From Ethics to effective Legal Protection: The Role of Legal Protection by Design in the EU AI Act

Thiago Guimarães Moraes
School of Law and Criminology
Vrije Universiteit Brussel (VUB), Brussels, Belgium
Universidade de Brasília (UnB), Brasília, Brazil
thiago.guimaraes.moraes@vub.be

Abstract

This paper examines two "by design" approaches relevant to the EU AI Act's effort to balance the risks and opportunities of artificial intelligence (AI): Ethics by Design (EbD) and Legal Protection by Design (LPbD). EbD seeks to embed ethical principles in AI development but lacks enforceability and mechanisms for closure and balancing, limiting its effectiveness in safeguarding fundamental rights. LPbD, by contrast, incorporates legally binding protections into technology, providing enforceable and systematic safeguards. The paper analyses the interplay between these approaches, showing how LPbD addresses EbD's shortcomings while strengthening accountability under the AI Act. It concludes that EbD can play a complementary role, but LPbD must remain central to ensuring that trustworthy AI respects fundamental rights.

1 Introduction

The European Union's Regulation 2024/1689, also known as the AI Act,[1] aims to find a balance between AI risks and opportunities. Its objectives include ensuring a high level of protection for health, safety, democracy, and fundamental rights, while simultaneously fostering innovation and competitiveness in the internal market. The middle ground presented is the notion of "human-centric and trustworthy AI." According to the European Commission's High-Level Expert Group on AI (AI HLEG), trustworthy AI must be lawful, ethical, and robust. [2]

But how can trustworthy AI be achieved? Among the various approaches to embedding values in technological systems, "by design" strategies have gained traction. These include Privacy by Design,[3] Security by Design,[4] Transparency by Design,[5] and, more broadly, Ethics by Design (EbD).[6] Also, Mireille Hildebrandt has proposed Legal Protection by Design (LPbD),[7] which emphasizes embedding legal protections and the Rule of Law into technology itself.

This paper explores the interplay between EbD and LPbD in the context of the AI Act. While EbD contributes ethical principles, it struggles to provide enforceable safeguards for fundamental rights. Ethical frameworks may be perceived as too "soft" to offer meaningful fundamental rights protection and are also prone to "ethics washing" both by private and public actors, where ethical guidelines are promoted without substantial enforcement mechanisms.[8]

That is why Hildebrandt proposed a new "by design" approach which she entitles as Legal Protection by Design (LPbD), that focus on how to enable legal protection provided by fundamental rights and the checks and balances of the Rule of Law in the design of digital technologies.[7] Hildebrandt contrasts LPbD with EbD, emphasizing that while EbD aligns technology design with ethical values, LPbD envisions embedding legally binding protections into technology.[9] Although

there may be overlap between some legal and ethical requirements, Hildebrandt reminds us that the two are not synonymous: law and ethics offer distinct approaches and have their own limitations. As she states: ethics is both more and less than law.[10]

This paper explores the interplay between EbD and LPbD in the context of developing trustworthy AI, with a particular focus on the EU AI Act. It will be structured as follows: Section 2 will explore the concept of ethics by design in the context of AI governance; section 3 will focus on the limitation of EbD regarding its implementation; section 4 will discuss how Legal Protection by Design addresses EbD's limitations. Finally, in Section 5, I analyse LPbD's contribution to the AI Act.

2 Ethics by Design in the context of AI Governance

Ethics by Design is an approach that integrates ethical principles into the design and development of technology. It relates to an even broader concept, that of Value Sensitive Design (VSD), a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process.[11] In this sense, EbD can be understood as one particular example of VSD, and it has become an approach that has been often mentioned in the context of responsible innovation on AI technologies.

In AI governance, EbD gained prominence with the EU's Horizon 2020 projects SHERPA[12] and SIENNA,[13] which developed frameworks linking ethical requirements to AI system design. These frameworks were later operationalized in the AI HLEG's guidelines and the European Commission's 2021 guidance, which mapped 63 specific tasks across the AI lifecycle.[14]

In 2023, Brey and Dainow updated this framework into" Ethics by Design for AI" or EbD-AI, [6] proposing design requirements that operationalize ethical principles into actionable steps. These guidelines encourage AI developers to embed values such as transparency, fairness, and accountability into their systems.

However, Ethics by Design remains limited as an approach. While it specifies what should be done, it refrains from prescribing how to implement measures, leaving interpretation to developers.[6] This flexibility can be beneficial for innovation but also opens the door to inconsistent application and "ethics washing," where values are invoked rhetorically without enforceable consequences.[8]

3 Limitations of Ethics by Design on ensuring effective legal protection

While EbD is useful for guiding responsible innovation, it falls short in ensuring effective protection of fundamental rights. Legal protection can be understood as an affordance offered by the Rule of Law established in democratic societies.[1] By its turn, ethics lacks some essential elements to ensure the protection of fundamental rights. Five key limitations illustrate why:

1st limitation: weak linkage to fundamental rights. Ethical values do not always directly relate to fundamental rights. For instance, while "privacy" corresponds to the right to privacy, values like "accountability" or "technical robustness" lack clear parallels to other rights. A possible approach to better understand if and how ethics may contribute to the protection of fundamental rights would be a mapping exercise that points out how each ethical value contributes (or interferes) to fundamental rights established in relevant legal frameworks, such as the European Union's Charter of Fundamental Rights (EUCFR). In Table 1, I present a starting point for this analysis by comparing the EUCFR with the EU AI HLEG ethical principles, as mentioned in the Recital 27 of the EU AI Act. This comparison does not aim to be comprehensive.

[1] Affordances are properties of a specific environment. In this case, legal protection is considered a property of democratic societies and their states structured upon the rule of law. See: [9] P. 120.

Table 1: Draft mapping exercise of Ethical principles proposed by the EU AI HLEG and some related fundamental rights of the EUCFR

EU AI HLEG Ethical principle	EUCFR
Human Agency and Oversight	Human dignity (Article 1); Prohibition of slavery and forced labour (Article 5); Freedom rights (Article 6 to 19); right to vote (Articles 39 and 40)
Technical Robustness and Safety	Integrity of the person (Article 3); Fair and just working conditions (Article 31); Consumer protection (Article 38)
Privacy, Data Protection and Data Governance	Respect for private and family life (Article 7), Protection of personal data (Article 8)
Transparency	Freedom of expression and information (Article 11); Access to documents (Article 42)
Diversity, Non-discrimination and Fairness	Equity rights (Articles 20 to 26); Justice rights (Articles 47 to 50)
Social and Environmental Well-being	Solidarity rights (Articles 27 to 38)
Accountability	Rights to good administration (Article 41); Ombudsman (Article 43)

2nd limitation: fragmented ethical theories. Ethics encompasses multiple, often conflicting approaches. The trend on "AI ethics" seems to ignore (naively or purposively) that there are different theories and approaches to Ethics. Therefore,: there is no unified ethical framework for fundamental rights protection. Hildebrandt warns how utilitarian reasoning may justify outcomes that harm minorities. [10] For example, an aggregate utilitarian perspective may conclude that unfair bias brings more utility to society even if at the cost of some minority or underrepresented communities and groups of individuals. As for deontological approaches she cautions that assuming a rational universal consensus is problematic, because people have different ideas about the value of their interests and how to prioritize them. Other approaches such as virtue ethics and pragmatic ethics seem to be better fit to address current challenges on AI. Virtue ethics emphasizes moral character and accountability, and pragmatist ethics highlights contextual, iterative reasoning. [15] However, like all other ethical approaches, they lack consistent enforceability, the third issue to be discussed.

3rd limitation: lack of enforceability. Unlike legal norms backed by courts and regulatory bodies, ethical norms depend on voluntary adherence. Violations may provoke disapproval but rarely carry tangible consequences. As argued by Habermas, ethics operates within the realm of deliberation and mutual agreement, relying on intrinsic motivation rather than external compulsion. [16] Legal norms, on the other hand, have "legal teeth", and they can be backed up with the force of the law. [7] The enforceability of law and legal norms lies in their binding nature, backed by institutional mechanisms such as courts, regulatory agencies, audition bodies and police forces. [17] These mechanisms ensure compliance through sanctions, fines, or other legal consequences, providing a framework for accountability. For example, Hart emphasizes that legal norms derive their authority from formal recognition within a legal system and the capacity to impose sanctions upon violations.[18]

4th limitation: lack of closure. Ethics is inherently open-ended. While both ethics and law share contestability (due to the multiple interpretations a given norm may have - be it ethical or legal – they allows for subjects of these norms to contest their interpretation, and their own actions as qualified in light of these norms),[9] only legal norms have closure, which permits the conclusion of a given contest. Multiple ethical theories can lead to conflicting solutions without authoritative resolution. By contrast, legal systems provide closure through judicial and regulatory processes, ensuring stability and legal certainty.[9] This is an important feature of legal frameworks, necessary

for individuals and society to coordinate their life and goals under a democratic state protected by the Rule of Law.[9]

5th limitation: lack of effective balancing mechanisms. While ethics may acknowledge the importance of human-centred values such as privacy, fairness and transparency, it does not provide a clear answer on what to do when these values clash. Should privacy always come first? Or fairness? What if for ensuring more privacy you must lose some transparency, or vice-versa? Is there an optimal point? Who has the final say on that?

There are several research that explore the trade-offs between ethical values. Take for example, fairness issues when privacy enhancing technologies (PETs) are implemented in machine learning. While PETs, such as differential privacy and federated learning, aim to protect data privacy during model training, their deployment can propagate biases and entrench inequalities.[19] It is also important to consider that in the computer science community there are different notions of fairness and their impact on privacy (and utility) may differ. Alves *et al.* analysed how there are tensions between different fairness notions (i.e. an AI model usually cannot satisfy different notions, such as individual-based or group-based simultaneously), as well as between these notions and privacy or accuracy.[20] They prove, through a theoretical analysis that it is arguably impossible of satisfying simultaneously fairness metrics and differential privacy while keeping a non-trivial accuracy.

Other works worth mentioning that attempted to find a better trade-off between privacy, fairness and utility in AI systems are from Cummings *et al*.,[21] which proposes learning differentially private classifiers that approximate fairness and privacy while maintaining utility; Gu *et al*.,[22] which demonstrated that carefully tuning differential privacy parameters can mitigate unfair outcomes in federated learning models, without overly sacrificing utility; and Cheng *et al*., which examines the use of differentially private generative adversarial networks to create synthetic data that balances privacy, utility, and fairness in machine learning. There is also research that show the clashes between privacy and cybersecurity,[23] privacy and transparency,[24] and fairness and transparency.[25] While their goal is to mitigate these trade-offs, the main issue is that ethical values cannot be properly measured. Furthermore, while we can talk about trade-offs between technical indicators fundamental rights cannot be traded off. Therefore, there is a mismatch on how to balance between ethical values (and AI system's utility) and how to ensure that the fundamental rights related to these values are effectively protected. While ethics highlights these tensions, it lacks structured procedures to resolve them - due to moral flexibility, context strongly influence which moral beliefs are applied in a situation[26].

The contrast between mechanisms for balancing in law and ethics reveals key differences in their structure and implementation. Legal frameworks are designed with robust mechanisms to balance competing interests systematically and transparently. For instance, Robert Alexy's proportionality test in constitutional law ensures that limitations on fundamental rights are justified, necessary, and proportionate to achieve a legitimate aim.[27] Another interest example of balancing exercise in law, is the Legitimate Interest Assessment (LIA) under data protection law, which balances a data controller's legitimate interests with the rights and freedoms of data subjects, providing a structured, accountable process. While not explicitly mentioned in the EU GDPR, it is widely adopted and acknowledge by the European Data Protection Board.[28] What is particular interesting of LIA, is that its assessment allows for contrasting several rights and freedoms (which may be associated with different ethical values) with controller's legitimate interests that can either represent their owns rights (i.e. intellectual property rights) or economic interests (i.e. economy, efficiency and effectiveness[29]). However, it must be clear that fundamental rights must always have the upper hand. That is why Article 6(1)(f) states that data controllers' legitimate interests are overridden by the interests or fundamental rights and freedoms of data subjects.

In short, these limitations show that Ethics by Design, while normatively significant, cannot on its own guarantee enforceable and balanced protection of fundamental rights. This highlights the essential role of law in reinforcing ethical considerations with structured and enforceable balancing mechanisms.

4 How Legal Protection by Design addresses EbD's issues

Prof. Mireille Hildebrandt developed the concept of Legal Protection by Design (LPbD) to address the issue of enabling effective legal protection of fundamental rights in the design of ICT infrastructure.[30] This approach addresses the issues of EbD. LPbD is embedded within legal frameworks, thus solving the first two issues – since it derives from established legal frameworks, it is inherently connected to the protection of fundamental rights, avoiding the ambiguity of loosely defined ethical values. Second, being constrained by the principles and procedures of the Rule of Law, LPbD mitigates the problem of fragmented or conflicting ethical theories, ensuring consistency and coherence in the protection of rights.[9] The three other issues of Ethics – lack of enforceability, lack of closure and lack of effective balancing mechanisms - are explored in more details in the next paragraph.

4.1 LPbD and enforcement

The enforceability of legal norms is a cornerstone of the legal protection of fundamental rights. Enforcement is ensured by (i) the binding character of positive law, (ii) the state's actual power to enforce it and (iii) the decision-making of state actors (e.g. legislators, public administration and judicial courts) on legal norms, either by issuing, interpreting or applying them.[9]

According to Hildebrandt, LPbD is an enabler for the Rule of Law in democratic constitutions, and in this way it's possible to argue that her vision connects better with John Rawls perspective. Rawls adopts a pluralist approach to legal theory by emphasizing the importance of reasonable pluralism in democratic societies. He argues that although having diverse, often conflicting, comprehensive doctrines (religion, moral, traditions, etc.), citizens can still share a commitment to toleration and respect for democratic principles.[31] Therefore, to him, enforcement's role is to maintain order and uphold justice, ensuring compliance with laws rooted in fairness and equality.[32]

It is important to consider that enforceability is not directly ensured by LPbD but by the Rule of Law (which is sustained by positive law).[9] What LPbD does is to ensure that ICT infrastructure is built with features that will better allow for the Rule of Law to be enforced thus providing for the protection of fundamental rights in the contexts these technologies are deployed. This can be done by assuring that legal requirements are present in the design of ICT systems, thus providing for the implementation of some legal rules directly by the technology.

Such approach relates to Lessig's argument that State must not only be able to regulate the Code, using the Law to establish rules of conduct for how the Code should be developed and used (indirect regulation, or regulating the Code); but also "regulate by the Code", that is, act directly in the development of digital systems that meet your interests.[33] However, it must be made clear that LPbD is not equivalent to Lessig's theory, since he looks at four normative forces as silos, in which code is one of them (the others are - social, economic and legal norms). In Hildebrandt's theory, these four forces overlap and must be considered together.[9]

Also, it must not be confused with 'legal by design', which refers to a specific type of techno-regulation, whereby legal norms are translated into code or into the design of computing systems such that compliance become automated or semi-automated.[9] Implementing legal requirements in software's architecture does not mean that law will automatically be enforced and met. Legal by design impedes that contestability, closure and balancing exercises can be performed outside software. This ultimately undermines the Rule of Law.[34] LPbD means that code will be written in a way that enables the enforcement of law, by constraining behaviour but while still allowing for these other features of legal protection to take place.

For instance, consider cookie banners. They are a very interesting example of LPbD applied in the context of data protection law. Cookie banners are designed in a way that aims to respect the principle of lawfulness of data protection legislations such as the GDPR. By doing so, they enable the fulfilment of (some of) GDPR's legal basis on Article 6, such as consent and legitimate interest. They constrain behaviour of the system (which allegedly will process personal data according to the settings of the cookie banner), and the users (which will react to the banner settings and may subject themselves to be targeted by certain ads, for example). But they never

ensure legal protection by itself. Several Data Protection Authorities, such as France's CNIL,[35], the UK's ICO,[36] and Brazil's ANPD,[37] have already provided guidelines to assist data controllers on how to design cookie banners and differ the "good" banners from the "bad" ones. There is also research on how deceptive patterns have been continuously being used in cookie banners, constraining user behaviour in a deceitful way, and leading them to accept more cookies than needed and more than they are aware of.[38] This is one of the reasons that legal by design can never be the answer. LPbD remembers us that regardless of how much code attempts to automate legal requirements, organisational measures (e.g., oversight mechanisms), which go beyond code, will always be needed.

4.2 LPbD and closure

Legal Protection by Design synergizes with Habermas and Dworkin perspectives on the closure of law. It ensures ICT systems embed compliance and accountability mechanisms, which reflects Habermas's view that legitimacy arises from democratic procedures enabling inclusive, rational discourse. Legal closure occurs when norms are validated through public deliberation, producing laws that are both binding and justified.[39] For Habermas, this participatory process aligns law with societal values, ensuring its adaptability to changing conditions. LPbD operationalizes these values, by considering legal norms in the development of ICT systems, and by facilitating multistakeholder participation in its governance, akin to Habermas's communicative rationality.[39]

Furthermore, LPbD contributes to Dworkin's theory of adjudication, which emphasizes law as an interpretive practice guided by principles of justice, fairness, and integrity.[40] Dworkin argues that judges should not merely apply rules but instead find the "best fit" between existing legal materials (statutes, precedents) and moral principles. When hard cases arise from conflicts involving ICT systems, LPbD support judges and regulators with practical evidence of implemented technical and organizational measures based on legal safeguards. For example, it can demonstrate how a system's design aligns with fairness principles, aiding judicial interpretation. LPbD thus helps make ICT systems responsive to democratic processes and judicial oversight.

4.3 LPbD and balancing mechanisms

Finally, LPbD also enables structured balancing exercises. Tools such as Data Protection Impact Assessments (DPIAs), Human Rights Impact Assessments (HRIAs), and Fundamental Rights Impact Assessments (FRIAs) operationalize balancing by requiring systematic evaluation of risks and stakeholder engagement.[41], [42]. Impact assessments create a structured process requiring organizations to systematically evaluate potential risks and impacts of their technologies on fundamental rights. This fosters compliance with legal norms and principles, particularly in contexts like the GDPR, where DPIAs ensure that privacy and personal data protection are treated not as an afterthought but as a central concern in technological development.[42] Similarly, HRIAs help ensuring that technologies are aligned with human rights frameworks, preventing them from enabling discrimination, surveillance, or other rights violations.[41] FRIAs are the EU AI Act version of HRIAs, thus being focused specifically on AI systems impact assessment and following the provisions of its Article 27.[41]

These processes can also promote a balance of information and power: although not a legal obligation, there are recommendations from the European Data Protection Board (EDPB) on transparency and consultation with stakeholders, including civil society and affected individuals, which limits unilateral decision-making on ICT systems behaviour by technology developers.[43] For example, the former GDPR's Article 29 Working Party (29WP) recommended that DPIAs should be made public to better inform the affected public about the processing operation.[44] These guidelines were endorsed by the European Data Protection Board – EDPB.[43] Furthermore, in Recital 96 of the AI Act, it is recommended that FRIAs involve relevant stakeholders, including the representatives of groups of persons likely to be affected by the AI system, independent experts, and civil society organisations in conducting such impact assessments and designing measures to be taken in the case of materialisation of the risks.

Impact assessments mitigate risks associated with the unpredictable evolution of data-driven technologies, such as AI. In doing so, they promote democratic legitimacy and public trust, ensuring that technological governance remains accountable to the societies it serves. Thus, assessments such as LIA, DPIAs and FRIAs are clear examples of how legal protection by design can enable balancing mechanisms that aligns innovation with social justice and the rule of law.

5 LPbD in the AI Act

The EU AI Act is the first comprehensive attempt worldwide to regulate AI systems with legally binding obligations.[1] It represents a clear move away from non-binding ethical guidelines toward enforceable requirements grounded in the Rule of Law. In this context, the Act can be understood as an expression of Legal Protection by Design (LPbD), since it integrates enforceability, closure, and balancing mechanisms into the lifecycle of AI systems. Importantly, the AI Act establishes general principles and operationalizes them through legal obligations for providers, deployers, and regulators.

5.1 Enforcement in Practice

One of the clearest manifestations of LPbD in the AI Act is the requirement for data and data governance practices for high-risk AI systems (Article 10). Providers must ensure that training, validation, and testing datasets are relevant, representative, and free of errors and bias, as far as possible. The AI Act is explicit how training, validation and testing data sets shall be subject to data governance and management practices appropriate for the intended purpose of the high-risk AI system, practices which must consider relevant design choices to ensure the conditions expressed in Article 10(2) are met. These obligations go beyond abstract principles of "fairness" or "non-discrimination," translating them into enforceable duties.

The practical impact of these rules can be illustrated by recent debates on data scraping practices. Data protection authorities, including members of the Global Privacy Assembly, have stressed the importance of technical measures such as rate-limiting, bot-detection, and CAPTCHAs to mitigate unlawful scraping.[45] Also, organisational measures such as assigning specific teams to oversee anti-scraping controls, pursuing legal remedies against infringers, and, where relevant, notifying affected individuals and regulators in case of breaches have been recommended.[45] These examples highlight how technical and organisational measures may be embedded in systems to enable the enforceability of legal protection on the rights to privacy and data protection.

More broadly, enforcement in the AI Act is supported by the threat of administrative fines and market surveillance mechanisms, which ensure that compliance is legally binding (Chapter IX of the AI Act). This layer of enforceability provides AI providers with clarity while ensuring individuals and society enjoy actual, not symbolic, protection of their rights.

5.2 Closure through Compliance Mechanisms

Moving on to closure, this can be easier understood by looking at governance elements of the AI Act. The AI Act incorporates such mechanisms by ensuring that obligations are not open to endless contestation but can be resolved through authoritative procedures. A central instrument here is the conformity assessment (Article 43). Providers of high-risk AI systems must demonstrate compliance with legal requirements before placing systems on the market. Depending on the use case, this assessment may take the form of self-assessment or third-party evaluation by a "notified body."

These conformity assessments operationalize LPbD by making AI providers accountable for their design choices, subject to regulatory oversight. Even when self-assessment is allowed, providers must maintain detailed technical documentation (Article 11), which can be requested and reviewed by authorities. This creates a procedural avenue for contestation: if regulators or stakeholders question compliance, there is a structured process to review, challenge, and resolve disputes.

Furthermore, record-keeping obligations (Article 12) and post-market monitoring systems (Article 61) ensure that accountability extends beyond the design phase into deployment. In this way, the AI Act ensures that legal oversight is not a one-off but a continuous process, and providers must maintain records of system operation and monitor ongoing risks. This reflects LPbD's principle that legal protection must remain in place throughout a system's lifecycle.

5.3 Balancing Mechanisms: Risk Management and Rights Protection

A third pillar of LPbD in the AI Act lies in its balancing mechanisms. High-risk AI systems must implement a risk management system - RMS (Article 9), which is iterative and continuous, running throughout the lifecycle of the system. The RMS requires providers to identify foreseeable risks to health, safety, and fundamental rights, and to adopt proportionate mitigation measures.

This approach raises questions about whether fundamental rights can be subject to "risk-based" reasoning. However, as in data protection law, the purpose of such a framework is not to weigh rights against efficiency but to calibrate the level of legal obligations according to the level of risk.[46] Fundamental rights remain non-negotiable; what varies is the number and intensity of obligations imposed on providers. Thus, a low-risk AI system may have fewer duties, while high-risk systems face a more extensive set of obligations.

Beyond RMS, the AI Act also introduces Fundamental Rights Impact Assessments - FRIAs (Article 27), required for certain deployments of high-risk AI, such as by public authorities or in sensitive domains like credit scoring. FRIAs ensure that providers systematically evaluate the potential impact of their systems on rights and freedoms before deployment.[41] Recital 96 even encourages stakeholder engagement, recommending consultation with civil society organizations, independent experts, and representatives of affected groups - reflecting LPbD's commitment to procedural fairness and the democratization of risk assessment.

6 Contrasting Ethics by Design and Legal Protection by Design through a hypothetical AI system

To clarify the implications of Ethics by Design (EbD) and Legal Protection by Design (LPbD), this section contrasts the two approaches through a hypothetical high-risk AI system. The example concerns an AI-based decision-support system used by a public authority to assess eligibility for social benefits. This scenario is inspired by the Dutch childcare benefit scandal ("*Toeslagenaffaire*") in 2021, in which a tax authority used an algorithmic risk-classification system to detect potential fraud in childcare benefit claims, treating factors such as nationality and dual citizenship as risk indicators [47]. This led to the wrongful recovery of benefits, severe financial hurdles for tens of thousands of families, and, in some cases, the removal of children from their homes. The resulting political crisis culminated in the resignation of the Dutch government in January 2021 and exposed how opaque and discriminatory algorithmic decision-making can undermine constitutional rights and the rule of law.

In our hypothetical system, a public authority from an EU Member State employs an AI model to support decisions about eligibility for social benefits, processing personal data to classify applicants as eligible or not. The system lifecycle comprises design and training, deployment, monitoring and review, and modification or withdrawal; this mirrors the stages at which the Dutch childcare algorithm was built, executed and sustained without adequate safeguards. High-impact decisions affecting fundamental rights — including the right to social security, non-discrimination, and due process — make such systems "high risk" by regulatory standards.

Under the EU AI Act, such a system would likely fall under Annex III of the AI Act, given its direct impact on access to social security, non-discrimination, human dignity, and the right to good administration. The system processes personal and socio-economic data and supports decisions taken by caseworkers throughout its lifecycle, from design and deployment to monitoring and review.

Following an Ethics by Design approach, the development and deployment of this system would be guided by ethical values such as fairness, transparency, and social well-being. During the design phase, EbD would encourage bias-aware data selection, internal testing for discriminatory

outcomes, and the inclusion of explainability features to make outputs intelligible to human operators. At the organizational level, EbD might also promote ethical guidelines, internal review boards, and training programmes for caseworkers to ensure responsible use. These measures can meaningfully influence design choices and organizational culture, particularly in early stages of development.

However, when ethical tensions arise, EbD offers limited tools to address them. Conflicts between efficiency (automating large volumes of applications), fairness (avoiding indirect discrimination), and transparency (explaining complex models) are left to internal interpretation and discretionary judgment. EbD does not specify who has the authority to decide which value should prevail, nor how such decisions can be challenged. Ethical commitments are typically voluntary, and their violation rarely triggers formal consequences. As a result, affected individuals have few avenues to contest outcomes beyond informal complaints, political pressure, or ex post reputational accountability.

Legal Protection by Design approaches the same AI system from a fundamentally different perspective. Rather than embedding values alone, LPbD seeks to embed legal safeguards that enable the Rule of Law to operate throughout the system's lifecycle, including after deployment. At the design stage, this requires anticipating legally binding obligations and translating them into technical and organisational measures that will later allow for effective legal protection. In the context of a public benefits eligibility system, this entails compliance with the GDPR and the AI Act as enforceable legal duties. Under the AI Act, this includes, in particular, the requirements set out in Section 2 of Chapter III for high-risk AI systems.

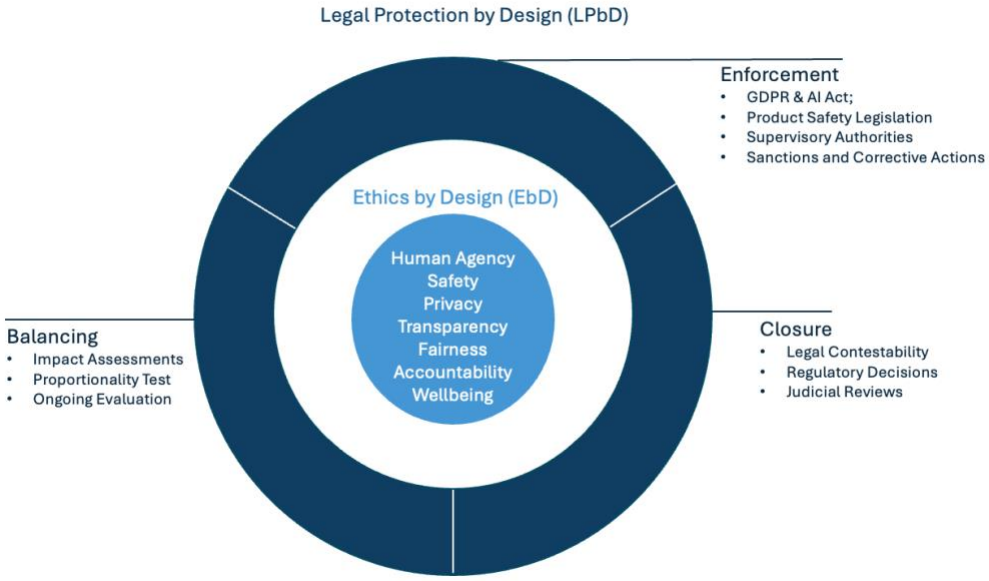
Before deployment, LPbD activates mechanisms of closure and balancing through legally defined procedures. Conformity assessments (Article 43), Fundamental Rights Impact Assessments (Article 27) and the GDPR's Data Protection Impact Assessments (Article 35) function as structured pre-deployment checkpoints that assess whether legal requirements have been met and whether risks to fundamental rights have been adequately identified and mitigated. At this stage, impact assessments play a central role as balancing mechanisms: they require providers to systematically evaluate their own legitimate interests (such as utility or efficiency) against the potential impact of the system on individuals' fundamental rights, including data protection, non-discrimination or health. To support future legal protection, design choices must also preserve contestability by enabling features such as logging, audit trails, and meaningful human oversight. These features do not themselves resolve disputes but ensure that, if a dispute arises, there is an evidentiary and procedural basis upon which it can be assessed by competent authorities.

Once the system is deployed, LPbD becomes fully operational through its interaction with the real world under the Rule of Law. Enforcement is no longer hypothetical: legal obligations under the GDPR and the AI Act must be actively enforced by supervisory authorities, under the rules of Chapter IX, Section 3. In the case of a public benefits eligibility system, this means that systemic errors, discriminatory outcomes, or unlawful data processing can trigger investigations by data protection or market surveillance authorities, independently of the provider's internal assessments. Human oversight ceases to be a matter of organisational ethics and becomes a legally structured requirement, while affected individuals retain legally guaranteed rights to lodge complaints to the relevant market authority, without prejudice to other administrative or judicial remedies (Article 85), and to obtain explanation of individual decision-making (Article 86). Complementarily, data subjects rights under the GDPR may also be requested. In cases of conflict, disputes between AI providers or deployers and affected individuals must be capable of reaching supervisory authorities or judicial bodies, where authoritative decisions can be taken. It is at this stage that closure is achieved, not by the system itself, but through institutional decision-making grounded in law.

Finally, balancing is not frozen at impact assessments but may be reassessed in light of real-world effects. Where, for instance, the deployment of an eligibility system disproportionately excludes certain groups from access to social benefits, courts and authorities may evaluate whether such outcomes constitute unjustified limitations on fundamental rights. This evaluation can rely on established legal mechanisms, such as Alexy's proportionality test. In this way, LPbD ensures that AI systems remain governable within democratic legal orders, where enforcement, closure, and balancing are exercised by legitimate institutions, instead of being arbitrary ethical choices embedded in technical systems.

The contrast between EbD and LPbD can be visualised in a simplified regulatory workflow (Figure 1). Under EbD, governance primarily operates through internal ethical feedback loops, with limited external accountability. Under LPbD, governance is structured around legally binding obligations, external oversight, and enforceable remedies. While EbD can meaningfully guide design choices, only LPbD ensures that norms are translated into enforceable, contestable, and balanced protection of fundamental rights in high-risk AI systems.

Figure 1: Embedding values in AI systems through Ethics by Design and Legal Protection by Design



7 Conclusion

Ethics by Design and Legal Protection by Design represent two complementary but distinct approaches to embedding values in AI systems. While EbD fosters awareness of ethical principles, it lacks enforceability, closure, and structured balancing mechanisms. LPbD, by contrast, grounds AI governance in the Rule of Law, ensuring that fundamental rights are not left to voluntary interpretation.

The AI Act illustrates how LPbD principles can be operationalized through enforceable obligations, accountability mechanisms, and structured balancing procedures. At the same time, EbD still has a complementary role, helping developers and deployers identify relevant principles and align them with legal obligations.

Looking forward, regulatory sandboxes under the AI Act may provide valuable environments for testing how EbD and LPbD can be combined to promote responsible innovation. Embedding both ethical reflection and legal safeguards into sandbox experimentation can help ensure that AI development is not only innovative but also fundamentally rights-respecting and trustworthy.

Nevertheless, practitioners must always bear in mind that their goal should be to protect fundamental rights when developing and deploying AI systems, and this cannot be done solely by considering ethics. The Rule of Law must be respected and for this Legal Protection by Design is key.

References

[1] European Union, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)Text with EEA relevance. Accessed: July 20, 2024. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[2] European Commission, "Ethics Guidelines For Trustworthy AI," 2021. Accessed: Dec. 20, 2023. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[3] A. Cavoukian, "Privacy by Design - The 7 Foundational Principles," Information & Privacy Commissioner, Ontario, 2011. Accessed: Dec. 20, 2023. [Online]. Available: https://iapp.org/media/pdf/resource_center/pbd_implement_7found_principles.pdf

[4] S. Gupta, "Towards secure-by-design artificial intelligence systems," Aug. 26, 2024, Toronto, Itally. Accessed: Nov. 06, 2024. [Online]. Available: https://www.authorea.com/users/686249/articles/1216496-towards-secure-by-design-artificial-intelligence-systems

[5] H. Felzmann, E. Fosch-Villaronga, C. Lutz, and A. Tamò-Larrieux, "Towards Transparency by Design for Artificial Intelligence," Sci Eng Ethics, vol. 26, no. 6, pp. 3333–3361, Dec. 2020, doi: 10.1007/s11948-020-00276-4.

[6] P. Brey and B. Dainow, "Ethics by design for artificial intelligence," AI Ethics, Sept. 2023, doi: 10.1007/s43681-023-00330-4.

[7] G. Gori, "Legal Protection by Design in the AI Value Chain What role for AI Metrics?," VUB, Brussels, Aug. 2024. Accessed: June 11, 2024. [Online]. Available: https://www.humane-ai.eu/wp-content/uploads/2024/08/HAI-NET-Report-Metrics_VUB_Gianmarco_Gori_2.pdf

[8] G. van Maanen, "AI Ethics, Ethics Washing, and the Need to Politicize Data Ethics," DISO, vol. 1, no. 2, p. 9, Aug. 2022, doi: 10.1007/s44206-022-00013-3.

[9] L. Diver, T. Duarte, G. Gori, E. van den Hoven, and M. Hildebrandt, "Cohubicol: Research Study on Text-Driven Law," VUB, Brussels, Sept. 2023. Accessed: June 11, 2024. [Online]. Available: https://publications.cohubicol.com/assets/uploads/cohubicol-research-study-on-text-driven-law-final.pdf

[10] M. Hildebrandt, "Closure: On Ethics, Code, and Law," in Law for Computer Scientists and Other Folk, 1st ed., Oxford University PressOxford, 2020, pp. 283–318. doi: 10.1093/oso/9780198860877.003.0011.

[11] B. Friedman, P. H. Kahn, and A. Borning, "Value Sensitive Design: Theory and Methods," 2001.

[12] "Project Sherpa – Shaping the Ethical Dimensions of Smart Information Systems a European Perspective." Accessed: Oct. 17, 2024. [Online]. Available: https://www.project-sherpa.eu/

[13] "SIENNA." Accessed: Oct. 17, 2024. [Online]. Available: https://www.sienna-project.eu/w/si

[14] European Commission, "Ethics By Design and Ethics of Use Approaches for Artificial Intelligence," European Commission, Brussels, Nov. 2021. [Online]. Available: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf

[15] C. Legg and C. Hookway, "Pragmatism," in The Stanford Encyclopedia of Philosophy, Winter 2024., E. N. Zalta and U. Nodelman, Eds., Metaphysics Research Lab, Stanford University, 2024. Accessed: Nov. 21, 2024. [Online]. Available: https://plato.stanford.edu/archives/win2024/entries/pragmatism/

[16] J. Habermas, Moral Consciousness and Communicative Action. MIT Press, 1990.

[17] C. H. (remco) V. Rhee and A. Uzelak, "Enforcement and Enforceability – Tradition and Reform," Jan. 2010, Accessed: Nov. 21, 2024. [Online]. Available: https://www.academia.edu/33857022/Enforcement_and_Enforceability_Tradition_and_Reform

[18] H. L. A. Hart and L. Green, The Concept of Law. OUP Oxford, 2012.
[19] A. Calvi, G. Malgieri, and D. Kotzinos, “The unfair side of Privacy Enhancing Technologies: addressing the trade-offs between PETs and fairness,” in Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, in FAccT ’24. New York, NY, USA: Association for Computing Machinery, June 2024, pp. 2047–2059. doi: 10.1145/3630106.3659024.
[20] G. Alves, F. Bernier, M. Couceiro, K. Makhlouf, C. Palamidessi, and S. Zhioua, “Survey on fairness notions and related tensions,” EURO Journal on Decision Processes, vol. 11, p. 100033, Jan. 2023, doi: 10.1016/j.ejdp.2023.100033.
[21] R. Cummings, V. Gupta, D. Kimpara, and J. Morgenstern, “On the Compatibility of Privacy and Fairness,” in Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, in UMAP’19 Adjunct. New York, NY, USA: Association for Computing Machinery, June 2019, pp. 309–315. doi: 10.1145/3314183.3323847.
[22] X. Gu, Z. Tianqing, J. Li, T. Zhang, W. Ren, and K.-K. R. Choo, “Privacy, accuracy, and model fairness trade-offs in federated learning,” Computers & Security, vol. 122, p. 102907, Nov. 2022, doi: 10.1016/j.cose.2022.102907.
[23] I. van de Poel, “Core Values and Value Conflicts in Cybersecurity: Beyond Privacy Versus Security,” in The Ethics of Cybersecurity, M. Christen, B. Gordijn, and M. Loi, Eds., Cham: Springer International Publishing, 2020, pp. 45–71. doi: 10.1007/978-3-030-29053-5_3.
[24] R. Meijer, P. Conradie, and S. Choenni, “Reconciling Contradictions of Open Data Regarding Transparency, Privacy, Security and Trust,” Journal of Theoretical and Applied Electronic Commerce Research, vol. 9, no. 3, Art. no. 3, Sept. 2014, doi: 10.4067/S0718-18762014000300004.
[25] J. Schoeffer, M. De-Arteaga, and N. Kuehl, “On the Relationship Between Explanations, Fairness Perceptions, and Decisions,” May 06, 2022, arXiv: arXiv:2204.13156. doi: 10.48550/arXiv.2204.13156.
[26] D. M. Bartels, C. W. Bauman, F. A. Cushman, D. A. Pizarro, and A. P. McGraw, “Moral Judgment and Decision Making,” in The Wiley Blackwell Handbook of Judgment and Decision Making, John Wiley & Sons, Ltd, 2015, pp. 478–515. doi: 10.1002/9781118468333.ch17.
[27] M. Jestaedt, “The Doctrine of Balancing—its Strengths and Weaknesses,” in Institutionalized Reason: The Jurisprudence of Robert Alexy, M. Klatt, Ed., Oxford University Press, 2012, p. 0. doi: 10.1093/acprof:oso/9780199582068.003.0007.
[28] European Data Protection Board - EDPB, “Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR | European Data Protection Board,” EDPB, 2024. Accessed: Nov. 21, 2024. [Online]. Available: https://www.edpb.europa.eu/our-work-tools/documents/public-consultations/2024/guidelines-12024-processing-personal-data-based_en
[29] K. Brunsson, “Economy, Efficiency, Effectiveness,” in The Teachings of Management - Perceptions in a Society of Organizations, 2017, pp. 37–43. doi: 10.1007/978-3-319-56120-2_4.
[30] M. Hildebrandt, “Legal Protection by Design: Objections and Refutations,” Legisprudence, vol. 5, no. 2, pp. 223–248, Oct. 2011, doi: 10.5235/175214611797885693.
[31] L. Wenar, “John Rawls,” in The Stanford Encyclopedia of Philosophy, Summer 2021., E. N. Zalta, Ed., Metaphysics Research Lab, Stanford University, 2021. Accessed: Nov. 25, 2024. [Online]. Available: https://plato.stanford.edu/archives/sum2021/entries/rawls/
[32] P. J. Rawls, A Theory of Justice: Revised Edition, 2nd Revised ed. edition. Cambridge, Mass: Belknap Press, 1999.
[33] L. Lessig, “Law Regulating Code Regulating Law,” Loyola University Chicago Law Journal, vol. 35, no. 1, Jan. 2003, Accessed: Dec. 20, 2023. [Online]. Available: https://lawecommons.luc.edu/luclj/vol35/iss1/2
[34] M. Hildebrandt, “‘Legal by Design’ or ‘Legal Protection by Design’?,” in Law for Computer Scientists and Other Folk, M. Hildebrandt, Ed., Oxford University Press, 2020, p. 0. doi: 10.1093/oso/9780198860877.003.0010.
[35] Commission Nationale de l’Informatique et des Libertés, “Cookies et autres traceurs : la CNIL publie des lignes directrices modificatives et sa recommandation.” Accessed: Nov.

27, 2024. [Online]. Available: https://www.cnil.fr/fr/cookies-et-autres-traceurs/regles/cookies/lignes-directrices-modificatives-et-recommandation
[36] Information Commissioner's Office - ICO, "Guidance on the use of cookies and similar technologies." Accessed: Nov. 27, 2024. [Online]. Available: https://ico.org.uk/for-organisations/direct-marketing-and-privacy-and-electronic-communications/guidance-on-the-use-of-cookies-and-similar-technologies/
[37] Autoridade Nacional de Proteção de Dados – ANPD, "Guia Orientativo - Cookies e proteção de dados pessoais," ANPD, Brasilia, Out 2022. Accessed: Nov. 26, 2024. [Online]. Available: https://www.gov.br/anpd/pt-br/documentos-e-publicacoes/guia-orientativo-cookies-e-protecao-de-dados-pessoais.pdf
[38] B. M. Berens, M. Bohlender, H. Dietmann, C. Krisam, O. Kulyk, and M. Volkamer, "Cookie disclaimers: Dark patterns and lack of transparency," Computers & Security, vol. 136, p. 103507, Jan. 2024, doi: 10.1016/j.cose.2023.103507.
[39] J. Habermas and W. Rehg, Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy, Reprint edition. Cambridge, Mass.: The MIT Press, 1998.
[40] F. H. S. P. of L. and P. R. Dworkin, Law's Empire, Reprint édition. Cambridge, Mass.: Belknap Press, 1988. Accessed: Dec. 16, 2025. [Online]. Available: https://www.filosoficas.unam.mx/~cruzparc/empire.pdf
[41] A. Mantelero, "The Fundamental Rights Impact Assessment (FRIA) in the AI Act: Roots, legal obligations and key elements for a model template," Computer Law & Security Review, vol. 54, p. 106020, Sept. 2024, doi: 10.1016/j.clsr.2024.106020.
[42] D. Kloza et al., "Data protection impact assessments in the European Union: complementing the new legal framework towards a more robust protection of individuals," d.pia.lab Policy Brief, pp. 1–4, May 2017, doi: 10.31228/osf.io/b68em.
[43] European Data Protection Board - EDPB, "Data Protection impact assessments High risk processing." Accessed: Nov. 25, 2024. [Online]. Available: https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/data-protection-impact-assessments-high-risk-processing_en
[44] Article 29 Working Party - WP29, "Guidelines on Data Protection Impact Assessment (DPIA) (wp248rev.01)," Article 29 Working Party - WP29, Brussels, Oct. 2017. Accessed: Nov. 25, 2024. [Online]. Available: https://ec.europa.eu/newsroom/article29/items/611236/en
[45] Information Commissioner's Office, "Joint statement on data scraping and data protection." Accessed: Nov. 27, 2024. [Online]. Available: https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2023/08/joint-statement-on-data-scraping-and-data-protection/
[46] Article 29 Working Party, "Statement on the role of a risk-based approach in data protection legal frameworks - WP 218," Article 29 Working Party - WP29, May 2014. Accessed: Nov. 29, 2024. [Online]. Available: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp218_en.pdf
[47] J. Arts and M. van den Berg, "What the Dutch benefits scandal and policy's focus on 'fraud' can teach us about the endurance of empire," Critical Social Policy, vol. 45, no. 1, pp. 177–187, Feb. 2025, doi: 10.1177/02610183241281346.