EU AI Act (European Parliament & Council of the European Union, 2024):

- **Prohibited uses:** YAGP data may not be used in public employment, the right to education, credit scoring, or political-merit assessments.
- **No automated processing:** YAGP data may not be fed as input to any algorithmic decision system.
- **Zero-Knowledge Proof (ZKP) architecture:** Participation is stored using ZKP cryptographic protocols and Decentralised Identifier (DID) standards; individuals can cryptographically prove the existence of their voluntary contributions without disclosing identifying personal data.
- **Principle of voluntariness:** No public institution may make YAGP participation mandatory.
- **Independent audit:** Data-processing operations are subject to annual independent bias auditing under Article 10 of the EU AI Act.

YAGP is a voluntary, cryptographically anonymised network of solidarity that adapts MacIntyre's (1984) virtue-ethics tradition to an institutional context.

3.4. Endowing AI and Autonomous Systems with Good Character

3.4.1. Global Regulatory Frameworks

An AI power devoid of a value framework clearly risks turning into an algorithmic domination (Génova et al., 2023; Horkheimer & Adorno, 2002). Three global normative frameworks anchor the AI-ethics component of the proposed architecture: the UNESCO Recommendation on the Ethics of AI (2021), the first global normative framework endorsed by 194 member states; the OECD AI Principles (2024), which add the principles of Information Integrity, Environmental Sustainability, and Global Interoperability; and the EU AI Act (European Parliament & Council of the European Union, 2024, Reg. 2024/1689), which prohibits social-scoring systems under Article 5.

3.4.2. Constitutional AI: Replacing RLHF Majority Approval with Universal Norms

Constitutional AI (Bai et al., 2022) is adopted as the technical means of endowing the algorithm with good character. In this approach, systematically proposed by Bai et al. (2022), the model evaluates and corrects its own outputs against a pre-encoded universal constitutional text — the UN Universal Declaration of Human Rights, UNESCO (2021) principles and OECD (2024) standards — rather than against the sycophancy weakness of RLHF, which rests on the momentary approval of the popular majority. (Just as populist politicians who lie to the crowd at the ballot box are rewarded with election, RLHF models are misaligned when rewarded for producing content that confirms user biases.) This makes it possible to project Landemore's (2013) ideal of collective reason onto technological systems.

Update procedure for the constitutional text. Changing the constitutional principles to which the model is bound is possible only through a broadly participatory, transparent, super-majority process: Synergistic Assembly + Citizens' Assembly + independent judicial review + a qualified majority of at least 75–80%. The core principles corresponding to Articles 1–3 of the UN Declaration are protected as non-amendable in a backward-looking sense.

3.4.3. Situational Value Alignment and Federated Morality

A universal conception of morality must not become a monolithic tyranny that erases cultural diversity. The Federated Morality framework — by analogy with federated-learning architectures — allows the roots of a universal tree of virtue to be preserved while its branches are shaped through local model fine-tuning (Page, 2011). The core–branch boundary is defined explicitly: inalienable norms such as the fundamental rights of the UN Declaration are positioned as the "core" (non-amendable), while elements such as worship, custom, and local priority weights are positioned as the "branches" (open to local adaptation).

3.4.4. The Eight Principles of AI Endowed with Good Character

- **Respect for human dignity:** No algorithmic decision may demean human dignity.
- **Transparency and explainability (XAI):** Decisions must be intelligible, auditable, and justifiable (Arrieta et al., 2020).
- **Justice and impartiality:** No discrimination on the basis of sex, race, religion, or political opinion; fairness metrics must be embedded in the training objective.
- **Accountability:** There must exist a human or institution to assume responsibility for every AI decision.
- **Solidarity and collective benefit:** AI serves humanity's shared welfare.
- **Reversibility:** Critical decisions cannot become final without human oversight.
- **Safety and security:** Systems are equipped with safeguards against cyberattack and manipulation.
- **Environment and ecological balance:** AI's carbon footprint is a matter of moral responsibility (OECD, 2024).

3.4.5. Updated Autonomous-System Laws

The concept of robot ethics has expanded in the 2024–2026 literature into autonomous-system ethics and embodied-AI ethics (Gogoshin, 2021; Yigit et al., 2018). Asimov's (1942) three laws are inadequate for the present day. From an engineering-implementation perspective, the following laws are operationalisable: (a) irreversible Constitutional AI constraints that block prohibited outputs at inference time (Bai et al., 2022); and (b) large positive loss-penalty terms that activate when the model produces content violating Articles 1–3 of the UN Declaration. The laws thus operate not merely as rule-based safety filters but as learned behavioural priors embedded in the reward model and linked to Constitutional AI oversight.

- **Law 0 (Universal):** A robot/autonomous system may not execute any command that violates the fundamental rights defined in Articles 1, 2, and 3 of the UN Universal Declaration of Human Rights — human dignity, non-discrimination, and the right to life, liberty, and security.
- **Law 1 (Updated):** The duty not to cause harm; this also covers psychological, social, and digital harm, including deepfake production, algorithmic manipulation, and breaches of information integrity.
- **Law 2 (Updated):** A robot obeys commands; but if a command clearly conflicts with universal human-rights norms, it bears the duty to report that conflict to its overseer.
- **Law 3 (Updated):** A robot protects itself, but only to the extent that this protection does not harm a human, an animal, or the environment.

4. Discussion

4.1. Institutional Architecture for Humane Democracy

4.1.1. The Synergistic Assembly: Constructive Conflict, Distributed Veto, and Anti-Deadlock Mechanisms

Habermas's (1996) theory of deliberative democracy shows that moral truth emerges from debate on a fair footing. The Synergistic Assembly internalises this through the following safeguards: each bloc holds veto power in a distributed manner; constitutional amendments require a 75% qualified majority; all deliberative processes are open to civil society. To prevent deadlock: (i) a time-bound veto — a bloc may not veto the same proposal a second time within 90 days; (ii) an override mechanism — in ordinary legislative decisions a veto can be overridden by a 60% majority of the other blocs. This structure is consistent with veto-players theory (Tsebelis, 2002).

4.1.2. Citizens' Assemblies and State-Funded Elections

Before major policy decisions, Citizens' Assemblies selected by lot are convened (Fishkin, 2009; Landemore, 2013). The Irish, Icelandic, and French experiences attest to the effectiveness of this model. A state-funded electoral model is a structural necessity for breaking the influence of capital over politics (Demirkuş, 2025a; Demirkuş, 2026).

4.2. A Global Vision: The Federated Universal Governance Network

Demirkuş's (2008) vision of a Universal State of Humanity is reframed, so as to prevent the concentration of power, as a Federated Universal Governance Network. This structure rests on the tradition of a federation of nations in Kant's (2003) Perpetual Peace: a network of republics that preserve their national sovereignties while being bound to one another by shared moral and legal covenants. The network's jurisdiction is initially limited to global collective-action problems such as cross-border AI accidents, climate change, and antibiotic resistance. Humane Secularism — the state standing at equal distance from different faiths and cultures (Casanova, 1994; Demirkuş, 2025b) — forms its normative ground.

4.3. Synthesis: Architectural Integration

Table 3 Synthesis: Good Character, Democracy, and the Technology Architecture

Domain	Core Problem	Contribution of Good Character	System Component
Humane Democracy	Leaders without a moral compass coming to power	Merit- and virtue-centred mechanisms	KSAOH (Demirkuş, 2026) + AODÇ + Citizens' Assemblies (Fishkin, 2009)
AI Systems	Lack of value alignment; RLHF sycophancy	Constitutional AI, XAI, fairness metrics	Bai et al. (2022) + Arrieta et al. (2020) + OECD (OECD, 2024)
Autonomous Systems	Unethical decisions; trampling human dignity	Universal virtue laws encoded as loss penalties	Updated Laws + Constitutional AI (Bai et al., 2022)
Education System	Amplifying instrumental reason while blunting value reason	Phronesis-MCDM education, ethics curriculum	YAGP (ZKP-anonymous) + Weber (Weber, 1978) framework
Global Governance	Nation-states' inability to solve existential problems	Federated moral covenant, polycentric structure	Federated Universal Governance Network (Kant, 2003)

5. Limitations and Future Work

5.1. Theoretical and Legal Limitations

- **Tension with liberal democracy:** The weighted-vote component of KSAOH is in theoretical tension with one-person-one-vote (Abizadeh, 2021).
- **Human-rights risks and non-ideal theory:** Paradoxically, KSAOH and AODÇ are most dangerous precisely in the contexts where they are most needed (weak judicial independence, illiberal tendencies), where they could become instruments of political repression. Their deployment is therefore conditional on the prior attainment of a threshold of judicial independence and constitutional safeguards; until that threshold is met, the weighting component should be held in abeyance.
- **Cultural relativism:** Universal good-character norms must be anchored to internationally ratified texts (the UN Declaration).

5.2. Empirical and Technical Limitations

- **Algorithmic bias:** KSAOH and AODÇ may reproduce structural biases in historical judicial data. Against this, data-fairness auditing, independent bias testing, XAI/SHAP explanations, counterfactual review, and mandatory human oversight (HITL) are stipulated.
- **Threat model and adversarial robustness:** Sybil attacks on the DLT voting infrastructure, gaming/manipulation of the YAGP, ZKP implementation errors and trust assumptions, the accuracy of off-chain data (the oracle problem), and model capture are open risks requiring formal verification, red-teaming, and independent audit.
- **Empirical validation:** KSAOH simulations, formal verification of Constitutional AI constraints, ZKP-YAGP privacy analysis, and pilot trials constitute the subsequent research stages.

- **Ethical pluralism in AI:** Reconciling the moral priorities of different cultures within a single optimisation objective is an open research problem.

6. Abbreviations and Terminology

Note on naming. The author's coined frameworks retain their original Turkish acronyms (e.g., KSAOH, AODÇ, YAGP) for consistency with the anchor work (Demirkuş, 2026), with English glosses provided below.

Acronym	Expansion (Turkish → English)
KSAOH	Kademeli Sivil Sorumluluk Ağırlıklı Oy Hakkı → Graduated Civil-Responsibility Weighted Vote
AODÇ	Ahlaki Olgunluk Değerlendirme Çerçevesi → Moral-Maturity Assessment Framework
YAGP	Yurttaş Ahlak Geliştirme Portföyü → Citizen Moral-Development Portfolio
DLT	Distributed Ledger Technology
ZKP / DID	Zero-Knowledge Proof / Decentralised Identifier
XAI / SHAP	Explainable AI / Shapley Additive Explanations
HITL	Human-in-the-Loop
MCDM / RLHF / LLM	Multi-Criteria Decision-Making / Reinforcement Learning from Human Feedback / Large Language Model

7. Conclusion

This study has transformed Prof. Dr. Nasip Demirkuş's (2026) Humane Democracy model — by synthesising it with deep philosophical foundations, comparative legal analysis, and the current AI-ethics literature — into a systematic, multi-component, and implementable sociotechnical governance architecture. The principal contributions can be grouped under three headings:

- **Contribution to democracy and electoral architecture.** A temporary, proportionate KSAOH system secured by DLT, made transparent through SHAP/counterfactual XAI, and safeguarded by HITL judicial review, together with AODÇ and a state-funded electoral model.
- **Contribution to AI and autonomous-system ethics.** Constitutional AI integration encompassing loss-function penalty encoding and eight embedded moral principles, combined with the updated autonomous-system laws (Bai et al., 2022).
- **Contribution to global governance and federated morality.** The Synergistic Assembly's anti-deadlock veto-override mechanism, the YAGP grounded in ZKP/DID anonymisation, and the vision of a Federated Universal Governance Network grounded in Kant's (2003) federation theory.

The question of whether or not we will embed values in technology now lies behind us. The question is this: *which values, through which mechanisms, and with which safeguards?*

Compliance with ethical standards

Acknowledgments

In preparing this study the author made use of the Claude (Anthropic), Google Gemini, and DeepSeek AI tools for *literature search, synthesis, critical-analysis support, and manuscript formatting.* The generation of all original ideas — including the design of all proposed system architectures (KSAOH, AODÇ, YAGP, Constitutional AI integration, the Federated Universal Governance Network) — and the overall conceptual framework belong exclusively to Prof. Dr. Nasip Demirkuş. After using these tools, the author reviewed and edited the content as required and assumes full responsibility for the content of the publication. My thanks go to all engineers who have contributed to this technology.

Disclosure of conflict of interest

The author declares that there is no conflict of interest.

References

[1] Abizadeh, A. (2021). Counter-majoritarian democracy. American Political Science Review, 115(3), 919–933. https://doi.org/10.1017/S0003055421000368

[2] Abou El Fadl, K. (2004). Islam and the challenge of democracy. Princeton University Press. https://doi.org/10.1515/9781400826339

[3] Aristotle. (2007). Nicomachean Ethics (S. Babür, Trans.). Bilgesu. (Original work composed c. 350 BCE)

[4] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[5] Asimov, I. (1942). Runaround. Astounding Science Fiction, 29(1), 94–103.

[6] Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv. https://arxiv.org/abs/2212.08073

[7] Bell, D. A. (2015). The China model. Princeton University Press. https://doi.org/10.1515/9781400865505

[8] Brennan, J. (2016). Against democracy. Princeton University Press. https://doi.org/10.2307/j.ctv7h0s6j

[9] Casanova, J. (1994). Public religions in the modern world. University of Chicago Press. https://doi.org/10.7208/chicago/9780226190204.001.0001

[10] Demirkuş, N. (2008). Brainstorming and fallow questions on religion [Lecture notes]. Nadidem.net. https://nadidem.net/ders/soru/din.htm

[11] Demirkuş, N. (2025a). Brainstorming and fallow questions on democracy [Lecture notes]. Nadidem.net. https://nadidem.net/ders/soru/demokrasi.htm

[12] Demirkuş, N. (2025b). Brainstorming and fallow questions on secularism [Lecture notes]. Nadidem.net. https://nadidem.net/ders/soru/laiklik.htm

[13] Demirkuş, N. (2026). Rethinking human democracy: From pluralism to merit. World Journal of Advanced Engineering Technology and Sciences, 19(1), 117–124. https://doi.org/10.30574/wjaets.2026.19.1.0203

[14] Estlund, D. M. (2008). Democratic authority. Princeton University Press. https://doi.org/10.1515/9781400831548

[15] European Court of Human Rights. (2005). Hirst v. United Kingdom (No. 2) (Application No. 74025/01). https://hudoc.echr.coe.int/eng?i=001-70442

[16] European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

[17] Fishkin, J. S. (2009). When the people speak. Oxford University Press. https://doi.org/10.1093/acprof:osobl/9780199604432.001.0001

[18] Foa, R. S., & Mounk, Y. (2016). The danger of deconsolidation: The democratic disconnect. Journal of Democracy, 27(3), 5–17. https://doi.org/10.1353/jod.2016.0049

[19] Fricker, M. (2007). Epistemic injustice. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198237907.001.0001

[20] Fukuyama, F. (1992). The end of history and the last man. Free Press.

[21] Génova, G., Moreno, V., & González, M. R. (2023). Machine ethics: Do androids dream of being good people? Science and Engineering Ethics, 29, Article 20. https://doi.org/10.1007/s11948-023-00433-5

[22] Gogoshin, D. L. (2021). Robot responsibility and moral community. Frontiers in Robotics and AI, 8, Article 768092. https://doi.org/10.3389/frobt.2021.768092

[23] Habermas, J. (1996). Between facts and norms (W. Rehg, Trans.). MIT Press. (Original work published 1992)

[24] Haidt, J. (2012). The righteous mind. Pantheon Books.

[25] Horkheimer, M., & Adorno, T. W. (2002). Dialectic of enlightenment (E. Jephcott, Trans.). Stanford University Press. (Original work published 1947)

[26] Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

[27] Kant, I. (2003). Perpetual peace and other essays (T. Humphrey, Trans.). Hackett Publishing. (Original work published 1795)

[28] Landemore, H. (2013). Democratic reason. Princeton University Press. https://doi.org/10.1515/9781400848515

[29] MacIntyre, A. (1984). After virtue: A study in moral theory (2nd ed.). University of Notre Dame Press.

[30] Mill, J. S. (1991). On liberty and other essays. Oxford University Press. (Original work published 1859)

[31] Mounk, Y. (2018). The people vs. democracy. Harvard University Press. https://doi.org/10.4159/9780674984776

[32] OECD. (2024). OECD principles on artificial intelligence (updated May 2024; originally adopted 2019). https://oecd.ai/en/ai-principles

[33] Page, S. E. (2011). Diversity and complexity. Princeton University Press. https://doi.org/10.1515/9781400835140

[34] Parray, T. A. (2024). Islam and democracy in the twenty-first century. Routledge. https://doi.org/10.4324/9781003310624

[35] Sandel, M. J. (2009). Justice: What's the right thing to do? Farrar, Straus and Giroux.

[36] Sandel, M. J. (2020). The tyranny of merit. Farrar, Straus and Giroux.

[37] Soroush, A. (2000). Reason, freedom, and democracy in Islam. Oxford University Press.

[38] Tocqueville, A. de. (2000). Democracy in America (H. C. Mansfield & D. Winthrop, Trans.). University of Chicago Press. (Original work published 1835)

[39] Tsebelis, G. (2002). Veto players. Princeton University Press. https://doi.org/10.1515/9781400831456

[40] UNESCO. (2021). Recommendation on the ethics of artificial intelligence. https://en.unesco.org/about-us/legal-affairs/recommendation-ethics-artificial-intelligence

[41] Weber, M. (1978). Economy and society (G. Roth & C. Wittich, Trans.). University of California Press. (Original work published 1921)

[42] Yigit, T., Kose, U., & Sengoz, N. (2018). Robotics rights and ethics rules. arXiv. https://arxiv.org/abs/1809.08885