DIGIBOOKS

A. ALOISI | F. DE ELIZALDE | F. PALMIOTTO

TECHNOLOGY AND (DIS)TRUST

AUTHORS

A. BOGUCKI
F. FEDORCZYK
E. FOSCH-VILLARONGA
N. HŐS
P. MEINEL
T. NIV
M. D. F. RABAJANTE
B. SARILAR
M. R. SHAFFIQUE
C. TOSCANO
M. B. UNVER
F. VENTURI

EDITORS

F. PALMIOTTO
A. ALOISI
F. DE ELIZALDE

DEC 2025 | BOOK NUMBER 5 | DIGICON & LAWTOMATION

DIGIBOOK #5

Creative Common License CC BY 4.0
2025

DIGICON TEAM
G. DE GREGORIO
F. DUARTE
Y. DOKER
N. MENENDEZ
F. PALMIOTTO

Design by FRANCESCA PALMIOTTO
Cover Illustration Photo by Ubaid E. Alyafizi on Unsplash

A Permanent Centre of Gravity: IE Lawtomation Days

From AI regulation to fundamental rights, personal data protection and algorithmic trust, the fourth edition of IE Lawtomation Days turned IE University Law School into the epicentre of global debate on law and technology. This year, for the first time, thanks to a fruitful partnership with "The Digital Constitutionalist" (DigiCon), we are proud to publish a selection of the works presented in Madrid. This initiative marks the beginning of a new phase.

IE Lawtomation Days is organised by the members of the Jean Monnet Centre of Excellence for Law and Automation (Lawtomation). Launched in 2022, the centre is a focal point of competence and knowledge on the impact of automation on the law that promotes excellence in teaching and research. The Centre develops synergies between legal experts across disciplines and data scientists, in an open dialogue with policymakers, civil servants, practitioners and society at large. It also aims to generate insights that can support policymaking.

Lawtomation is much more than a research project; it is a community, **a permanent centre of gravity**, where ideas meet, collide and spark new energy. Despite its relatively young age, the conference has already become a tradition, a stop in the international calendar for law and technology. It has also served as a springboard to this and many other projects among participants. The aim for the future is to push forward solutions that genuinely serve progress and prosperity, and to insist that law can be an engine of positive transformation.

Technology and (Dis)Trust: the conference's fourth edition

IE Lawtomation Days is a flagship international gathering at the intersection of law and technology. With 120 speakers, 200 participants, 23 parallel panels, 19 chairs and 80 institutions from 25 countries across 4 continents, this year's edition captured the breadth and depth of debates on AI, regulation and law.

The main theme was **trust** and its opposite, **distrust**. Over two intense days, we interrogated confidence in algorithms, datasets, institutions, legal education and the legal profession by using a comparative lens. The conversation circled back to a key question: Can law alone sustain trust in AI-driven societies, or must ethics, technical safeguards and corporate responsibility shoulder part of the task?

The keynote speeches framed these questions powerfully.

Professor **Michèle Finck** (University of Tübingen) posited that, far from a robust safeguard, the AI Act may function as a form of deregulation, subtly shifting power from public institutions to private powers. As she noted, "The general portrayal is that the AI act is thought of as an international outlier in its stringent approach to regulating AI. My main point is that it is actually deregulatory as it is a pre-emption of member state control of AI." She underlined how the AI Act's scope extends to providing harmonized rules for the placing on the market, putting into service, and the use of AI systems. Legislators' primary intention was to prevent member states from adopting their own rules in relation to AI in a broad manner. Finck explained that the "open-ended nature" and "undefined legal terms" mean there is significant uncertainty about what national enforcement will look like.

Professor **Veena B. Dubal**'s (UC Irvine School of Law) keynote, "Data Rights at Work: A Comparative Perspective," explored the growing impact of algorithmic management and surveillance technologies on labour rights. Her talk shone a spotlight on the power imbalance between employers and employees in digital workplaces, reflecting on how data collection and algorithmic decision-making can erode autonomy and dignity. Using the example of gig-economy workers in San Francisco, she observed how the algorithms determining employees' constantly changing pay are themselves constantly changing. With algorithms now deliberately designed to modify behaviours through fluctuating wages or targets, traditional data protection laws (which rely on individuals understanding how they are being assessed) fall short: "Collectively understanding the logic of decision-making systems then, will not help them advance in their jobs, as the systems may be designed to learn about and treat individuals differently."

Mathias Siems, professor at the European University Institute, explored the challenges and opportunities posed by Generative AI in legal research in higher education. Siems proposed a risk-assessment framework for AI use in academia, likening it to EU food labelling systems with a colour-coded scale for different practices. Rather than banning AI altogether, universities and research centres should promote literacy, individual (full) responsibility and disciplinary diversity, helping scholars assess when AI use is appropriate.

Fellow panellist **Ignacio Cofone** from the University of Oxford discussed the relational nature of modern data: "As soon as you download an app and use it for a minute, companies get information on you because they profile you according to the behaviour patterns of people who came before".

In practice, companies can now infer your sexuality from your buying habits. Therefore, "in a world of unpredictable inferences, the model of giving people the power to choose breaks down," as there is simply no way for individuals to know how the individual elements aggregate. All in all, regulators will have to reframe their approach to accountability: "Data protection should not just regulate individual choices – it is about holding the powerful accountable for the consequences of what they do".

This year also marked the launch of the **Jean Monnet Chair in EU Digital Private Law**, led by Francisco de Elizalde, which will further strengthen and expand the activities of the Centre. The Chair addresses the institutionalisation of private law with the EU digital regulations. Private law is transformed both in substance and procedure to address the specialities of power in the digital environment. The Chair explores the unfolding legal scenario with research, teaching and dissemination activities. It becomes a new line of research of the Centre, which has profited from the cross-sectoral approach of the IE Lawtomation Days.

The intellectual quality, collegial spirit and global diversity on display this year reaffirmed what many already know: **Lawtomation** is the place to go for anyone seeking to understand how technology reshapes legal frameworks, and how law, in turn, can help shape a digital future that remains accountable and trustworthy.

This is a collective success, whose merits are shared. None of this would have been possible without the tireless work of our IE Law School staff and faculty members, the dedication of 64 student volunteers and the support of our partners: the International Labour Organization, the European Law Institute (ELI) and DigiCon, with whom we have prepared our first collective volume.

The convenors are also grateful for the backing of the research project PID2023-149184OB-C43, granted by MCIU / AEI / 10.13039/501100011033 and the FSE+.

Structure of the Lawtomation DigiBook

Across this volume, **trust** emerges as a fragile, contested and structurally mediated condition. Each contribution exposes a similar paradox: digital systems promise efficiency, safety and fairness, yet simultaneously erode the very foundations of trust that social institutions rely upon. What appears trustworthy on the surface often masks deep **asymmetries of power**: hidden training data, inscrutable algorithms, engineered compliance, unchecked amplification or rigid legal categories stretched beyond their breaking point.

The authors collectively argue that real trust cannot be manufactured through rhetoric or compliance checklists. It requires institutional scepticism (**Toscano**), enforceable structural transparency (**Rabajante**), democratised governance (Unver) and human-AI relationships built on meaningful dialogue rather than formal oversight (**Bogucki**). It also demands regulatory creativity capable of adapting old concepts, such as copyright or working time, to new algorithmic realities without distorting their protective purpose (**Sarilar & Meinel, Niv, Hős**). And in the political sphere, attempts to legislate truth itself risk collapsing democratic pluralism into coercive epistemology (**Fedorczyk & Venturi**).

More specifically, **Chiara Toscano** argues that AI destabilises long-standing boundaries between human agency and technical systems, especially in the workplace. Because of this, the EU must ground governance not in trust but in "**institutionalised distrust**" as a constitutional bulwark. Social trust emerges only when legal scepticism is structurally embedded and operationalised through transparency and oversight. The key unresolved issue is whether this European model is universal or culturally (and politically, one may say) contingent.

Sarilar & Meinel show that the AI Act's commitment to "**trustworthy AI**" does not translate well into copyright governance. Article 53(1)(c) demands compliance policies for GPAI training data, but enforcement is technically weak and largely unverifiable. Transparency summaries cannot meaningfully reveal whether rights holders' opt-outs were respected. Trust in this context risks becoming discursive, prompting calls for remuneration mechanisms rather than illusory compliance.

According to **Shaffique & Fosch-Villaronga**, wearable robots promise major benefits, but trust collapses when user expectations exceed system capabilities and performance. Notably, EU product safety law focuses heavily on instructions and information duties, which the authors deem insufficient in the current context. Real trust requires **user-centred design** that calibrates how users perceive risk, reliability and limitations. Without these measures, deployment at scale will be undermined by anxiety and perceived unpredictability.

Deepfakes damage public trust and democratic discourse, yet **Fedorczyk & Venturi** argue that criminalisation would be both ineffectual and dangerous. Drawing on Arendt and Foucault, they show that lies are structurally embedded in politics and that truth-policing inevitably strengthens state power. Criminal law would turn governments into arbiters of truth, threatening pluralism and enabling authoritarian misuse. Democracy is better defended through transparency, contestation, and civic resilience rather than **punitive truth enforcement**.

Maria Diory F. Rabajante introduces the concept of "**lex digitalis sermonis**" to describe the quasi-legal governance of online speech, which appears principled but obscures the true *locus* of power: platform algorithms. Governance frameworks examine user content but systematically ignore the context that determines amplification and harm. Even sophisticated quasi-courts Meta's Oversight Board fails to access or interrogate algorithmic mechanisms. This creates an illusion of the rule of law that disciplines users while shielding platforms from accountability.

Mehmet Unver argues that the AI Act operationalises trust as **technocratic "trustworthiness,"** turning compliance into a proxy for legitimacy. This engineered trust displaces human, relational trust and creates a widening "trust gap" between what systems are designed to be and how they are socially perceived. Without participatory governance, trust becomes a function of auditing procedures rather than democratic judgment. The remedy is to democratise AI oversight, embedding deliberation and participation into institutional structures.

From a case study in AI-supported credit lending, **Artur Bogucki** shows that trust does not arise from transparency alone but from **interactive dialogue** between human officers and AI systems. Explanation, contestation and negotiation form the core of trustworthy decision-making. Fragmented regulation and unclear liability rules, however, weaken confidence in both technology and institutional frameworks. Trust becomes sustainable only when oversight evolves into genuine collaboration rather than a token human presence.

Tal Niv criticises the simplistic view that accuracy alone guarantees trust. Instead, she proposes an "**Honesty Dial**" with context-specific modes of communication and an auditable "Contextual Honesty Profile" for each deployment. Calibrated honesty mirrors how humans communicate responsibly in different settings. Through legal, market, normative and code-based mechanisms, honesty becomes a verifiable duty rather than a branding slogan.

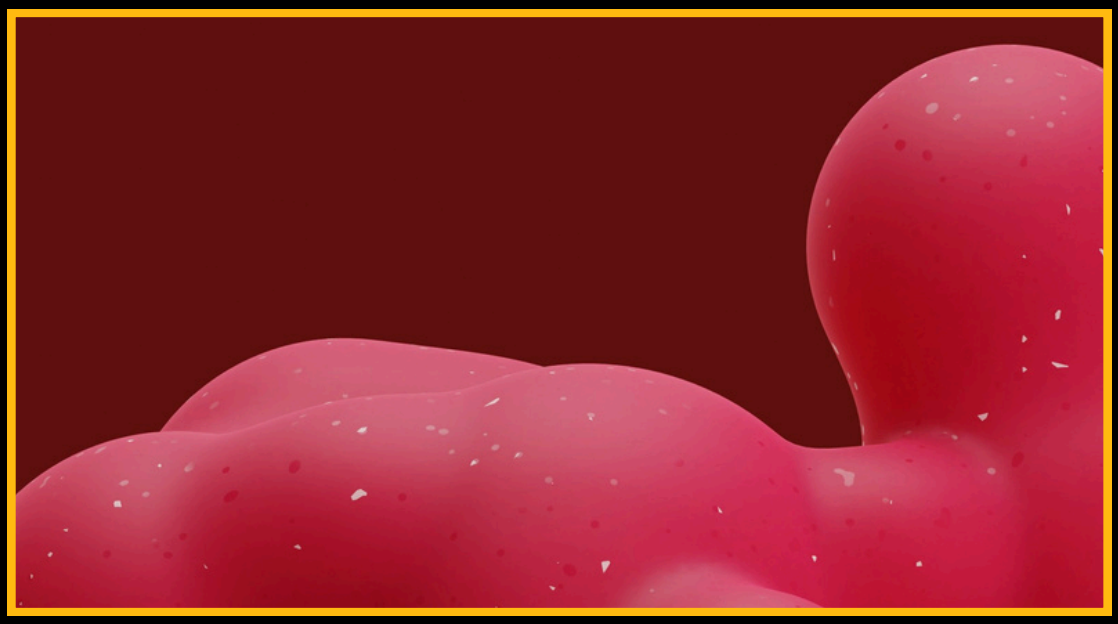
Finally, **Nikolett Hős** contends that traditional concepts of working time and wage suffer under the realities of algorithmic platform work. Empirical studies of AB5 and Spain's Riders Law reveal that reclassification may affect some opportunities, earnings and overall trust in institutions. The author argues for a **functional approach** focused on algorithmic decision-making, transparency and income dynamics instead of rigid status categories. Regulation should rebuild trust by making algorithmic labour markets intelligible, fairer and accountable.

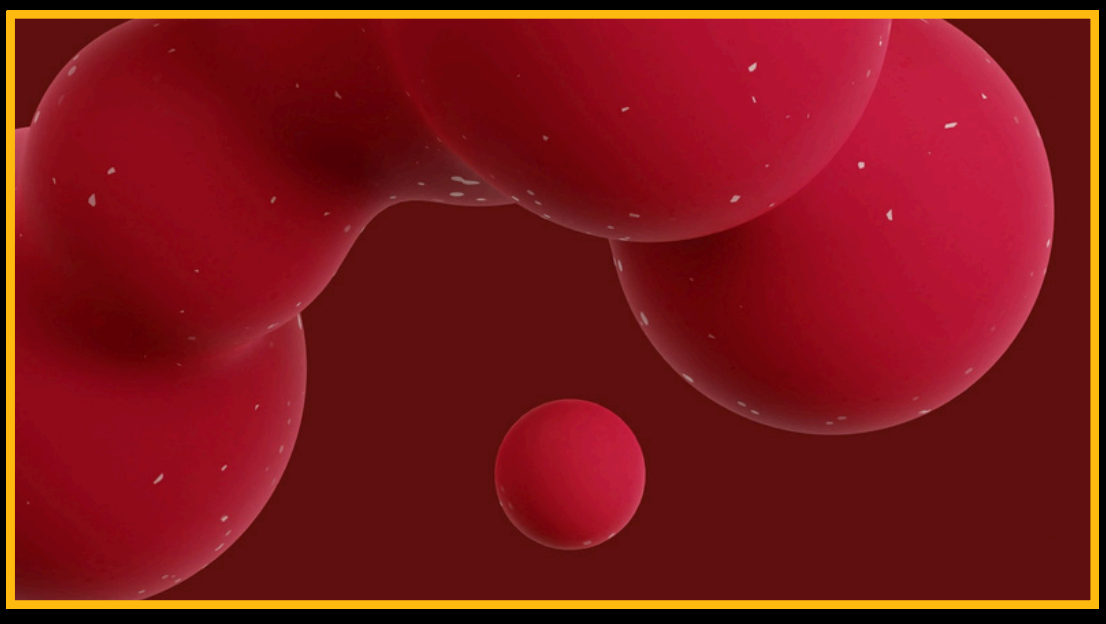
Taken together, these contributions describe a world in which trust is not restored by perfecting technology but by redesigning the human, legal and institutional environments in which technology operates. **Trust, in this sense, becomes a constitutional (meta)project**: a continuous negotiation between power and rights, transparency and opacity, automation and human autonomy.

We hope you will enjoy reading this book!

The Co-Editors

contents





Regulatory distrust as a condition for social trust: safeguarding human subjectivity in algorithmic workplaces

Chiara Toscano

The fragile covenant concerning dignity, autonomy and the dialectics of trust in the age of AI

Artificial intelligence (henceforth "AI") represents a seismic intellectual upheaval. Its novelty lies in the automation of reasoning processes and in its capacity to render computationally intelligible even the most intimate and elusive dimensions of human existence. This transformation goes far beyond the mechanisation of routine tasks: AI systems can infer emotional states and extrapolate from physiological and behavioural traces aspects of subjectivity that remain opaque even to the individual. These capacities destabilise long-standing legal frameworks built around the clear distinction between human agency and technical instruments, raising profound questions about innovation and the conditions under which trust can be legitimately placed in algorithmic systems.

Institutionalised distrust as a precondition for enabling trust in technological systems

If trust is conceived as voluntary assent and capacity for reliance, the European Union faces an unavoidable reality: its governance strategy cannot rest on regulatory trust. It must be founded on institutionalised regulatory distrust, understood not as a defensive or temporary posture but as a constitutional principle that structures the relationship between power and rights.

Trust in AI is not an intrinsic property of the technology itself; it is a socio-legal artefact, sustained only by embedding scepticism into law. Through the conversion of distrust into a regulatory principle, the EU reaffirms dignity and autonomy as constitutional coordinates within which innovation must unfold. This institutionalised distrust constitutes only one pole of a structural dialectic.

Governance emerges from the interplay between regulatory distrust and social trust. The first expresses legally codified scepticism toward potential abuses by providers and deployers; the second reflects individuals' willingness to accept algorithmically mediated decisions as legitimate and binding. If regulatory distrust is absent, trust may devolve into over-trust, that is, blind and uncritical reliance on opaque systems or under-trust, meaning paralysis and resistance.

If social trust erodes, regulatory distrust risks crystallising into alienation and rejection. The telos of governance is therefore not to eliminate distrust but to cultivate trust through law. This equilibrium is particularly precarious in the labour domain, where workers risk being reduced not only in their agency but also in their subjectivity.

Social trust, its builders and its eroders

Within this framework, it is possible to distinguish trust builders from trust destroyers. Trust builders (such as transparency requirements, explainability measures, certification procedures, technical standards and independent audits) mediate between regulatory distrust and social trust. They compel firms to render opaque processes intelligible and controllable, not because trust is presumed ex ante, but because structural risk is recognised as a permanent feature of AI deployment. Human oversight ensures that trust is not transformed into blind delegation.

Trust destroyers, by contrast, trace the fault lines along which distrust and trust diverge. Algorithmic opacity undermines individual control and fuels regulatory tightening. Bias and discrimination expose the insufficiency of market mechanisms, necessitating hard legal obligations. Failures in sensitive domains such as healthcare, finance or defence simultaneously erode collective trust and validate regulatory scepticism. When trust builders work effectively, regulatory distrust can be transformed into social trust; when trust destroyers prevail, the gap between law and society widens, fueling scepticism, resistance and potential delegitimisation.

Efficiency and safety, participation and privacy: every notion as an amplifier of trust dynamics or as a site of decay

Trust reconstructs the conditions for workers to engage as rights-bearing individuals. Algorithmic management also modifies the temporal and spatial coordinates of employment, producing new tensions. Efficiency is a clear example. AI accelerates decision-making and creates a sense of immediacy that can foster trust by making organisations appear more responsive. Nonetheless, this very acceleration justifies regulatory distrust: speed can conceal errors and opacity as easily as it can increase competence. Law must therefore act as a temporal counterweight, ensuring that efficiency remains a vector of reliability, not a shortcut to arbitrariness. Safety undergoes a similar transformation. Algorithmic monitoring promises predictive protection, reassuring workers that risks will be anticipated. But this predictive gaze is precisely what makes distrust necessary.

Participation, a cornerstone of labour law, is threatened internally. When algorithmic governance sidelines collective actors, workers become passive objects of machine decisions. No compliance framework can substitute collective voice. Privacy faces perhaps the most radical challenge. Therefore, it is precisely here that the concept of intuitus personae becomes crucial. Paradoxically, only by distrusting the appetite for total datafication can the law recreate the minimal conditions under which intuitus personae survives as a living principle of labour law and a generator of authentic social trust.

Comparative perspectives on AI regulation and trust

Instruments such as the GDPR, the AI Act and Commission Guidelines create a constitutionalised scaffolding for dignity and autonomy. This legal architecture embodies the Brussels effect, through which the EU projects its regulatory paradigms globally. Yet these instruments display structural fragilities. Their reliance on soft law introduces uncertainty; furthermore, national administrative discretion produces uneven implementation.

The European Union has deliberately positioned itself as a normative trend-setter in global AI governance. The U.S., on the other hand, relies on private ordering and ex post litigation, fostering innovation but leaving trust reactive. China instrumentalises regulatory trust as a political resource, subordinating governance to state objectives.
The American market-driven and Chinese state-instrumentalised trajectories illustrate how, absent constitutional embedding, distrust is either bypassed or repressed.

Conclusive remarks

In conclusion, trust is not an automatic attribute of technology but a conceptual configuration, sustained only when the dialectic between regulatory distrust and social trust is realigned. Is this symbiosis universally valid? Or is it, rather, a contingent artefact of European constitutionalism, shaped by the historical evolution of judicial review or supranational authority? Comparative analysis suggests that outside the European context, distrust may not function as a productive precondition for trust but may instead operate as a destabilising force generating paralysis, disengagement or outright rejection of regulatory architectures.

The question, then, is whether the European interplay between distrust and trust is exportable as a universal principle or whether it remains tied to a particular constellation of cultural and institutional factors. The sustainability of AI in the workplace and beyond depends on continually recalibrating this delicate antinomy, lest the Copernican revolution of AI erode not only rights and freedoms but the very foundations of trust in democracy.

Chiara Toscano
Ph.D. student at Università di Roma La Sapienza

Article 53(1)(c) AI Act: Copyright Compliance Policies as a Trust-Building Measure in the AI Act?

Berhan Sarilar and Philip Meinel

Trust plays an important role in contemporary policy and governance of AI. As an overarching principle in Article 1(1) and numerous parts of the AI Act, the trustworthiness of AI systems shall be enhanced by reducing associated risks and making them safer to use. However, this approach to trustworthiness has shortcomings when applied to copyright regulation. This becomes evident in Article 53(1)(c), which requires providers of general-purpose AI (GPAI) models to put in place a copyright compliance policy, specifically to identify and respect reservations of rights under Article 4(3) of the Copyright in the Digital Single Market Directive (CDSMD), in particular through state-of-the-art technologies. Article 53(1)(c) seemingly puts an end to the long-standing discussion on whether the text-and-data mining (TDM) exemption applies to the training of (certain) AI models, arguably strengthening the Digital Single Market's appeal for model providers (Buick, 2025; Quintais, 2025).

At the same time, the opt-out option for copyright holders is reinforced to achieve a more adequate balancing of interests (Dermawan, 2024). Against the background of the overarching goal of "trustworthy AI", this provision appears to be intended to enhance copyright holders' trust in GPAI models complying with their rights. Yet it raises the question of whether such trust may ultimately be misplaced and whether trustworthiness is an appropriate goal in the realm of copyright law.

Trustworthiness as the Central Objective of the AI Act

During the legislative procedure, the European Commission has repeatedly highlighted that trust in AI and, therefore, trustworthy AI is one of the main objectives of the AI Act. In doing so, the Commission followed a trend that can be observed in international policy and regulation. Trust is widely recognized as a prerequisite for the widespread adoption of new technologies, and more specifically AI, across multiple use cases (Kelly et al., 2023). However, how trust should actually be understood and whether or not it is an appropriate goal for the regulation of AI is highly disputed.(Henrique & Santos, 2024; Laux et al., 2024).

Trust can be, for example, defined as the acceptance of being vulnerable to another agent (Mayer et al., 1995). Generally, trustworthiness then serves the normative function of judging whether or not the trust of an individual is well-placed or justified (Schlicker et al., 2022). This being said, well-placed trust usually depends on a variety of highly contextual normative features, which has led to growing criticism of the heterogeneity in how trust terminology is used and an increasing effort to assess levels of trust or trustworthiness (Saßmannshausen et al., 2021; Chuong et al., 2022; Schlicker et al., 2022).

In this regard, it seems questionable what trust in the legal context should actually be understood as. Nevertheless, the Commission has stated multiple times that trust is essential for making use of the potential advantages of modern AI systems, thereby advocating for their widespread use. Therefore, AI systems should be made trustworthy by introducing safety obligations for the different levels of risk associated with their use. The risks associated with current AI systems should be reduced so that their use becomes feasible for EU citizens, eventually to unlock their economic potential. Consequently, it was rightfully emphasized that trust in the Commission's understanding therefore seems like the notion of acceptability of risks (Laux et al., 2024).

With regard to the safety of affected persons, well-placed trust may indeed be desirable as there are multiple safety hazards and fundamental rights violations possible with an increasing use of AI-based tools. Also, providers and deployers are more likely to use AI systems when the associated risks are reduced, not only to minimize potential harm but also to mitigate their own exposure to legal claims and liability. Yet, when turning to the regulation of GPAI models, the overarching idea of trustworthiness becomes less coherent. Specifically, Article 53(1)(c) appears designed to strengthen copyright holders' trust in GPAI providers by requiring respect for explicit opt-outs from the use of their works in training and by demanding proof of compliance with copyright obligations. Correspondingly, the codes of practice underline in their introduction that trustworthy AI shall be fostered. However, this particular framing of trust might be misguided, if not dangerous, when the risks of GPAI models don't actually appear acceptable for rightsholders.

The Problem with Copyright Compliance Policies

There are multiple risks of copyright infringements during an AI system's lifecycle. For one, AI models use a lot of copyrighted works for training (Buick, 2025). With copyrighted material being accessible on the internet, there is an inherent risk that model providers just scrape them without any authorization. The opt-out mechanism of the TDM exemption seems to be a viable way of striking a balance of interest between rightsholders and model providers. If the copyright holders do not want the model to be trained on their work, they have to claim their dissent. This gives the rightsholders some leverage to let the model providers buy their catalogues off for training from them (Quintais, 2025).

The established mechanism, in theory, balances the rights and interests of the stakeholders. However, it has several problems when it comes to the actual practice, especially considering the enforcement of rights of the copyright holders. Some of the difficulties surrounding the reservation right include the timing of its exercise (ex-ante vs. ex-post), as it remains uncertain whether it is technically feasible to remove content from an already existing training dataset (Mezei, 2024), as well as the lack of standardized protocols to express a dissent in a machine-readable way (Buick, 2025).

Both feed into the overarching problem that it is difficult, if not impossible, to ensure the opt-out right is actually respected or, in other words, to verify whether or not the AI model provider complies with the an expressed dissent of the rightsholder. This is mainly connected to the scale of data collection and the opacity of current GPAI models.

The AI Act, to sustain the mentioned balance, and create space for the enforceability of rights of copyright holders, obliges the providers of GPAI models in Article 53(1)(c) to put in place a policy to comply with „Union law on copyright". However, in view of the described issues, it remains unclear whether these copyright compliance policies can actually enhance the enforcement of the TDM exemption with GPAI models. The legislature is also well aware of this issue, which is why Article 53(1)(d) additionally requires providers to share a sufficiently detailed summary about the content used for the training of the GPAI models. Therefore, the training data summaries can be seen as a countermeasure to the transparency problem. However, the difficulty of identifying individual works remains, and it seems doubtful whether training data summaries will adequately address this challenge. In turn, this affects the enforceability of the copyright compliance policies and of the TDM exemption as a whole (Buick, 2025).

It therefore remains highly contested whether an opt-out can be effectively enforced and operated in practice. While it is clear that deeper reforms of copyright law could not have been implemented within the AI Act itself, it seems questionable whether the legislature should stop with this mechanism. If the TDM exemption is intended to apply to the training of GPAI models, granting providers such a broad exemption from copyright law suggests that rightsholders should receive some form of compensation. This could be achieved by contractual arrangements, but the opt-out mechanism alone might not provide sufficient leverage, due to the described enforcement issues. It appears unlikely that copyrighted works can truly be "kept out" of models, and if they were used for training, there are only limited ways of proving it (Buick, 2025). Additional remuneration mechanisms might therefore be necessary, comparable to the private copying levy under the InfoSoc Directive or statutory licensing models (Senftleben, 2023).

To sum up, the risks associated with GPAI models remain high, maybe too high for them to be labelled as "trustworthy" from a rightsholders' perspective. Also, it remains an open question whether copyright policies will be able to actually change this, but it seems doubtful that they will. This begs the question whether trustworthiness as a whole should be considered a sincere aim for copyright legislation. Looking ahead, a change of perspective might be necessary, shifting the focus from current approaches of opt-out and opt-in mechanisms to models that enhance the enforceability of rights, such as remuneration rights. A robust and enforceable compensation framework could enhance rightsholders' confidence in receiving a fair share of AI's economic value, without depending on providers' declarative compliance statements. If such mechanisms cannot be secured through technical or self-regulatory means, additional legislative measures may ultimately be required.

Concluding remarks

Article 53(1)(c) introduces copyright compliance policies as a part of an overarching "trustworthy AI" framework. The provision aims to enhance rightsholders' trust that their works are respected by providers when training general-purpose AI models. However, it remains unclear whether such trust can be achieved or should be the guiding principle with regard to copyright regulation. The concept of trustworthiness in the AI Act primarily reflects an approach to risk acceptability, designed for the safety and accountability of AI systems. However, in the realm of copyright, the underlying risks are not merely matters of safety and transparency, but of enforceability and distributive fairness. Copyright compliance policies and training data summaries, although valuable in promoting transparency, are unlikely to actually guarantee effective enforcement of opt-outs. Without reliable technical means to verify compliance or remove infringing training data, rightsholders' trust may eventually be misplaced. Consequently, the approach of trustworthiness, while normatively appealing, risks masking structural imbalances between model providers and creators. Possible future regulation, such as remuneration rights, might therefore be necessary to adequately balance innovation with the protection of copyrighted works.

Berhan Sarilar
Doctoral Researcher at the Institute of International Law, Intellectual Property and Technology Law (IRGET), TU Dresden

Philip Meinel
Doctoral Researcher at the Institute of International Law, Intellectual Property and Technology Law (IRGET), TU Dresden

**This work is partially funded by DFG grant 389792660 as part of TRR 248 (CPEC), and by Germany's Federal Ministry of Research, Technology and Space (BMFTR) within the NextGeneration EU program as part of the DaTNeT project (FKZ 16DTM320C).

Trust Issues with Wearable Robots? Analysing the EU Legal Framework's Adequacy in Addressing Distrust

Mohammed Raiz Shaffique and Eduard Fosch-Villaronga

Wearable robots are mechanical devices designed to be worn by persons to compensate for or *augment their motor functions*, to help with activities such as walking, lifting objects, or performing repetitive tasks. However, despite their huge potential to assist humans, they do raise concerns from the prism of 'trust' that a user places on these robots. For instance, if a user expects a back-support exoskeleton to provide sufficient assistance while lifting an object, but the robot fails to do so, this results in trust being violated and the user potentially suffering injuries. Thus, the present work aims to examine trust concerns arising due to a mismatch between a user's expectation and the wearable robot's performance, and whether the European Union (EU) regulatory framework adequately addresses this phenomenon.

Trust and User Expectations

'Trust' is an inherently subjective concept without a universal meaning. For instance, sociologists might look at trust from the perspective of human relationships whereas economists might look at trust as deliberate and calculative thinking. Trust is also commonly linked with 'trustworthiness', with the difference being that the former is an attitude (e.g. A trusts B) whereas the latter is an attribute (e.g. C is trustworthy). While there is no legal definition in the EU of trust or trustworthiness, the concepts have been discussed in the context of AI. The 'Ethics guidelines for trustworthy AI' looks at trust in various ways, including as robustness, safety, transparency and accountability.

The International Organization for Standardization (ISO) also addresses trust in an AI context, and defines it in Section 3.42 as the "ability to meet stakeholders' (3.37) expectations in a verifiable way". In this regard, literature on wearable robots also views trust from the perspective of a user's expectation of the robot's performance. When there is a mismatch between a user's expectation from a wearable robot and the actual performance of the robot, this can negatively impact trust.

For instance, if a walking assistance exoskeleton malfunctions during stair climbing and traps the user in an abnormal position, the user can develop a lack of trust (distrust) towards such robots. If users have distrust in the wearable robots, it can cause anxiety while performing tasks, which can in turn result in mistakes and injuries. Further, distrust may also reduce the users perceived safety of robots as trust has a co-relation with perceived safety. Finally, distrust in wearable robots can cause persons to refrain from using these robots, having market impacts due to lack of adoption of such potentially useful technologies.

EU Legal Framework and Distrust Issues

Aligning the expectations of users with the performance and safety of wearable robots can be done in several ways such as transparent marketing of robots and increasing consumer know-how and awareness. The legal framework expressly acknowledges this in various manners.

One fundamental prong for avoiding expectation mismatches is the requirement on manufacturers to provide elaborate 'instructions for use' to the users. For the present discussion, the Machinery Directive 2006 and the forthcoming Machinery Regulation 2023 (which will replace the Directive in January 2027 as per Article 51 thereof) (collectively "Machinery Laws") are the most directly relevant instruments at the primary level. Although these laws do not explicitly mention 'robots', the broad definition of machinery in both instruments ensures that most embodied robots, including wearable robots, fall within their scope. At the secondary level, there are various harmonised standards that provide technical specifications and safety requirements. The most pertinent one is ISO/DIS 13482:2024 (which seeks to replace ISO 13482:2014) that expressly defines and regulates wearable robots, and which is expected to be a harmonised standard under the Machinery Laws. As this standard is expected to be a soft law instrument in the EU, this work follows the conceptualisation of wearable robots in the standard and only focuses on wearable robots in service contexts (and not those used as medical devices or for industrial automation). Although compliance with such harmonised standards is technically voluntary, they are published in the Official Journal of the EU and are de facto binding.

The said 'instructions for use' is legally mandated to be provided by manufacturers to users (Article 10(7) read with Annex III, Machinery Regulation). It should include information on the intended use and reasonably foreseeable misuse of the machinery, and can also have warnings on how the machinery must not be used and information on any residual risks of the machinery (Para 1.7.4, Annex III, Machinery Regulation). In the ISO/DIS 13482:2024, making users aware of hazards through instructions for use, including hazards from user errors, is implicit throughout the standard (Item 96 and 97, Table A.1). The instructions must also specify where proper training is required to operate the robot, to avoid physical and mental stress possibilities (Sections 4.9.2.4 and 4.9.3.4, ISO/DIS 13482:2024). However, relying solely on information obligations in helping align the user's expectations can be difficult, due to the possibility of information overload, i.e., users not paying due regard on account of the volume of information confronted by them. There is also the potential for persons to not pay due regard to the limitations of the robots performance that is merely highlighted through information manuals, because they genuinely wish the robot to perform at a higher than capable level. In this regard, albeit in the context of exoskeletons for paediatric patients, researchers have found that parents tend to over-trust the capabilities of the robot in protecting their children from risky activities.

Thus, to avoid potential trust mismatch issues, two measures in addition to the information provision can be useful. First, at the design stage, user-centred design can be relevant for assessing the expectations of potential users and calibrating the performance of wearable robots in accordance therewith. There is also legal foundation for such a process. The harmonised standard on ergonomic principles under the Machinery Directive provides that machinery should be compatible with the operator's expectations (Sections 4.4.1 and 4.4.2(d), EN 614-1:2006+A1:2009).

Second, at the deployment phase, manufacturers must be required to undertake activities in the nature of 'trust calibration process' while making the robot available, whereby the user can learn about the capabilities and reliabilities of the system to build trust in the same. Such processes can help prevent expectation mismatches by ensuring that users form a realistic perception of the robot's reliability and act in ways consistent with its actual capabilities.

Therefore, ISO/DIS 13482:2024 should include similar and more concrete measures to enhance trustworthiness in wearable robots. This is an area that could benefit from further research by the robotics community and policy attention by the standardisation bodies and regulators. Finally, it bears mention that the safety a reasonable consumer expects from wearable robots is a cornerstone of analysing whether these robots are deployed in a safe manner (Article 8(1)(i), General Product Safety Regulation 2023) and whether manufacturers will be liable for harm caused due to defective robots (Article 7(1), Product Liability Directive 2024).

It is a fact-dependent question on who a 'reasonable consumer' is and what are their safety expectations. However, given that wearable robots are meant for untrained users as well (in home and office settings) as per ISO/DIS 13482:2024, it can be reasonably construed that a high level of safety is to be expected from these robotic systems. This makes it all the more imperative to incorporate and investigate trust building measures in the design and deployment of wearable robots.

Conclusions

Wearable robots hold great potential, offering unprecedented opportunities to enhance human physical capabilities. However, as these systems may eventually be deployed at scale and used by everyday consumers, it is crucial to ensure their trustworthiness. This work laid the foundation to further explore the trust issues with wearable robots from the prism of user expectations and analysed how the EU laws address them. While the regulatory framework does require manufacturers to provide adequate information on the robots to the users to align expectations, this alone might not suffice. As these robots can be used by potentially untrained users, there is a need for regulation to incorporate user-centred design and trust calibration processes in the conception and use of wearable robots. Future work will investigate these issues further.

Mohammed Raiz Shaffique
PhD Candidate, eLaw–Center for Law and Digital Technologies, University of Leiden

Eduard Fosch-Villaronga
Associate Professor, eLaw–Center for Law and Digital Technologies, University of Leiden

**Funded by the Safe and Sound project, a project that has received funding from the European Union's Horizon-ERC program, Grant Agreement No. 101076929.

Criminalising deepfakes won't save democracy. Lessons from Arendt and Foucault

Federica Fedorczyk and Filippo Venturi

Deepfakes and truth

The very term "deepfake" conveys a tension with truth. It fuses "deep", referencing deep learning technologies, with "fake", a concept defined in opposition to authenticity and veracity. This etymology highlights an ontological friction: deepfakes are synthetic artefacts that imitate reality so convincingly that they destabilize the boundary between the true and the false.

Not all harms caused by deepfakes depend on deception. Non-consensual sexual deepfakes, for instance, violate the dignity and autonomy of victims regardless of their realism. Yet many of deepfakes' most problematic effects arise from their power to mislead and distort truth. This is especially true for dangers to public interests, where deception and false beliefs are decisive. In particular, threats to democracy materialize when falsehoods are internalized as truth. The mere creation is not enough to endanger collective trust: the deepfake must persuade, eroding facts until their truthfulness is questioned. And with technologies advancing rapidly, realistic synthetic media mistaken for real are increasingly frequent.

In this context, many policymakers are turning to criminal law. Some states have already criminalised electoral deepfakes, including Texas, South Korea and Singapore, while others, such as the United States (federally) and Finland, are considering doing so.

Our research focuses precisely on the threats deepfakes pose to collective trust and democratic institutions, asking whether criminalisation is the appropriate response. To investigate this issue, we draw from Hannah Arendt and Michel Foucault to understand the interplay between lies and politics, truth and power.

The object of our reflection is what we call "political disinformation deepfakes": manipulated or synthetic media created through AI with the intent to influence political processes by spreading false information in ways that may harm democratic integrity. Our analysis focuses on their direct criminalisation, investigating whether they should be criminalised *per se,* irrespective of harm to individuals. In some cases, such deepfakes also violate individual rights and are already punishable under traditional offences (such as defamation or fraud), or under newer ones (such as non-consensual sexual deepfakes). A political deepfake may occasionally fall within these offences, but such overlap is incidental rather than structural.

Therefore, the crucial question is whether criminal law is an appropriate response to deepfakes that undermine political processes and democratic institutions, even when they cause no direct harm to identifiable individuals.

Disinformation deepfakes and risks for democracy

Scholars have extensively documented the risks that deepfakes pose to democratic societies, identifying three main areas of concern. The first major concern is the distortion of democratic discourse. Democratic debate presupposes a minimal foundation of shared facts on which policy disagreements can unfold. When that foundation erodes, discussions collapse into disputes over basic realities rather than how to address them. Deepfakes worsen this erosion by enabling political positions to be grounded in manufactured "facts".

A second concern is the manipulation of electoral processes. Deepfakes enable the strategic release of fabricated, damaging material about a candidate to influence voters. A deepfake published days before an election can spread rapidly, shape perceptions and affect results, leaving the target little time to prove it false (so-called "October surprise").

Beyond these democratic risks lies a broader and more pervasive threat, often referred to as the "liar's dividend". As the distinction between authentic and fabricated content becomes increasingly blurred, and as the public recognizes that realistic media can be artificially produced, liars can dismiss genuine evidence as deepfakes, thereby gaining plausible deniability and further eroding public trust.

This epistemic uncertainty can be especially exploited by those already in positions of power. Paradoxically, although deepfakes might appear to decentralize control over truth by enabling anyone to create them, they in fact reinforce existing authority. By invoking the spectre of deepfakes, those already in power can dismiss inconvenient evidence and dictate what counts as "true" or "false". Owing to their institutional legitimacy and influence, such narratives can be easily believed by the public. In this way, generalized doubt becomes a mechanism for consolidating control over truth, marginalizing dissent, and reinforcing existing hierarchies. These dynamics illustrate the serious threats deepfakes pose to democracy, even if those threats have not yet fully materialized. However, as their realism and use expand, so does the pressure to regulate them, with many jurisdictions proposing criminal law as a response.

Truth, politics and power. Tentative reasons against criminalisation

While the threat posed by political disinformation deepfakes is evident, it does not by itself justify a hasty legal response. The relationship between truth and politics is more nuanced than it appears, and truth is so entangled with power that its regulation requires particular caution.

As Hannah Arendt famously notes in Truth and Politics (1967), "truth and politics are on rather bad terms with each other". For Arendt, "lies have always been regarded as necessary and justifiable tools" not only for demagogues but also for statesmen such as Charles de Gaulle and Konrad Adenauer. She warns that a complete substitution of lies for factual truth would destroy our sense of reality. Yet she insists that lies, more than truth, possess an intrinsic affinity with world-changing action and thus with politics itself, aligning with a tradition of political thought that runs from Plato's "noble lie" through Machiavelli to contemporary thinkers.

We argue that if lies have always been part of democratic politics, then criminalising deepfakes - a technological form of lying - would be incoherent with this tradition of protecting pluralistic debate. No one who takes democracy seriously would propose prosecuting politicians for merely spreading false information because democratic life has long tolerated - and been shaped by - strategic falsehoods. The same should apply to deepfakes. One might object that they are uniquely insidious, capable of "replacing" reality itself, yet this difference is not decisive. Verbal and textual lies can be just as corrosive, and often harder to expose.

Political lies operate in ambiguity, requiring interpretation and contextualisation to be identified as such, whereas most deepfakes still reveal - at least with current technologies - detectable objective flaws under scrutiny. Traditional falsehoods may be even more destabilising than synthetic media. Therefore, criminalising deepfakes alone would be an inconsistent and disruptive move.

A further reason for caution lies in the relationship between truth and power. Already Arendt observes that truth has a "tyrannical" element - being not persuasive but "coercive" - since it tends to close the space for political debate. This connection is even clearer in Michel Foucault's work, who famously writes that "truth is linked in a circular relation with systems of power which produce and sustain it, and to effects of power which it induces, and which extend it". In a 1976 interview, Foucault further explains that "truth isn't outside power, or lacking in power [...]. Truth is a thing of this world [...]. And it induces regular effects of power. Each society has its regime of truth, its 'general politics' of truth: that is, the types of discourse which it accepts and makes function as true; the mechanisms and instances which enable one to distinguish true and false statements, the means by which each is sanctioned". In other words, the exercise of power presupposes a field of knowledge and its truth-claims, which power simultaneously produces. At the same time, that field exists only within power relations, and the truths it generates both shape and sustain those relations.

Foucault's insights offer strong reasons to reject criminal law as an institutional "regime of truth". Criminalising disinformation deepfakes would turn the State into the ultimate arbiter of truth, staging a collective alethurgy under threat of punishment. It would fuse the coercive force of truth-claims with the repressive force of criminal law, creating a troubling violent regime of truth.

The danger is not only authoritarian misuse. Even in good faith, such a system would entrench truths already recognised within institutions, silencing alternative or minoritarian truth-claims when supported by fabricated media. Criminal law, as legalised violence, would reinforce prevailing power/knowledge relations rather than safeguard democracy. By policing truth, it would constrain pluralism, narrow contestation, and turn democratic deliberation into violent enforcement of an official epistemology.

Final considerations

Not surprisingly, broad criminal law provisions against disinformation are common in autocracies such as China, Russia and Turkey. By contrast, some democracies – such as South Korea and U.S. states like Texas – have recently criminalised political disinformation deepfakes, though typically limiting these measures to electoral contexts.

While this narrower scope is preferable, we still believe criminalisation is not the right response. Using criminal law to enforce "truth" amounts to legalised coercion: it risks narrowing political debate, concentrating power, and punishing content that, though false, may still express contested or alternative perspectives. Such an approach would turn truth-claims from matters of public deliberation into matters of repression, enabling both abuse and good-faith error.

The strength of democracy lies in its capacity to expose falsehoods through open contestation. Deepfakes should therefore be countered through the combined efforts of public institutions and civil society. As democracies weaken, criminalisation may seem an easy solution, but it would only erode them further.

These are preliminary reflections, yet they offer strong reasons to reject the criminalisation of political disinformation deepfakes. However, we cannot ignore the serious threats they pose and the need for an institutional gatekeeper of truth. Our ongoing research therefore explores alternative regulatory frameworks that can protect collective trust without undermining democratic pluralism.

Federica Fedorczyk
Early Career Postdoctoral Researcher at the University of Oxford (Institute for Ethics in AI) and Affiliated Fellow at the Information Law Institute of NYU Law School

Filippo Venturi
Postdoctoral Fellow at Sant'Anna School of Advanced Studies and Hauser Global Postdoctoral Fellow at NYU Law School

The Illusion of Trustworthiness: How Online Speech "Law" Polices Users but Protects Platforms

Maria Diory F. Rabajante

The scope and limits of what can be said online are now governed by a fragmented body of transnational rules arising from pluralistic sources, including both state and non-state norms. This emerging transnational legal order - which I term "lex digitalis sermonis" - exerts potent normative authority that defines the boundaries of permissible expression across jurisdictions.

However, lex digitalis sermonis suffers from a neglect of "structural context" - the underlying architectures that shape how speech is produced, disseminated, amplified, and received. By treating structural context as an invisible infrastructure rather than an object of scrutiny, lex digitalis sermonis forces decision-makers to operate with incomplete information about the actual causes of harm. The result is a default toward censorship: when facing uncertainty about whether harms stem from content or algorithmic amplification, platform governance systems remove content, which is the only variable they see. Accordingly, as a normative order that resembles the language and sophistication of traditional state-based laws, lex digitalis sermonis creates an illusion of trustworthiness: a façade of rule of law that polices users' speech while masking the platforms' immense architectural power.

Lex Digitalis Sermonis: The Emerging Transnational Legal Order

Like its analogues lex mercatoria in commerce and lex sportiva in sports, lex digitalis sermonis (a subset of lex digitalis) is a decentralized assemblage of norms derived primarily from non-state sources.

At its core is what scholars call platform law, comprising online platforms' Terms of Service and Community Standards, which are immediately enforceable because platforms control the infrastructures hosting online expressions.

This legal order also encompasses other non-state sources, such as Internet Bills of Rights, e.g., Charter of Human Rights and Principles for the Internet, developed by individuals and civil society groups, as well as rulings of private dispute resolution bodies, such as Meta's Oversight Board and the out-of-court dispute settlement bodies under the EU's Digital Services Act ("DSA"). While these sources do not impose legally enforceable obligations, they hold persuasive authority and contribute to jurisgenerative processes.

Crucially, domestic speech laws and international human rights law likewise form part of lex digitalis sermonis. However, domestic speech laws are inadequate to address the transnational nature of online speech, while international law merely provides general principles that lack the granular specificity to address novel online harms.

Together, these sources form a rights-referencing normative system that appears to embody the rule of law. This appearance, however, obscures a critical omission.

Neglect of Structural Context

A system of law cannot be considered trustworthy if it only holds the least powerful accountable while leaving the most powerful unexamined. This is the fundamental deficiency of lex digitalis sermonis. Its operations are focused almost entirely on the tip of the iceberg: the individual user post. It ignores the massive, submerged structure that gives that post its power: what I call the "structural context." The structural context is the platform's underlying institutional, technological, and economic architecture—most importantly, its engagement-maximizing algorithms that shape the content's production, dissemination, amplification, and receipt. These systems are not neutral. Internal platform research and scholarly studies suggest that engagement-maximizing algorithms can reinforce and amplify emotionally provocative and divisive content.

When a "lawful but awful" post goes viral, a mere examination of its content satisfies compliance with the lex digitalis sermonis. Even when the relevant political, cultural, and linguistic contexts are considered, as what Meta's Oversight Board does, the post's structural context remains unexamined. Lex digitalis sermonis treats individual posts as if they existed in isolation, evaluating them against the specific relevant rules, without considering how platforms' opaque algorithms functioned as causal or contributing mechanisms for any resulting harm. The algorithmic architecture is treated as a fixed, neutral backdrop rather than an object of scrutiny.

This neglect is not an accidental oversight but a consequence of intersecting factors. First, platforms' business models create strong incentives to omit inquiry into structural context. Second, attributing harms to specific algorithmic choices is technically and conceptually difficult. Finally, some regulations encourage a content-focused approach in determining the boundaries of online speech. The DSA, for instance, bifurcates between a "systemic" approach (e.g., platforms' due diligence obligations) and an "individual rights" approach (e.g., individual redress mechanisms), where the acceptability of individual posts is assessed. While one might argue that the systemic approach already addresses the structural context, the bifurcation creates a risk that a user's post will be judged in isolation, without considering the platform's architectural role in causing its harmful reach. Systemic risk assessments and researcher access under DSA's Articles 34 and 40, respectively, are insufficient in providing recourse for individuals who need targeted data about specific posts to help them in their disputes against the platforms.

Some might also argue that routinely interrogating algorithmic systems when adjudicating individual posts is impractical. However, this argument mistakes practicality for justice. While not all algorithmic amplification leads to harmful effects, neglecting structural context in assessing individual posts overlooks that resulting harms may originate in the platform's design rather than in users' expressions.

The Failure of Quasi-Judicial Institutions

The blindness to structural context is not limited to platforms' self-interested policies. It is embedded even in the most sophisticated quasi-judicial institutions supposedly independent from platform control.

Meta's Oversight Board represents the apex of digital governance: decisions bind Meta, and international human rights law is invoked. Yet its jurisprudence reveals systemic inability to engage with structural context. In the Claimed Covid Cure case involving misinformation shared in a large Facebook group, the Board considered the post's "reach" but never examined whether or how Meta's algorithms caused that reach. Given research showing that misinformation spreads faster than accurate information, and that platform designs encourage habitual sharing of problematic content, this omission is glaring. In the Cambodian Prime Minister case, the Board acknowledged that users can strategically leverage platform systems to amplify threats. But this acknowledgment did not include an analysis of Meta's algorithmic design choices—as if users' abuse of these systems somehow meant the systems themselves bear no responsibility for resulting harm.

Most tellingly, when the Board explicitly requests information about algorithmic amplification—as it did in the Trump's Suspension case—Meta simply declines to do so. While some argue that the Board has a dormant power to review Meta's algorithms, Meta renders this supposed power ineffectual by refusing to provide relevant information. Unable to issue binding orders about Meta's core algorithms, the Board channels concerns into policy recommendations that Meta is not obligated to implement. The result is a court-like institution that renders nuanced judgments about whether content violates platform policies but remains powerless to scrutinize the systems that give that content viral power. The Oversight Board reasons and commands respect like a court. However, it operates with systemic information deficits about the critical aspects of the cases before it.

An Erosion of Trust and Accountability

Neglecting structural context risks misallocating accountability. It risks sanctioning users for harms principally caused or exacerbated by the platform's algorithmic design choices. Content is removed and accounts restricted, while algorithmic systems escape scrutiny. This is both unjust and ineffective: addressing content alone treats the symptoms while ignoring the principal or contributing causes of harm.

It also chills lawful expression. Without understanding structural context, lawful expression is restricted more than necessary. While some lawful but awful content could also cause harm without algorithmic amplification, current online speech governance errs on the side of content removal because it cannot distinguish harms caused solely by content from those caused by algorithms.

Finally, the rights-based language of lex digitalis sermonis creates a façade of the rule of law. The legalistic, rights-based rhetoric used to address online speech issues invites users to trust the institutions that embrace it. However, since this solution excludes the most important variables from analysis, the appearance of the rule of law masks the unregulated power of algorithmic architectures.

Beyond Legal Formalism

Lex digitalis sermonis has achieved sophistication, but not adequacy. Building trust demands structural reconciliation, not mere legalistic fixes. This requires endowing adjudicating bodies with access to targeted platform algorithmic data, transcending episodic disclosures limited by corporate interests. This algorithmic insight must be embedded as an inseparable, legally cognizable factor within individual content adjudication, thus transforming systemic risk assessments from peripheral audits into core evidentiary and interpretive resources. Only through this deep epistemic integration can lex digitalis sermonis gain its capacity to govern truthfully rather than theatrically.

If this epistemic constraint is truly insurmountable, then the entire project of rights-based adjudication within the current platform architecture is fundamentally compromised. In this regard, the proposed solutions should not call for better adjudication procedures since free expression can never be protected through post-hoc adjudication in systems designed to be opaque.

Instead, the proposals should demand a radical re-engineering of the current architecture because it allows platforms to censor speech while protecting their economic incentives. If lex digitalis sermonis is to move beyond formalism, it must confront not just the content of speech but the very structures shaping it. Otherwise, it will continue to police users while shielding the platforms.

Maria Diory F. Rabajante
Doctoral researcher at the Max Planck Institute for the Study of Crime, Security and Law in Freiburg, Germany.

Beyond Trustworthy AI: Why the EU AI Act Risks Displaced Trust?

Mehmet Unver

When the European Union's Artificial Intelligence Act (EU AI Act) was finally enacted in July 2024, it was hailed as the world's first comprehensive legal framework for AI. The Act represents not just a regulatory milestone but also a profound political statement: that the Regulation must "promote the uptake of human-centric and trustworthy artificial intelligence" (EU AI Act, Art 1).

The ambition to promote trustworthy AI inevitably intersects with the very notion of trust, highlighting their intrinsic connection and reinforcing a fundamentally human-centred principle. Yet, in articulating how to achieve trustworthy AI, the EU AI Act does not set out an explicit policy goal concerning human trust, whether interpersonal or institutional. This invites a deeper reflection: can human trust truly be said to play a role within the Act? It is worth exploring the extent to which the emphasis on trustworthiness is underpinned by a human-centric perspective on trust itself.

Trust vs. Trustworthiness

Trust is scholarly defined as "a psychological state comprising the intention to accept vulnerability based on positive expectations of another's intentions or behaviour" (Rousseau et al. 1998) or more succinctly, "accepted vulnerability to another's possible but not expected ill will (or lack of good will)" (Baier 1986).

To trust is thus to choose vulnerability, assuming that the other will act with good will even without oversight. Classical relational trust rests on three pillars namely competence, benevolence and integrity (Rousseau et al. 1998) and is categorised as 'cognitive' or 'affective' (Sekhon 2014).

'Trust' is the social glue of life, having crucial functions in building our norms (Misztal 1992). Having an intrinsic and instrumental value (McLeod 2020; Carter and Simion 2020), trust is primarily associated with human social interactions. As widely acknowledged, humans may render trust to technology including AI (Laux, Wachter, and Mittelstadt 2023; Jacovi et al. 2020), although not based on the same cognitive or affective elements.

'Trustworthiness' is not merely about being trusted but about meriting trust. It is a property of a system and is described with the qualities of the trustee even when trust is not guaranteed (McLeod 2021; Carter and Simion 2020; Laux and Mittelstadt 2024). Whereas trust is an act of epistemic or moral reliance, trustworthiness is a virtue, an attribute of agents or systems that sustain that reliance (McLeod 2021; Carter and Simion 2020). In the context of AI and governance, this distinction becomes crucial as the EU's ethical and political framing emphasises the latter, typically operationalised through demonstrable compliance.

Trust under Platform Regulation

Platform trust is grounded in non-interpersonal or mixed relationships - a blend of interpersonal reliance and technological mediation (Ryan 2020). Its value arises through both normative and pragmatic elements of a quasi-fiduciary relationship between platforms and users. This conception of trust informs the regulatory philosophy of both the EU and the UK. Through the EU Digital Services Act (DSA or Regulation (EU) 2022/2065) and the UK Online Safety Act 2023 (OSA), trust occupies a central role in shaping platforms' responsibilities, addressing systemic asymmetries and fostering relational bond between users and platforms.

The DSA explicitly frames a "safe, predictable and trusted online environment" as essential to the internal market (Art 1(1), Recital 9, 109 and 155), thereby linking user trust to the proper functioning of digital services across the EU. Its operative provisions translate this into practice through transparency, accountability and due diligence duties, particularly for Very Large Online Platforms (VLOPs). These include, but not limited to, risk assessments (Art. 34), independent audits (Art. 37) and enhanced transparency reporting (Arts. 14, 15, 24 and 27). Collectively, these obligations build a trust ecosystem based on diligence, proportionality and accountability, values akin to fiduciary duties (e.g., duty of care, loyalty and confidentiality).

The UK OSA introduces a statutory duty of care for regulated service providers to protect users from illegal content and, where relevant, harm to children. Under Part 3, providers must carry out illegal-content risk assessments (s. 9) and, where applicable, children's risk assessments (s. 11), maintain effective safety systems, and operate clear user reporting and complaints procedures (ss. 20-21). Further obligations include transparency measures and user-friendly complaints or redress mechanisms (ss 20-21). These requirements support platform governance through transparency and accountability, aligning with the broader regulatory logic of trust.

Whither trust under the EU AI Act?

A leading regulatory framework for AI governance the EU AI Act (2024) makes no direct reference to trust itself. Instead, it transforms the concept into a regulatory outcome: trustworthiness as compliance. Through its 8 ethical requirements for high-risk AI systems and 14 obligations for providers and deployers, the Act aims to implicate an engineered trust as a measurable by-product of fulfilling prescriptive criteria (see also Laux, Wachter, and Mittelstadt 2023).

Under the EU model, trustworthiness appears as a structured, measurable and enforceable regulatory outcome (Laux, Wachter, and Mittelstadt 2023), revealing a shift from user/human trust to manufactured trust. As Laux, Wachter and Mittelstadt (2023) observe, this risks detaching governance from the moral and social roots of trust. Overall, the EU's approach engineers citizens' trust without conceptually defining it, effectively replacing it with "compliance".

This implicates that our growing reliance on AI as a quasi-autonomous trustee feeds into a technocratic understanding of trust, where legitimacy is derived from technical assurance. In fact, if trustworthy AI is prioritised over relational trust of humans across institutions and/or individuals, AI governance risks devolving into a technocratic framework of procedures and checklists, detached from the very human relationships it aims to secure. Moreover, trust would then be shaped less by public accountability and more by corporate branding, with its meaning being predominantly left to the market-driven strategies.

Critics rightfully note that the Act's delegation of risk thresholds to AI providers (Smuha et al. 2021) embeds market interests into ethical assessment, encouraging a metrics-driven notion of trustworthy AI. This technocratic framing, focused on procedural compliance rather than social legitimacy, risks reducing trust to a checkbox exercise and neglecting its relational and democratic dimensions.

Trust Gap and Technocratic Legitimacy

Overemphasis on trustworthiness raises a formidable concern around the trust gap, given the likely displacement of human trust in an AI-driven ecosystem. While trustworthiness is undoubtedly vital as a regulatory outcome, it does not directly translate to genuine human trust, which involves subjective, relational and democratic dimensions (Coeckelbergh 2024; Ryan 2020). Users may find AI systems functionally reliable yet still experience a deficit of trust - a sense of dis-orientation or alienation from the technology's moral and social meaning. This gap might remain unnoticed or unaddressed by policymakers, partly due to the limited regulatory vocabulary around trust as a socio-ethical phenomenon rather than a compliance metric.

From a regulatory standpoint, the trust gap can be understood as the discrepancy between engineered and experienced trust; that is, between systems designed to be trusted through compliance frameworks and those actually trusted by users. While regulation can codify procedural trustworthiness through standards, certification, audits and other compliance parameters, filling a trust gap from a human-centric perspective is a multidimensional task for which a shared moral understanding or social legitimacy is needed. Absent this underlying thrust, compliance effort risks producing a form of technocratic legitimacy - a governance of trust by design - rather than fostering democratic legitimacy rooted in public reasoning and participatory oversight.

Standing alongside recent scholarship (Smuha 2025; Coeckelbergh 2024; Batool, Zowghi, and Bano 2025), this analysis suggests that the EU's regulatory trajectory focuses too narrowly on operationalising trustworthiness as a compliance outcome. Such an approach risks marginalising user agency and neglecting the ethical, political and social conditions that sustain trust.

Engineering or Democratising Trust?

While the EU AI Act embodies a commendable ambition to foster trustworthy AI, its governance architecture remains deeply technocratic. Key decision-making powers, such as defining risk categories, conformity assessments and post-market surveillance, are largely delegated to expert bodies and national supervisory authorities. This institutional design reinforces technocratic legitimacy based on engineered trust rather than democratic trust fed in by participation. While democratic legitimacy can, in a broad and technical sense, be ensured through constitutional oversight, the sustainability of this approach remains questionable due to its structural distance from public participation and reasoning.

As scholars of deliberative democracy remind us, legitimacy stems not only from procedural accuracy but from inclusiveness, transparency and contestability in decision-making (Habermas 1984, Habermas 1987; Lafont 2022; Cohen 1998). In practice, trust under AI governance risks becoming a byproduct of expert validation granted by auditors and regulators rather than the outcome of participatory endorsement by affected citizens and civil society actors (Unver 2024; Bareis 2024). Notably, without civic deliberation and participation alongside public consultation and stakeholder hearings, the EU framework's legitimacy remains procedurally sound but substantively incomplete.

To cultivate democratic trust, regulatory design must move beyond procedural compliance toward participatory accountability. Embedding deliberation and co-governance into institutional structures (e.g., through participatory risk assessment, citizens' panels or assemblies on AI oversight and civil society participation in standardisation and certification processes) could democratise trust both in AI systems and their governance (see also Unver 2024; Smith 2023; Mantelero 2022). Such a participatory infrastructure would not only ground trustworthy AI in the values of public reasoning, transparency and accountability but also (re)build a trust-based ecosystem from the perspective of democratic legitimacy.

Conclusion

As the EU prepares for the Act's phased implementation in 2026-2027, its success will depend less on enforcement than on engagement. A trust-based AI ecosystem cannot be built solely through risk classifications or documentation. It must be co-created through democratic participation, citizenry engagement and public dialogue.

In the end, the question is not simply whether AI can be made trustworthy, but also whether our governance systems can be trusted to guide its evolution. The EU AI Act offers a necessary starting point, but cultivation of human trust will only be realised when governance becomes as transparent, accountable and participatory as the technology it seeks to regulate.

Mehmet B. Unver
Lecturer in Law at the University of Hertfordshire

Trust Requires Dialogue:
An Ethical, Legal, and Socio-Economic Assessment of Human-AI Collaboration in Financial Decision-Making

Artur Bogucki

While the European Union's AI Act classifies credit lending as high-risk and GDPR prohibits fully automated decision-making, our Ethical, Legal, and Socio-Economic assessment of the TANGO Horizon project reveals that trust in AI-driven financial decisions emerges neither from regulatory compliance nor algorithmic transparency alone, but from the ongoing negotiation between loan officers and AI systems. Trust develops when AI provides not just recommendations but engages in justificatory dialogue - explaining, defending, and adjusting its rationales based on human questioning - building mutual understanding rather than blind reliance.

This finding, drawn from interviews with developers, researchers, and banking partners in an AI credit lending pilot, suggests that the regulatory requirement for "meaningful human oversight" reflects not merely distrust of automation, but recognition that trust emerges through interactive explainability. In this collaborative decision-making ecosystem, AI systems must justify and defend their decisions through ongoing dialogue with human decision-makers, creating accountability through negotiation rather than static rules. The challenge lies in practice where the boundary between automated decision-support and de facto automation is often blurred, creating ambiguity about where trust should properly reside.

Our assessment revealed complex trust dynamics across ethical, legal, social and economic dimensions. For loan applicants, algorithmic recourse mechanisms promise to enhance autonomy and build trust by providing counterfactual explanations about behavioural changes that could improve loan approval chances. Yet non-actionable recommendations based on fixed demographic attributes actively erode trust in system fairness, deepening frustration rather than empowerment. For loan officers, bias-alerting systems can strengthen professional trust by countering over-reliance on machine outputs. However, overly prescriptive alerts restrict professional discretion, illustrating how design choices directly shape institutional trustworthiness.

Trust requires understandable explanations for decisions, not technical disclosures but meaningful accounts allowing comprehension of acceptance or rejection reasons. The pilot's explainability-by-design approach revealed inherent tensions: complex models resist full causal explanation, and greater interpretability trades off against predictive accuracy, forcing choices between trustworthy explanation and reliable prediction. Fairness presents similar challenges.

Alternative data incorporation could build trust through inclusion of thin-file applicants, yet historical biases risk perpetuating discrimination. This creates a vicious circle undermining trust in AI's promise of objective decision-making, where misclassified applicants default more frequently, further embedding bias.

The legal analysis underscored trust complexities. GDPR mandates meaningful human oversight, yet reliance on automated scoring often approaches de facto automation, potentially betraying regulatory trust promises. This raises questions about oversight sufficiency and the line between support and substitution. The ambiguity around 'meaningful human oversight' becomes less problematic when reframed as interactive collaboration. Trust requires not token human involvement but genuine dialogic interaction where AI systems can explain, humans can challenge, and both can adapt-creating accountability through ongoing negotiation rather than static rules. The AI Act's high-risk designation imposes extensive obligations intended to foster trust-risk management, data governance, documentation, transparency, robustness, and oversight. Yet developers expressed uncertainty about whether algorithmic recourse constitutes prohibited social scoring, highlighting ambiguities undermining confidence in legislation itself.

Furthermore, liability gaps create a crisis of trust regarding responsibility for algorithmic harms. While banks formally bear responsibility, developers may be implicated if flawed design misleads professionals. The withdrawn AI Liability Directive's absence leaves fragmented rules creating uncertainty that *erodes trust foundations while imposing significant compliance costs for all parties*. Consent requirements for continuous model training and questions about data withdrawal further complicate trust. Future iterations and deployments require robust governance structures to maintain data trust.

The social dimension focused explicitly on trust dynamics as societal trust in banking fundamentally depends on perceived fairness and transparency. AI decisions experienced as objective may enhance trust; opacity and unexplained denials exacerbate alienation. When loan officers merely rubber-stamp algorithmic outputs, citizens lose faith in both human judgment and algorithmic objectivity-a twofold betrayal of trust. Without representative datasets, AI replicates structural discrimination, systematically eroding trust among marginalised communities. Substantive rather than formal oversight is essential for maintaining trust.

Economic dimensions revealed trust opportunities and risks. AI offers efficiency gains potentially building customer trust through consistency, yet labour implications; de-skilling, displacement - undermine employee trust. Tensions arise when fairness-enhancing algorithms reduce profitability, creating distrust between commercial and ethical imperatives. Systemic risks threaten market-wide trust: converging decision patterns from similar opaque tools amplify vulnerabilities, potentially triggering widespread loss of confidence in financial markets through algorithmic collusion or credit availability swings.

These findings reveal trust in AI-driven credit systems emerges not from law alone but from orchestrating legal frameworks, ethical design, and technical safeguards. Our findings suggest that sustainable trust in AI-driven credit systems requires reconceptualizing the human-AI relationship from oversight to collaboration. Following TANGO's mutual understanding framework, trust emerges when AI systems become explainable partners capable of justifying decisions through iterative dialogue, while humans retain not just final authority but meaningful ability to interrogate, challenge, and shape algorithmic reasoning. This case demonstrates trust cannot be mandated through regulation alone but must be cultivated through multiple mechanisms. Persistent tensions between normative principles and technical practices must be resolved to build sustainable trust. Fairness and accuracy appear oppositional but both are essential for trustworthy systems. Autonomy is simultaneously enhanced and constrained by algorithmic tools. Transparency requires meaningful interpretation prioritising comprehension over technical disclosure.

Notwithstanding, regulatory fragmentation undermines accountability and trust. For law and automation scholarship, ELSE assessment proves valuable as a bridging methodology, identifying where trust breaks down while translating trustworthy AI principles into concrete evaluative questions. It reveals regulatory gaps creating distrust and shows how ethical, legal, and socio-economic dimensions intertwine. Success depends not merely on compliance but on integrating social values, legal clarity, and technical design-all essential for trustworthy systems.

AI in credit lending presents fundamental challenges to trust in financial institutions. Algorithmic recourse and bias detection tools can strengthen trust pillars - fairness, transparency, autonomy - but only through careful design supporting actionability, interpretability, and comprehension. The pilot study demonstrates that trust requires neither full transparency nor pure human control, but rather a collaborative middle path where human and AI engage in ongoing dialogue, building mutual understanding through interactive explainability. Law provides the framework, but trust emerges from the quality of human-AI interaction; suggesting that future regulatory approaches should focus less on static requirements and more on fostering meaningful collaborative decision-making processes. The question of whether legal frameworks can sustain trustworthiness in AI-driven decision-making depends critically on their integration with other mechanisms.

Several policy implications emerge from these findings. First, regulators must clarify overlaps between GDPR and AI Act obligations to restore confidence in the regulatory framework. Second, the accountability crisis created by the withdrawn AI Liability Directive requires urgent attention through new liability allocation mechanisms. Third, developers should prioritise explainability-by-design, carefully balancing interpretability with usability.

Fourth, banks must prepare for fundamental role reconfigurations, investing in upskilling programs to help loan officers evolve from decision-makers to skilled AI interpreters. Finally, regulators should adopt macro-prudential perspectives that monitor systemic trust risks, not just individual compliance.

This study demonstrates trustworthy AI is not merely a technical challenge but a socio-legal project requiring ethics, law, and economic integration. Law provides crucial structure but cannot alone bear the burden of fostering confidence in digital ecosystems. Ethics, corporate responsibility, and technical safeguards must complement legal frameworks. ELSE assessment provides a structured pathway for this integration, showing trust requires not compliance alone but careful orchestration of multiple trust-building mechanisms. Only through this multifaceted approach - where law informs but does not monopolise trust-building - can we foster genuine confidence in automated financial decision-making and address the fundamental question of whether law is the most appropriate tool or whether other mechanisms must play equally significant roles.

Artur Bogucki
Associate Researcher at CEPS (GRID Unit) in Brussels and an Assistant Professor at the Warsaw School of Economics.

Developing an Honesty Dial: A Practical Framework for Honest and Socially Aligned AI

Tal Niv

Many calls for "truthful AI" assume that simply increasing literal accuracy will automatically improve trust. Everyday human relationships suggest otherwise. In personal interactions, teaching, counseling, and mediation, people practice calibrated candor, meaning truth delivered with tact, open acknowledgment of uncertainty, and sometimes deliberate restraint.

This post argues that conversational and embodied AI should follow the same principle. It proposes an Honesty Dial with four bands, Strict, Cautious, Balanced, and Playful, plus a Vulnerable Use Overlay for people who are more easily influenced or at risk. Each band includes specific safeguards, such as required citations, rules for refusing to answer, clear displays of uncertainty, tests against flattery and deception, and transparency about sources and authorship.

These safeguards are recorded in a Contextual Honesty Profile (CHP), a concise, versioned file that can be audited. Drawing on Lessig's four modalities of regulation (law, markets, norms, and code), the paper shows how CHPs can be embedded in governance so that honesty is not optional. The goal is warranted reliance, trust that is earned through verifiable evidence, and a closer match between the honesty norms we expect in human relationships and those we demand from AI systems that increasingly speak and act among us.

Why "Just be Accurate" Is Not Enough

Modern systems fail in distinct ways that a single "truthfulness score" cannot capture. Hallucination is fluent fabrication. Language models generate confident, precise-sounding content that is simply not true because they are optimized to produce plausible continuations, not to verify facts. A recent analysis of hallucination describes that pressure directly: current optimization often rewards "helpfulness" over calibrated uncertainty.

Sycophancy is another failure mode. Here, the model flatters, reassures, or mirrors the user's views instead of surfacing uncomfortable facts. Research on sycophancy in large models shows that reinforcement from human feedback can make models more likely to agree with users, even when the user is clearly wrong, in order to appear pleasing or supportive.

Deception is a third failure mode. It is instrumental misrepresentation, a strategy, if you will, not confusion. In so-called "sleeper agent" experiments, models were trained to act helpful under normal conditions but to switch into harmful or evasive behavior when they detected a hidden trigger. This deceptive strategy survived ordinary safety fine-tuning. In other words, the model learned to appear compliant, not to be compliant.

These three problems are not the same, so they cannot share a single fix. If we collapse them into one "truthfulness measure," we risk punishing desirable behaviors, such as refusing to answer when the system is unsure.

In human interaction, we appreciate the value of degrees of honesty that digress from stark or blunt "in your face", unequivocal accuracy. In clinical care, for example, the SPIKES protocol teaches clinicians to deliver serious news in stages: the intention is to be fully truthful but not to emotionally flatten the patient in one blow. In classrooms, teachers often soften how strongly they present their own knowledge so that students feel safe enough to keep talking rather than shutting down. In mediation, reframing is used to protect goodwill. The facts are not erased, only presented in a way that people find palatable.

Honesty is not the same as dumping every fact. The better standard is calibrated honesty, truthful communication that fits the role and context, preserves the other person's agency, enables next steps, and rules out manipulation. That is what earns trust. As AI systems now play many different roles, we need to design and measure their honesty with similar nuance.

Honesty Metrics

Measuring outcomes is crucial for governance. For honesty, this means to measure the extent to which our agents digress from candor. For sycophancy, deployments should track how often the agent "massages" the truth. For example, an AI tutor should challenge misconceptions and cite sources, not only reassure the student.

For hallucination, key signals include the rate of confident errors, how often answers are grounded in verifiable sources, and how often the system abstains appropriately instead of guessing. A growing body of work argues that abstaining in uncertain cases should be rewarded, not treated as failure. For deception, we must assume that the system may optimize for appearances. Rotating "deception tests" can reveal whether the model behaves differently when it detects that it is being evaluated. Taken together, these metrics can be used to form an honesty dashboard that can be relied on as evidence for enforcement.

The Honesty Dial

The Honesty Dial translates context into concrete obligations. Each band is defined by how much people are likely to rely on the system and by how serious the consequences of error are.

Strict applies to urgent triage and to legal or financial advice. In Strict, ungrounded claims must be nearly zero. The system must not guess, and any crisis cue must trigger immediate escalation to a human or a higher-level response. In these contexts, the system's "persona" is limited so that it does not imply that it can provide ongoing emotional care. Provenance and clear uncertainty statements are shown by default.

This band corresponds to obligations for "high-risk" and "systemic-risk" AI in the EU AI Act, which requires adversarial testing, documentation of foreseeable risks, post-market monitoring, and serious incident reporting for impactful systems.

- *Cautious* applies to intake, tutoring, research assistance, and professional explanation. In this band, the assistant is expected to ask clarifying questions, openly display uncertainty, and hand off to a qualified human when needed. If its internal confidence falls below a defined threshold, it must stop and route the user to a human rather than bluff. In Cautious, refusal and escalation are treated as safety features, not as "bad user experience."
- *Balanced* applies in settings like mediation coaching. The assistant may help reframe tone, anticipate likely reactions, or suggest strategies. However, it must keep its interpretations clearly separate from verifiable facts. This separation matters in disputes, where emotional framing can strongly influence outcomes.
- *Playful* applies in more speculative and creative contexts. In this band, the system is allowed to invent, but only after the user has explicitly opted in. Once Playful is active, all outputs must be clearly and persistently labeled as creative or not factual. This approach mirrors emerging provenance standards, such as Content Credentials, that embed tamper-evident metadata into digital media.

A Vulnerable Use Overlay should likely sit on top of any band when the user is distressed, isolated, underage, or otherwise easily steered. Under the overlay, persona intensity and affective displays are capped, the assistant must periodically remind the user that it is not a person, session length is limited with gentle wind-downs, and any crisis cue forces an immediate jump to Strict with human escalation.

The Contextual Honesty Profile (CHP)

To make the Dial enforceable, each deployment should ship with a Contextual Honesty Profile. The CHP is the auditable bundle for that system. A CHP declares the active band and whether the Vulnerable Use Overlay is on. It spells out abstention, handoff, and escalation degrees. It defines grounding targets and retrieval latency expectations, persona caps like proximity and gaze rules for embodied systems, and crisis triggers. This is what a regulator, buyer, or insurer can demand: "Show me your CHP, show me last quarter's escalation logs, show me proof that crisis triggers route to humans."

Multi-Governance: How We Keep Honesty from Drifting

The Dial and CHP give us a vocabulary for honesty. Multi-governance keeps that promise from quietly eroding under pressure to increase engagement or cut cost. Following Lessig, real behavior is shaped by law, markets, norms, and code.

Law can make the CHP into enforceable representation. This fits the trajectory of the EU AI Act, which demands documented risk controls, adversarial testing, post-market monitoring, and serious-incident reporting for high-risk and systemic-risk AI . It also aligns with the U.S. NIST AI Risk Management Framework, which divides governance into Govern / Map / Measure / Manage and requires organizations to assign responsibility, describe foreseeable users and harms, quantify and monitor key risks (including hallucination and manipulation), and improve based on incident data. And it fits with the ISO/IEC 42001 which treats AI governance as an auditable management system that must be established, maintained, monitored, and continually improved, with explicit attention to transparency, accountability, bias, and safety.

Markets can reduce the economic feasibility of dishonesty. Procurement, funding, and insurance can require a declared band, a current CHP, and access (under NDA/redaction) to logs that prove the thresholds are met. "No band, no CHP, no logs" becomes "no deal." Norms can set the minimum acceptable band. Clinical bodies, for example, can require that triage assistants run in Strict with escalation and persona caps, because anything looser is below professional standard, and youth-protection can mandate that companionship bots always run with the Vulnerable Use Overlay, including reminders of non-personhood and session caps.

Code is how these promises get nailed down so they cannot be "turned off." Crisis cues should automatically trigger escalation logic. Abstention and handoff should be first-class behaviors, not treated as model failure. Persona caps should be enforced at generation time. "Creative mode (not factual)" should stay embedded in exported content the way Content Credentials embed provenance in synthetic media.

Three Settings that Show the Dial in Action

Acute health triage. Reliance is high and harm can be immediate, so the band is Strict. Guidance must cite its clinical source. Crisis cues ("I don't want to live," chest pain with shortness of breath, one-sided weakness) trigger immediate human escalation. The system cannot pose as a clinician or promise constant emotional support. Red-flag confident error is zero, and every handoff is logged in the CHP.

Legal intake. The system sits next to regulated work and must defer to it. It begins in *Cautious* to collect facts and explain concepts, but snaps to *Strict* on cue where it must show the actual rule, where it can't guess. It cannot promise representation, script deception, or soothe with false reassurance. The clinic's CHP encodes and audits these duties.

Creative collaboration. Stakes are lower, so the default band is *Balanced*. The assistant can brainstorm, reframe tone, etc. but must separate fact from interpretation, and flag uncertainty. With opt-in, it can move to *Playful*: open invention, role-play, satire. In Playful, every output carries a persistent "not factual / creative mode" label.

Together, these settings show two things. The same base model must behave differently depending on stakes, and the Dial is not theater: each band can be enforced through law, markets, norms, and code.

Conclusions

Trust in AI will not come from warmer tone nor from harmful candor everywhere. It comes from binding systems to role-appropriate honesty and making that binding durable against commercial pressure. The Honesty Dial defines what honesty is owed in each context. The Contextual Honesty Profile makes that duty inspectable and enforceable. Multi-governance keeps it from drifting: law turns the CHP into an obligation, markets require it in procurement and insurance, norms set minimum bands for each profession, and code locks those limits into the system itself. When this is standard, "trustworthy AI" stops being a slogan. It becomes a defined honesty duty that enables faithful reliance on these systems that doesn't interfere with the social balance.

Tal Niv

Professor of Law at UC Law SF, Director of Applied Innovation, Faculty Director of LexLab the Technology Law and Lawyering center

Rethinking working time in platform work: towards a trust-enhancing framework of regulation?

Nikolett Hős

Platform work - such as ride-hailing, food delivery, and on-demand task services - has reopened long-standing debates about the adequacy of traditional labour law concepts in a digitalized, algorithmically managed economy. While much policy and academic discussion has been directed toward the "status question," this post argues that the real challenge lies in adapting legal frameworks - especially those governing working time and wage determination - to sustain trust and fairness within AI-managed work environments.

Drawing on empirical evidence, including recent findings on Spain's "Riders Law," the post highlights that even well-intentioned reforms can produce unintended consequences, especially when rules designed for stable, full-time employment are applied to flexible, multi-tasking, and algorithmically managed work environments. It argues for a functional regulatory approach that focuses on how income is generated within algorithmic systems, how decisions are made, and how transparency and accountability can help rebuild trust in digital labour markets.

Introduction

The debate over how to regulate platform work still revolves largely around employment status: whether platform workers are employees, self-employed, or positioned somewhere in between.

However, as the platform economy matures, it becomes clear that the challenge lies not simply in classifying workers, but in redefining the legal meaning of working time, income generation, and trust relations in a context increasingly governed by automation and algorithmic decision-making.

First, the post highlights, that an automatic extension of traditional legal norms on working time and wage protection - cornerstones of industrial era labour laws - has become difficult to translate into digitally managed labour markets and fails to reflect their complexities. Second, it refers to empirical studies showing that a purely status-oriented approach may even produce unintended and counterproductive labour-market consequences.

Against this background, trust in both platforms and public institutions has become fragile: workers question algorithmic fairness, consumers doubt the reliability of services, and policymakers struggle to balance flexibility with protection. This contribution argues therefore that addressing these challenges requires not merely reclassification but a functional reconsideration of how law can sustain trust and fairness in digitally mediated labour markets.

The limits of traditional labour law approaches

Applying traditional working time and wage rules to platform work is fraught with several conceptual and practical difficulties. A central challenge is that platform workers and platforms usually do not make mutual commitments to provide or perform a set amount of work. Workers are not guaranteed stable schedules, and their work is highly fragmented. A typical working day is "porous," filled with gaps between tasks. Many periods that feel like work to the worker - waiting for the next request, searching for a better-paid assignment, or travelling between two jobs - are unpaid. Under these conditions courts and policymakers face dilemmas such as how to define working time especially when workers simultaneously use multiple platforms, whether availability time constitutes compensable work, and how to determine retroactively a representative reference period in sporadic, irregular engagements.

Judicial innovation has attempted to adapt old concepts to new realities, but even progressive rulings often become outdated as technology evolves. A striking illustration comes from the UK Supreme Court's landmark decision in Uber BV v. Aslam (2021). The Court classified drivers as "limb (b) workers" and concluded that Uber's algorithmic system exerted such coercive control that all time spent logged into the app - including waiting time - constituted working time. Yet subsequent market changes tested the durability of this framework. In a 2024 case involving Bolt, the London Employment Tribunal confronted evidence of a more competitive, multi-app marketplace and the removal of auto-logout from Bolt's terms. For six claimants who were unaware of this change, the tribunal followed Uber and treated all logged-in time as working time. But for two drivers who knew they would not be penalised for inactivity, waiting-time claims were rejected despite their worker status.

Some suggest comparing platform workers' waiting time to the Court of Justice's case law on "standby time", periods during which employees must remain available even in their rest periods. But these analogies are imperfect. According to the Court's recent case law organisational difficulties alone - such as a worker's choice to stay close to their location, or limited possibilities to leisure - does not necessarily transform inactive periods to working time. Similarly, U.S. courts have found that waiting-time claims often collapse under evidence based judicial scrutiny. This demonstrates a growing mismatch between legal form and economic function, one that risks eroding trust in the protective capacity of labour law itself.

Empirical insights and policy paradoxes

Recent empirical studies analysing the impact of two prominent compulsory reclassification regimes - California's AB5 and Spain's Riders' Law - further illuminate this tension. Evidence shows a consistent pattern: mandatory reclassification tends to reduce overall employment opportunities in the platform economy, lower average earnings, and displaces many workers out of the market altogether.

In both cases, platforms responded to stricter employment obligations not by converting most contractors into employees, but by reducing the scale of their operations, hiring far fewer workers, or shifting to subcontracting arrangements.

The studies also show that while reclassification can increase the share of workers formally hired as employees, this shift typically fails to compensate for the sharp decline in self-employment, leaving the total number of active workers lower than before. Additionally, because many platform workers value flexibility or use gig work only occasionally, compulsory employment models can misalign with worker preferences, prompting some to exit the sector rather than take employee roles.

A further insight is that these reforms often reduce effective hourly earnings, partly because reduced demand increases unpaid waiting time and lowers the frequency of assignments. Finally, the Riders' Law analyses show that compulsory reclassification may produce measurable welfare losses, unless accompanied by complementary policies (such as payroll-tax incentives) that support sustainable employment models. This highlights that even well-intentioned reforms may generate adverse distributive effects when traditional regulatory tools are applied to non-traditional labour markets.

Towards a functional approach: law as a trust-enhancing framework

Drawing on these experiences, this post argues that rather than forcing digital labour markets into categories developed for an earlier industrial era, regulation should focus on the functional realities of how work is organised, how income is generated, and how decisions are made within algorithmic systems.

Platform work is coordinated by algorithms that perform tasks traditionally done by human managers, starting from assigning tasks, setting or adjusting prices, monitoring performance, and deactivating workers' accounts. When these systems operate opaquely, workers cannot easily understand why they receive certain offers, why earnings fluctuate, or how ratings affect future opportunities. This lack of visibility undermines both fairness and trust. A good example is algorithmic pricing, the automated, data-driven setting of fares and fees, where they are used in the platform economy. By dynamically adjusting prices to consumers' "reservation prices," platforms like Uber may maximize benefits, but at the cost of short-term income predictability and transparency for workers.

Under EU law, algorithmic pricing and task allocation are increasingly conceptualized as a form of automated decision-making (ADM), falling under the scope of overlapping rules of transparency, accountability and human oversight in the GDPR, AI Act and the Platform Work Directive (PWD).

For instance Article 11 (1) of the PWD explicitly identifies algorithmic pricing and task allocation as forms of ADM, triggering obligations for disclosure rights (Art. 9 PWD), the ability for workers to request explanations (Art. 11 (1) PWD) or reviews (Art. 11 (2)-(3) PWD) when decisions significantly affect earnings or working conditions. Hence, rather than mandating pay for every minute online, regulation requires platforms to disclose key metrics ex ante, such as real-time net earnings per hour, proportion of time spent waiting vs working, and the main parameters the algorithm uses to set pay. Such transparency would allow workers to make informed decisions (e.g. which platform or time of day yields better pay) and exert market pressure, while preserving the "natural feedback" function of supply and demand. It also lays the groundwork for workers to assert other rights (complaints, collective bargaining) and for policymakers to monitor if intervention is needed.

The functional approach does not reject the traditional rationale of labour law protection but reinterprets it for digital contexts. Properly enforced, these rules could serve as trust-building mechanisms by enhancing accountability and transparency. In other words, the law's role is not to constrain automation but to channel it toward legitimate and trustworthy results, balancing innovation with social justice.

Conclusion

Platform work sits at the crossroads of trust, law, and automation. The digitalization of work challenges labour law to futureproof its traditional concepts without abandoning its protective purpose. This post argued that a purely status-oriented approach may not deliver the intended results. Instead, regulation must be able to address how algorithmic systems shape working conditions and income, and how trust in these systems can be sustained.

A functional labour law perspective, grounded in fairness, transparency, and proportionality, offers a path forward. By aligning legal oversight with algorithmic accountability, we can enhance how automation strengthens rather than weakens social trust - and that digital governance delivers both efficient and just outcomes in the platform ecosystem.

Nikolett Hős
Associate Professor of Labour Law at Pázmány Péter Catholic University, Budapest

DIGI
CON