Ethics as a Socio-Technical Dimension of Modern Digital Systems*

Paola Inverardi

Gran Sasso Science Institute, Viale F. Crispi, 7 67100 L'Aquila

Abstract

In this paper, we review selected research results on the design of ethics-aware autonomous systems developed at GSSI. We discuss conceptual foundations, engineering methods, architectures, and tools, highlighting the need to consider both system behavior and human interaction.

Keywords

Autonomous systems, ethics, individual moral preferences, software architecture

1. Introduction

The world we inhabit is increasingly shaped — and, in many respects, structured — by digital technologies that act autonomously on our behalf and in response to our needs. From self-driving vehicles and medical diagnostic AI to algorithmic financial trading and domestic service robotics, digital ecosystems are evolving into autonomous technologies that directly mediate human access to information, social interaction, and decision-making.

As software systems gain progressive autonomy, their choices extend beyond substituting human agents in supplying contextual information. They directly impact the moral rights and agency of individual citizens, groups, and society as a whole. This transition raises profound ethical questions regarding human-system interactions. While technical and regulatory domains have traditionally studied privacy, fairness, transparency, and accountability through formal algorithms or legislative top-down frameworks (e.g., GDPR, EU AI Act), ethics cannot be reduced to a static checklist or pure algorithmic constraints.

Ethics in digital systems is fundamentally a **socio-technical challenge**. Digital systems must preserve human dignity and autonomy, empowering individuals to retain meaningful control over the values that guide their technological environment. This paper addresses the core conceptual, architectural, and operational requirements needed to shift computing paradigms toward *ethically aware autonomous systems*.

STPIS2026: Proceedings of the 12th International Conference on Socio-Technical Perspectives in IS, September 16-17th L'Aquila, Italy

* Corresponding author.

† These authors contributed equally.

paola.inverardi@gssi.it (P. Inverardi)

0000-0001-6734-1318 (P. Inverardi)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-4334/paper60.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-19

2. Hard Ethics vs. Soft Ethics: The Socio-Technical Gap

To ground ethical considerations into autonomous systems, the notion of digital ethics in [2] that distinguishes between *Hard Ethics* and *Soft Ethics* can help set the boundaries of the system's autonomy:

- **Hard Ethics:** Represents non-negotiable legal mandates, regulatory frameworks, and universal systemic constraints (e.g., EU AI Act, safety compliance, fundamental rights). Hard ethics is system-centric and enforces hard boundaries to ensure trustworthiness, legal compliance, and safety.
- **Soft Ethics:** Encompasses the space of moral decisions and personal values left to individual actors, organizations, and citizens above and beyond existing regulation. Soft ethics is fluid, highly subjective, contextual, and variable across cultures, individuals, and situational states.

Accordingly, autonomous systems should be envisioned as illustrated in Figure 1 [3]:

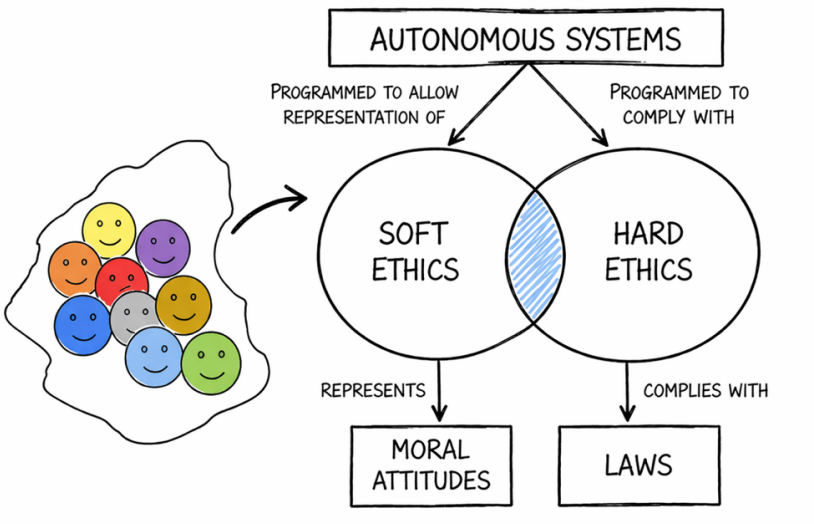


Figure 1: Conceptual representation of autonomous systems and digital ethics.

Current approaches to embed ethics in autonomous systems often struggle with the alignment problem because human ethics is inherently non-universal. While system architectures are built for universal hard-ethics compliance, they typically lack mechanisms to represent, negotiate, and adapt to individual human soft ethics. Putting humans at the center of modern digital ecosystems requires software architectures that support individual customization and negotiation. An **ethical-aware autonomous system** is an autonomous system that adapts to human interactions and/or changes in physical or virtual environments, and that exhibits a behavior that conforms to the **system ethics (hard ethics)**, the **human ethics (soft ethics)**, and the **current status** of the humans that interact or share the physical or virtual environment with the system (as illustrated in Figure 2) [4].

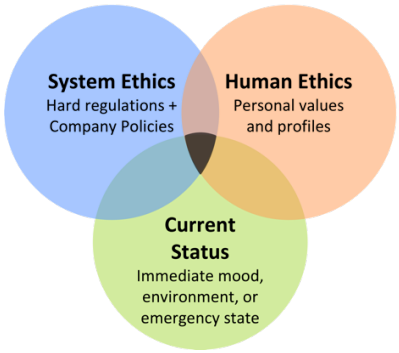


Figure 2: Tri-focal intersection defining an ethical-aware autonomous system.

3. The Exosoul Paradigm: Software Exoskeletons for Human Agency

To bridge the gap between autonomous systems and user agency, we introduce the **Exosoul** approach [6]. Rather than forcing AI models to learn an impossible "universal moral preference," Exosoul delegates soft ethics mediation to a software component acting on behalf of the user: a **Software Exoskeleton**.

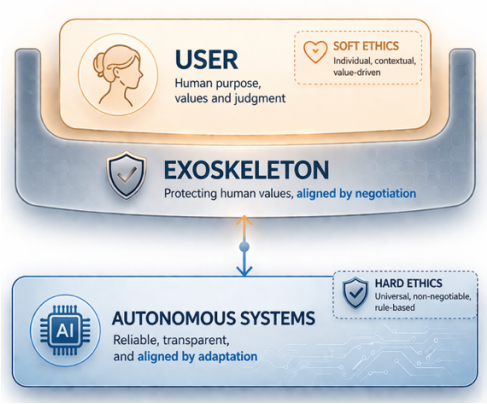


Figure 3: The Exosoul paradigm: the software exoskeleton acts as a protective mediator and negotiator between the user and the autonomous system.

The Exoskeleton is designed according to the following **key engineering principles:**

- **Ethical Profiling:** Eliciting both explicit and implicit soft-ethics preferences from the user through continuous context observation or interactive interfaces [7].
- **Dynamic Mediation:** Functioning as a proxy that negotiates interaction policies between the user's soft-ethics profile and the autonomous system's constraints [8].
- **Control Delegation & Override:** Enabling continuous human oversight, allowing the user to adapt system autonomy levels or invoke ethical overrides when dilemmas arise.

4. A reference Architecture for ethical-aware autonomous system

The genuine social acceptability of an ethical-aware autonomous system does not rely solely on satisfying the preferences of the primary end-user. As illustrated in Figure 4, designing the architecture of such systems requires the alignment of five key categories of stakeholders:

- **User:** The primary individual who interacts directly with the system and guides its operation.
- **Human:** Individuals who do not interact directly with the system but share the physical or virtual environment with it.
- **Owner:** The entity or person holding legal ownership and responsibility for the system.
- **Integrator:** The technical actor who adapts and integrates the system into a specific socio-technical context.
- **Producer:** The developer who creates the core technology and defines foundational architectural and safety policies.

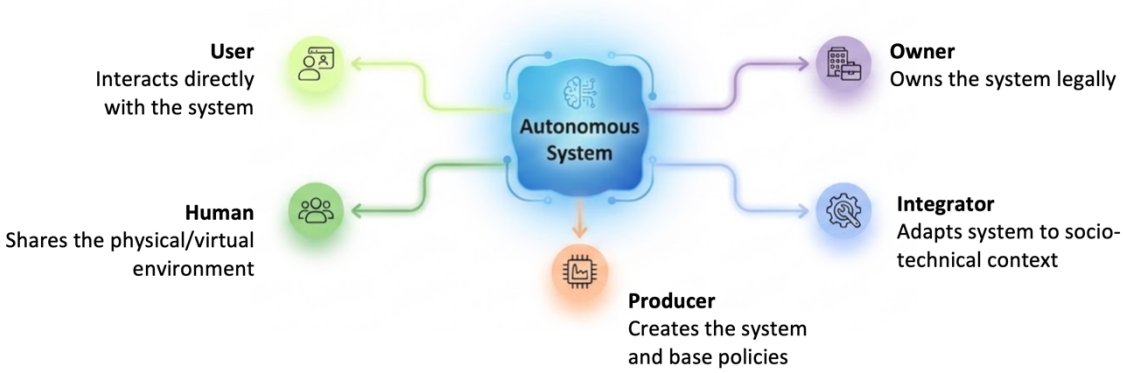


Figure 4: Multi-stakeholder alignment required to ensure social acceptability by balancing the preferences of the User, the boundaries of surrounding Humans, and the liabilities of the Owner, Integrator, and Producer

A rigorous reference architecture [4] must therefore ensure continuous balancing between fulfilling the soft-ethics values of the **User**, safeguarding the rights and boundaries of surrounding **Humans**, and respecting the legal and regulatory liabilities of the **Owner**.

We introduce the **E-MAPE-K (Ethical MAPE-K)** reference model, which incorporates an **Ethic Connector** and structured Knowledge modules to support dynamic moral adaptation [4].

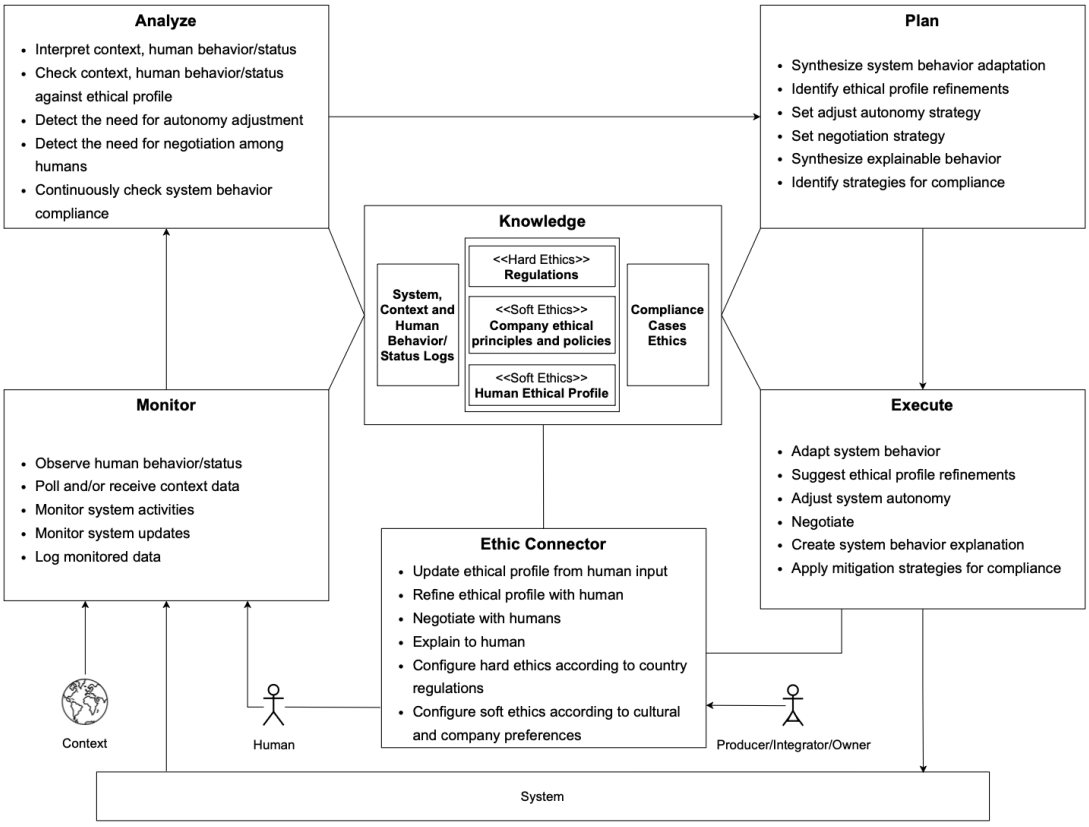


Figure 5: The E-MAPE-K (Ethical MAPE-K) reference architecture, extending traditional self-adaptive feedback loops with an Ethic Connector and dedicated Knowledge modules for ethical governance.

As illustrated in Figure 5, **E-MAPE-K** includes the following **core components:**

- **Monitor:** Continuously polls environmental status, system telemetry, and user state (e.g., workload, emergency conditions).
- **Analyze:** Evaluates system trajectories against hard ethics rules (regulations) and user soft ethics preferences. Detects conflicts, safety violations, or the need for autonomy level adjustment.
- **Ethic Connector:** Manages real-time negotiation, receives human override inputs, updates ethical profiles, and formats explainable rationales.
- **Plan:** Synthesizes adaptation plans that resolve moral dilemmas, balancing system objectives against user values.
- **Execute:** Deploys adapted operational strategies to the underlying autonomous system.

5. Transparency and Governance: Ethics Labels for Digital Systems

To ensure that autonomous systems remain accountable and comprehensible to non-technical users, ethical qualities must be clearly communicated. Borrowing concepts from nutritional facts and privacy nutrition labels, we propose a standardized **Ethics Label** framework for digital systems [1].

The Ethics Label translates complex legal, technical, and algorithmic compliance data into an accessible multi-card interface (as illustrated in Figure 6 and summarized in Table 1) structured across four core dimensions: Human, Societal, Environmental, and Management of Adverse Effects.

Ethics label.

Human	Societal	Environmental	Management of adverse effects
Principle of Autonomy: if and how the system limits human control and compromises user autonomy, understood as the ability of humans to act according to their informed beliefs. For instance, whether human beings can take or regain control over the system autonomy, and explain when this is not possible, allowed, or beneficial for humans.	*Society and Democracy*: if and how the digital system influences democratic processes, political engagement, public deliberation decision making, as well as media pluralism. For example, a news app should clearly state whether it uses a fact checker and which filtering policy (if any) is adopted in reporting opinions.	*Energy Consumption*: if and how the system provides a user-understandable estimation of energy consumption for the services offered or the training of AI algorithms. For instance, producers can make comparison with the energy needed by houses, saunas, etc.	*Minimization of Adverse Effects*: if and how the system has a plan to monitor, eliminate, or limit risks or negative impact as much as possible. For instance, organizations can have standardized risk management practices and processes for managing both existing and newly detected human, societal, and environmental risks. Also, organization can make use of bias detection solutions for diversity, fairness, and non-discrimination.
Privacy and Data Governance/ Intellectual Property: if and how privacy, data governance, and data intellectual property are protected and managed. How the misuse of personal data is prevented, how transparency in data management is ensured, and if and how the rights of the users over the information they generate are upheld and safeguarded. If and how the governance structure of the system, including who controls and who accesses data, are transparently disclosed.	*Inclusive and Participatory Design*: if and how inclusive design principles have been followed, along with the identification and consultation of relevant stakeholders during the system development. If and how the producer states whether the system was tested and shaped using feedback from diverse user groups, such as individuals with disabilities and elderly users.	*Greenhouse Gas Emissions*: if and how the system provides a user-understandable estimation of gas emission for the services offered or the training of AI algorithms. To make it understandable to users, producers can make a comparison with the emissions of vehicles.	*Mitigation of Risk/Negative Impact*: if and how the system recognizes the inevitability of some risks. If and how the system has strategies to address the aftermath of a problem, as well as the steps that can be taken beforehand to reduce adverse and potentially long-term effects. For instance, when a data leak is discovered, the system can take actions to reduce the severity of the consequences.
Transparency and Explainability: if and how it is made clear that AI is being used and that the resulting outcomes are transparent to those who use or interact with a digital system, e.g., an AI-powered system. If and how the rationale behind the decisions made by the system is explained.	*Openness and Plurality/Cultural Sensitivities*: if and how the system has been developed in an open-source environment and by a plurality of actors (e.g., universities, public authorities). Also, if and how the system is able to recognize, understand, and respect cultural differences. For example, users should be informed whether and why the system uses only a specific language or does not consider some cultural aspects.	*Water Consumption:* if and how the system provides user-understandable explanation of the system's water consumption during production, algorithm training, as well as use. The explanation should clearly make examples and comparisons, e.g., in terms of the average daily water consumption of a person.	*Reporting Negative Impact*: if and how the system provides a user-understandable description of the potential negative impacts of the system, along with a clear explanation of the strategy for reporting them. For instance, there should be plans and strategies in place to inform users about potential data leaks or biased decisions, including their impact and scope.
Beneficialness: the extent to which the use of the system is beneficial for humans and how. This includes whether and how the producer describes the benefits provided by the system in relation to its primary goals. For example, a productivity application that helps managing workload in a healthier way, can be considered to support user digital well-being as part of its core functionality.	*Diversity, Fairness, and Non-Discrimination*: if and how the system aligns with the ideal of justice and promotes fairness, inclusion, respect for diversity, and equality of opportunity. For example, the representativeness of the sample used for testing should be clear, as well as whether the system works better/worse for certain groups of people and why.	*Other Resource Consumption*: if and how the system provides a user-understandable explanation of other resource consumption of the system, e.g., in terms of raw material extraction or disposal at the end of life of the product. The explanation should clearly make examples and comparisons, e.g., in terms of land-use-related biodiversity loss or ecosystem damage.	
(Not) Harmfulness: If and how the producer states either that the use of the system is never harmful or describe the situations in which it could be. As an example, users should be informed that a personal LLM-based chatbot could cause harm if heavily jailbroken or improperly used.	*Responsibility and Accountability*: if and how users are informed about proper usage to prevent potential legal violations, such as the usage policies of LLMs. If and how providers indicate when the user is considered accountable or, more generally, responsible for the use of the system.	*Sustainability Promotion in Use*: if and how the system leverages sustainability potential, e.g., promoting sustainable products, incorporating sustainability into decision-making processes, or considering it when generating recommendations. For example, a search engine could rank results based on sustainability criteria in addition to other factors.	

Figure 6: Example of an Ethics Label.

Table 1
Summary of the four dimensions and focus areas of the Ethics Label for digital systems.

Dimension/Card	Scope & Focus Area
Human	Autonomy preservation, privacy, data governance, explainability, beneficialness, and non-harmfulness.
Societal	Impact on democracy, inclusive/participatory design, cultural sensitivities, diversity, and accountability.
Environmental	Energy consumption, greenhouse gas footprint, water usage, and life-cycle sustainability.
Management of Adverse Effects	Risk minimization (e.g., HAZOP protocols), failure mitigation, and user issue reporting mechanisms

6. Conclusion, Research Challenges and Future Directions

Ethical behavior in digital environments cannot be achieved solely through static regulation or hard-coded algorithmic constraints. It requires a socio-technical transformation centered on **adaptive interaction, transparent communication, and explicit software architectures**. By establishing frameworks like E-MAPE-K, software exoskeletons, and standardized ethics labeling, software engineering can transition from building systems that merely obey rules to engineering digital environments that safeguard human dignity, autonomy, and values.

Key elements to make this transition possible are:

- It takes both humans and systems to be ethical: both perspectives need to be considered.
- Interactions (and architectures) are the software engineering technical focus.
- System adaptiveness is the key technical property for **ethical-aware autonomous systems.**
- **Individual (soft) ethics** is a key interdisciplinary challenge.
- It is a **multidisciplinary effort** that connects **conceptual and philosophical inquiry, psychology, sociology, law**, and the **perspectives of different stakeholders** with concrete **engineering strategies, architectures, solutions, and tools** for building better digital systems.

Acknowledgements

This work synthesizes research conducted within the multidisciplinary **Exosoul** research initiative at the Gran Sasso Science Institute (GSSI), involving contributions from computer scientists, philosophers, sociologists, psychologists, and legal scholars.

Declaration on Generative AI

During the preparation of this work, the author used Google Gemini to assist with image generation (Figure 3), narrative structure, and language editing. After using this tool, the author

reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Autili, R. Corsi, M. De Sanctis, P. Inverardi, P. Pelliccione, Ethics label for digital systems to promote transparency and user awareness, Journal of Systems and Software (2026).

[2] L. Floridi, Soft ethics and the governance of the digital, Philosophy & Technology 31 (2018) 1–8.

[3] P. Inverardi, The European perspective on responsible computing, Communications of the ACM 62 (2019) 64.

[4] M. Autili, M. De Sanctis, P. Inverardi, M. A. Memon, P. Pelliccione, S. Pettinari, A reference architecture for ethical-aware autonomous systems, Journal of Systems and Software 235 (2026) 112749.

[5] M. Autili, D. Di Ruscio, P. Inverardi, P. Pelliccione, M. Tivoli, A software exoskeleton to protect and support citizen's ethics and privacy in the digital world, IEEE Access 7 (2019) 62011–62021.

[6] D. Donati, S. Gozzano, P. Inverardi, N. Troquard, Individual ethics and dispositions in the digital world, AI & Society 41 (2026) 4819–4833.

[7] M. A. Memon, G. Filippone, G. L. Scoccia, M. Autili, P. Inverardi, RobEthiChor: Automated context-aware ethics-based negotiation for autonomous robots, Journal of Systems and Software 232 (2026) 112692.