Cambridge Forum on AI: Law and Governance (2025), **1**, e18, 1–9
doi:10.1017/cfl.2025.10

Cambridge Forum

AN EDITORIAL

Individual safeguards in the era of AI: Fair algorithms, fair regulation, fair procedures

Ljupcho Grozdanovski[1] and Jerome De Cooman[2]

[1]FNRS, University of Liège, Liège, Belgium and [2]EU Legal Studies Institute, University of Liege, Liège, Belgium
Corresponding author: Jérôme De Cooman; Email: jerome.decooman@uliege.be

The need for individual safeguards in the face of artificial intelligence (AI) may seem uncontroversial: since AI technologies present risks of violating fundamental rights, conventional legal wisdom would require that, to prevent or mitigate those risks, new safeguards should be created and existing ones, reinforced. Fundamental rights protection has certainly been at the heart of global regulatory and academic concerns regarding AI technologies. In Europe, the original regulatory roadmap was drawn by the European Commission's High Level Expert Group (HLEG). With human dignity as cornerstone of the human-centric approach to AI regulation, the HLEG derived the principles of AI ethics from the fundamental rights enshrined in the EU Treaties, as specific expressions of the EU's foundational values like dignity, freedoms, equality, solidarity and justice. Based on this dignitarian and human-centric premise, the HLEG seminally selected four principles (respect for human autonomy, prevention of harm, fairness and explicability), laying down the normative framework within which the AI Act (AIA)[1] eventually took shape.

Following the HLEG, the concept of "safeguard" appears to be consubstantial, if not synonymous with that of "right." This alignment is understandable: rights establish subjective entitlements that provide individuals with specific benefits while imposing obligations to guarantee those benefits on public and private actors (MacCormick, 2008, 110). It, therefore, seems reasonable to assume that protecting individual rights and freedoms requires the (legal) recognition of subjective entitlements. However, as illustrated by the AIA – an instrument focused on safety and risk-prevention – as well as by the contributions to this special issue, rights and freedoms can also be protected through alternative mechanisms, such as safety standards. With this in mind, when we chose the theme for this issue, we understood the notion of "individual safeguard" to mean both a *regulatory goal* (a framework seeking *to safeguard* individual rights) and an *entitlement* (a right *to claim* a specific type of safeguard or protection).

In this context, the compelling question is: how are individual safeguards and fairness connected? Once again, the debate might appear to be uncontroversial. Any reasonable person would no doubt agree that individual safeguards (associated with, say, fundamental rights protection) contribute to a broader sense of fairness. Be that as it may, the safeguards/fairness interrelationship has remained relatively unexplored in AI scholarship. We can point to the digital constitutionalism strand which, indeed, places its focus on fundamental rights protection, but seldom through the fairness lens (De Gregorio, 2022). Building on these observations, our overview of the articles featured in this issue will be organised around two key ideas: (1) fairness requiring legal mechanisms to enhance individual safeguards and (2) individual safeguards designed to promote fair outcomes.

[1]Regulation No 2024/1689 laying down harmonised rules on artificial intelligence (AI Act – AIA) [2024] OJ L 1689.

1. Fairness requiring law to guarantee individual safeguards

1.1 "Fair" law protecting individual freedom and equality

In the social sciences and humanities, fairness has always been seen as a foundational but elusive concept (Klijnsma, 2015; Stith, 1982). As AI technologies showcased their ability to cause harm (typically, discrimination), they raised issues of fairness that standard regulatory and savant theories could not properly apprehend. Ethicists' reflex was then to (re)turn to the classics in view of laying down the tenets of a "new" ethical framework within which AI legislation could eventually be enforced. The "classics" we allude to are the three topical currents in ethics scholarship: virtue, utilitarian and deontic. Our goal here is not to provide detailed outlines of each current but to highlight and comment on two historically persistent intuitions those currents share.

First, an uncontroversial historic stance has been to view fairness as a specific form of societal "good"[2] considering its purpose to, ultimately, enhance human flourishing (Aristotle, 1956, 3). Depending on the currents considered, fairness-as-a-good was thought to be pursued by persons trained in the virtues (Aristotle, 1956); it was viewed as a utility to be maximised to achieve general happiness (Bentham, 2000) as well as a universal precept for morally "correct" action (Kant, 1997) … Rich, multilayered and intellectually stimulating, these currents have profoundly influenced the prevailing legal understanding of fairness, by giving it a foundational expression through the principle of equality. This is the second of the two historically persistent intuitions. Though the ontology of fairness is difficult to fully grasp, equality has traditionally been seen as its most tangible expression across modern justice, dignitarian (Kazim & Hanna, 2021) and fundamental rights strands (Aizenberg & van den Hoven, 2020).

The justification to equality being best suited to "capture" the essence of fairness can be found in the postulate – cardinal in modern rule-of-law theories – that all individuals are moral equals (Smuha, 2024). Moral equality acts as a defensive idea and is meant to preclude specific groups from perceiving (and treating) others as "lesser" or "inferior." It also implies that, although individuals operate in different socio-economic contexts and have varying talents, affinities and degrees of luck (Anderson, 2015, 21), they should benefit from an *equal claim* to entitlements that support the pursuit of a meaningful life. This creates a demand for decision-makers to devise political, social and legal frameworks where people can have an equal shot at effectively fulfilling their "internal potential" (Gasper, 2014, 107). The Herculean task – taken on by a select few – has been to devise a framework that would succeed in "equalising" socially sustained forms of inequality like unfavourable treatment based on gender, ethnic background, socio-economic status, etc. (Anderson, 2015, 29). In practice, conceptualising a gets-it-right-each-time distributive pattern[3] has, unsurprisingly, proven to be challenging (Dworkin, 1981, 186). We will refrain from commenting on the numerous doctrines in that regard, although Rawls' liberty and difference principles certainly deserve an honorary mention (Rawls, 1999, 24).[4]

Our intention with outlining the two intuitions of fairness scholarship (1. fairness is a social good; 2. law typically synonymises fairness with equality) is to, ultimately, identify what we previously termed as the "fairness requirements," i.e. the constraints that fairness poses on legislation that aims to protect individuals. In that regard, we contend that a legislation's level of fairness is a function of how well it supports individuals' *free* (i.e. uncoerced) and *equal* (i.e. non-preferential) opportunity to *exercise meaningful agency*. "Meaningful" agency is, itself, an elusive concept. For simplicity, let us

[2]The consideration that fairness is conceptually linked to the concept of "good" is arguably the oldest, dating to Aristotle who, in essence, premised his virtue ethics on the fact that reasoned action spontaneously tends towards some good.

[3]Dworkin, e.g., considered that fairness-epitomising distributive patterns should operate allocation in a way that no other allocative alternative would enhance the equality of people's shares of the overall resources.

[4]The liberty principle implies that basic liberty can be restricted for the sake of one or more other basic liberties, but not for a greater net sum of social and economic advantages for society. The difference principle means that, with equal division as a starting point, the more advantaged should not be better off than the less well off.

consider it as a person's ability of self-determination, goal definition and the pursuit of "meaningful" (i.e. meaning-seeking and meaning-giving) action (Cavalcante Siebert et al., 2023).

Against this backdrop, and with AI in focus, we could argue that "safeguarding the individual" as a goal of *fair* regulation is warranted precisely because AI technologies have the potential to undermine the key requirements of fairness (individual freedom and equality), by generating new forms of inequality and limiting individuals' ability to exercise free and meaningful agency. Our developments hereafter offer explanations on the nature and tenor of that threat.

1.2 AI's threats to fairness

1.2.1 The threat of restricting individual freedom

Let us begin with individual freedom. The threat of AI coercing people into making decisions they would presumably not make had they exercised full (free-from-AI) discernment has become somewhat of a commonplace. This is because new technologies, particularly AI, have become consubstantial to how we conceptualise "the individual." Digital humanities and the Science and Technology Scholarship (STS) have long argued that digital artefacts increasingly support self-understanding, shape people's subjectivity and promote various forms of social interaction (Pierosara, 2024, 1786; Rakover, 2024, 2139–2140). They have been called "extended minds" (Jones, 2024, 26), never fixed but always evolving because they, themselves, have a human base and are mediated by human values (Taylor et al., 2024, 2418). Following a techno-optimist view, new technologies like AI can, indeed, help gain a better understanding of the flaws and limitations of human intelligence (Soeffner, 2024, 2214), thus playing an important role in enhancing general human flourishing. Simondon's theory of individuation, for instance, proposes the concept of transduction (Grosz, 2012, 38–42)[5] to capture the key ways in which technologies support the development of a coherent sense of "self" (Bardin, 2015, 4). The implication here is that, given the degree of human/AI entanglement, a clear distinction between the archetypal human Master and the archetypal AI slave is no longer tenable. Both humans and non-human intelligent systems possess the capacity for mutual influence, suggesting a horizontal, rather than a hierarchical dynamic between agents and technologies (Deleuze, 1988). This type of horizontality is *mutatis mutandis* echoed in Latour (2014)'s "actant" concept and Floridi (2012, 6)'s view of technologies' re-ontologising of reality.

While many scholars have rightly highlighted AI's undeniable (both positive and eroding) impact on individual self-determination, legal discourse continues to uphold human autonomy as the prevailing norm (Boddington, 2021). The previously mentioned moral equality principle justifies this: people are moral equals not in their natural talents and abilities, but – again – in their claim to uncoerced and unmediated capacity to reason, decide and act. "The use of AI systems should never lead to people being deceived or unjustifiably impaired in their freedom of choice" – the HLEG (2019, 12) exclaimed.

Moral reasoning is perhaps the realm of human thought where AI systems are most often perceived as intrusive, manipulative and dangerous (Kim, 2023, 1764). The opacity characterising AI decisional and predictive processes exacerbates that danger: individuals may be "coerced" into adopting specific forms of conduct, without being aware that in doing so, they had not, in fact, exercised unconstrained discernment (Brey, 2010, 41). To avoid such scenarios, some theoreticians have advised us to develop technomoral wisdom, i.e. the ability to know how and where to direct our moral attention and conduct (Farina et al., 2024, 1136). Lawyers were more pragmatic and to the point: to deal with AI opacity, regulatory frameworks were to be based on the twin principles of transparency and human control and oversight. In that vein, Article 14 AIA cautions against the "possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (automation bias)"

[5] As Grosz noted, Simondonian individuality is the establishment of a mode of resonance among disparate forces that would otherwise be in a relationship of tension. This "bringing into resonance" occurs through a process Simondon called "transduction" which is, in essence, a specific type of self-structuring. Technology can act as an amplification of that process.

(Art. 14(4)(b) AIA). In essence, human oversight implies the conferring, on a human, the responsibility of *moral arbitration*, i.e. the choice to align, or not, one's own decisions with specific AI output.

The modalities of human oversight have been up for debate in connection with the so-called high-risk AI systems, but also in connection with content moderation. **Vargas Penagos** (*Platforms on the hook? EU and human rights requirements for human involvement in content moderation*) analyses the requirement for human involvement in several EU Secondary Law provisions (General Data Protection Regulation [GDPR], Copyright Directive, Terrorist Content Online Regulation [TERREG], Digital Services Act [DSA], AIA), suggesting a *meaningful* (or qualified) human intervention in the moderation cycle, framed by principles such as diligence, objectivity, non-discrimination, timeliness, non-arbitrariness and proportionality. Specifically, in the field of AI, **Corrêa, Garsia and Elbi** (*Better Together? Human Oversight as Means to Achieve Fairness in the European AI Act Governance*) stress the importance of human oversight within the AIA's framework, to the extent that it contributes to mitigating AI bias, facilitates the allocation of legal responsibility and increases empathy in the contextual understanding of specific AI output, leading to a more adequate assessment of that output's level of fairness.

Of course, not all individuals showcase equal awareness of the threats associated with AI technologies. Some groups are characterised by features that make them more likely to fall prey to AI manipulation. We allude here to the concept of *vulnerability*, generically defined by Coeckelbergh (2013, 44) as a risk of harm associated with a specific interaction between individuals and "the world" (the "world" extending to new technologies like AI). Some scholars have perceived vulnerability as an altered capacity of free (moral) discernment. In his seminal work in the field of personal data processing, Malgieri (2023, 166) distinguished four archetypal "incapabilities" that mark or enhance people's vulnerability: the inability to understand information about data processing; the inability to understand the risks, significance and effects thereof; the inability to give valid consent and the inability to exercise data protection rights adequately. In the AIA, the concept of vulnerability appears as a defining feature of systems considered to present unacceptable risks because of their ability to materially distort the behaviour of persons (Art. 5(1)(a) AIA) or exploit their vulnerabilities due to age, disability and specific social or economic situation (Art. 5(1)(b) AIA). In this special issue, **Taimur** (*Cognitive Freedom and Legal Accountability: Rethinking the EU AI Act's Theoretical Approach to Manipulative AI as Unacceptable Risk*) offers an innovative and structured approach to identifying the origins of manipulation, suggesting Cognitive Impact Assessment (CIA) as a mechanism that could help identify and map out the manipulative tendencies, particularly of the so-called high-risk AI systems. In a similar vein, **Lechevalier and Pottel-Saville** (*Fairness by Design. Combatting Deceptive AI-Driven Interfaces*) make a laudable plea for transparency, by submitting an original and thought-provoking proposal of a Fair patterns model which operationalises the fairness-by-design principle by providing users with specific knowledge and tools necessary to preserve their autonomous decisional capacity.

1.2.2 The threat of violating the principle of equality

Amongst the instances of unfairness AI technologies can trigger in subtle and covert ways, discrimination remains as the most prominent and topical example. In the HLEG's Guidelines, this threat to equality is, above all, showcased by AI's ability to expand and exacerbate historic biases (Hassani, 2021), resulting in the exclusion and marginalisation of specific groups (HLEG, 2019, 18). In many ways, unfair biases are the quintessential feature of the so-called high-risk AI systems. A cursory overview of the eight high-risk sectors listed in Annex III of the AIA confirms this. Indeed, a common thread running through sectors like biometrics, critical infrastructure, education and vocational training, employment and access to essential public services is, indeed, the likelihood of discrimination. Additionally, unfair biases play an important role in identifying the systems that present unacceptable risks such as systems that rely on persons' biometric data to infer their race, political

opinions, trade union membership, religious or philosophical beliefs, sexual orientation, etc. (Art. 5(1)(g) AIA).

Considering the threat of unfair biases, be they explicit or implicit, embedded or machine-learnt, it seems essential for providers to carefully select and "clean" the data used during a system's development and training.[6] The practical translation of fairness remains, however, problematic. This is because of the lack of consensus, first, on how equality should be understood and represented in the field of AI and how it should be, second, translated into specific, mandatory requirements. It is, indeed, difficult to imagine a bullet-proof checklist of criteria that AI providers and deployers would be called to observe when seeking to detect and/or eliminate unfair biases arising in specific AI technologies and those technologies' uses.

It is curious to note that the AIA does not contain any *binding*[7] provisions that explicitly mention fairness. Instead, its binding provisions include references to several conceptual avatars to fairness like non-discrimination, inclusivity and diversity. For instance, fairness-as-inclusivity can be detected in Article 10 ("*Data and Data Governance*") which states that training and validation data should be relevant, and "sufficiently representative" (Art. 10(3) AIA) – take into account the specific geographical, contextual, behavioural or functional factors that frame the use of high-risk AI (Art. 10(4) AIA) – and ensure bias detection and correction (Art. 10(5) AIA). Additionally, Article 15(4) AIA states that high-risk systems that continue to learn after being placed on the market should be developed so that they can eliminate or reduce the "risk of possibly biased outputs influencing input for future operations."

The general meaning of these allusions to fairness does not facilitate or significantly support the compliance with the AIA. Perhaps, the standards developed by the CEN/CENELEC will be more enlightening on the steps to be taken so that fairness (as diversity, inclusivity, non-discrimination … or whatever) can be practically achieved (Agarwal et al., 2023). Some scholars have argued that the AIA does not sufficiently consider inclusivity. For instance, **Karagianni** (*Gender in a Stereo-(Gender)Typical EU AI Law: A Feminist Reading of the AI Act*) argues that the AIA fails to adequately address the structural biases embedded in AI systems which disproportionately impact marginalised groups. She advocates for a feminist re-interpretation of fairness, aiming to ensure that equality – one of the AIA's core normative foundations – encompasses more fully the inclusivity of LGBTQ+ individuals.

In line with this general search for clarity, **Chowdhury and Klautzer** (*Shaping an Adaptive Approach to Address the Ambiguity of Fairness in AI: Theory, Framework and Case Studies*) aim to transcend the conceptual polysemy of fairness by suggesting a method that allows its concrete (and workable) conceptualisation, formalisation, coding and output. To illustrate the solidity of their methodology the authors chose to focus, as a case study, on COMPAS, a recidivism-predicting system. In a similar vein, and reactionary to currents assigning a "fixed" meaning to fairness, **Basu and Das** (*The Fair Game: Auditing and Debiasing AI Algorithms Over Time*) offer a compelling view of fairness as being a dynamic concept. By proposing a fairness assessing mechanism (Fair Game) that puts together an Auditor and a Debiasing algorithm, the authors argue that reinforcement learning can be used to adapt the fairness goals over time, supporting the achievement of those goals prior, during and post AI deployment.

[6]The HLEG (2019, 17) stressed that ensuring that data are free from "socially constructed biases" is crucial for the performance and accuracy of AI systems.

[7]We stress this because the term "fairness" is mentioned only in the AIA's preamble (Rec. 27, 74, 94, 110).

1.2.3. How the law safeguards individuals from AI's threats to fairness

To mitigate the above-mentioned threats to fairness associated with AI, one would expect regulators, particularly the EU's, to ensure individual protection through the recognition of subjective entitlements. The referent here is, of course, the GDPR.[8] This instrument particularised the fundamental right to data protection (Art. 8 CFR) by recognising a series of specific subjective rights (transparency, access, rectification and erasure, data portability, etc.) intended to enhance the guarantees and safeguards associated with the benefit of that right. In the field of AI, the EU legislature chose a different approach: the AIA is a risk-regulating instrument,[9] modelled after standard safety regulation that tends to place the burden of individual protection on market operators, standardisation organisms and surveillance authorities. The AIA's overall design and governance model align with this trend, allowing us to make two observations. First, the protection ("safeguard") of individuals is not primarily achieved through rights but is also the goal of safety standards. Second, the proper of safety standardisation is that it is primarily *preventive*. Heavily relying on compliance, this approach assumes that if predefined standards are observed, individual rights and freedoms (and, with that, general fairness) will not be jeopardised. That assumption is, of course, not absolute. The provider of a biometric identification system can commit to a religious observance of the AIA and the system can still discriminate on the basis of, say, ethnic background. In other words, though safety standards may integrate a rationale of fairness in seeking to protect individuals, they are not systematically conducive to fair outcomes. Hence, the need to reinforce individual safeguards through subjective entitlements; a point further explored in the following section.

2. Individual safeguards designed to promote fair outcomes

Fair outcomes are achieved when the *equal* or *non-preferential enjoyment* of rights remains undiminished throughout the deployment of AI technologies. Based on the HLEG (2019, 12)'s Guidelines, *substantive* fairness is satisfied when unfair biases, discrimination and stigmatisation are eliminated and individuals enjoy an effective equal opportunity to access specific benefits such as education, goods and essential services. To achieve this, and in addition to the above-mentioned safety standards, individual safeguards (understood here as "entitlements") also needed to be enhanced and created. Our terminology ("enhanced *and* created") is intentional. Indeed, the nature and tenor of the risks of fundamental rights violations linked to AI technologies warranted the enhancement of *already existing* individual safeguards associated with rights such as non-discrimination, privacy, data protection, freedom of expression and judicial redress. Additionally, the specificity of AI warranted the establishment of *new* individual entitlements that complement the existing ones, offering another layer of individual protection. The topical example in that regard is the right to an explanation, guaranteed under Article 86 AIA.

It should be stressed, however, that individual entitlements can sometimes (perhaps often) be in tension with the market-oriented entitlements of economic operators, especially in a field like AI, where excellence (implying enhanced innovation and increased AI use) and trust (implying adequate fundamental rights protection) are simultaneously pursued (European Commission, 2020). Reconciling economic growth with the fundamental rights concerns associated with AI has, indeed, been the "double-edged sword" of innovation and a conundrum regulators were called to address (Kasparov & De Cremer, 2022). The HLEG highlighted those tensions, stating that, in any case, the overall benefits of AI should substantially exceed the foreseeable individual risks (HLEG, 2019, 13). From the perspective of fairness, this suggests that the effective application of individual safeguards should lead to a state of *workable* equality, whereby the risk of fundamental rights violations is not fully eliminated but is "framed" so as to not disproportionately alter the benefit from those rights.

[8]Regulation No 2016/679 of the European Parliament and of the Council, of 27 April 2016, on the protection of natural persons with regard to the processing of personal data and on the free movement of such data [2016] OJ L 119, 39.

[9]We allude to Rec. 27 AIA which states that this instrument follows a clearly defined risk-based approach.

That said, the fact that the compliance with the AIA is meant to be trust-inflating and implies a level of acceptance, from end-users, that risks of harm might nevertheless materialise, should not be taken to mean that there are "tolerable" cases of unfairness (Laux et al., 2023).

Several contributors to this special issue highlight timely examples of AI technologies disrupting the equal enjoyment of safeguards associated with specific rights. By drawing attention to these cases, they offer valuable insights into the various forms of unfairness associated with AI, while providing practical solutions on how such harms can be prevented or mitigated. In the field of robotics, **Fosch-Villaronga, Mut-Piña, Shaffique, and Schwed-Shenker** (*Exploring Fairness in Service Robotics*) analyse five fairness dimensions for service robotics, namely, enhancing legal certainty; preventing bias and discrimination; protecting users from exploitation; promoting transparency and accountability and fostering reflexivity among stakeholders. In the field of education (as in "access to education"), **Tieleman** (*Fairness in tension: a socio-technical analysis of an algorithm used to grade students*) offers an interesting case study of the system developed by the UK's Office of Qualifications (OfQual) used to grade students in view of their admission to higher education programmes, showcasing a continuity of the "historic" admissions trends based on the students' socio-economic backgrounds. Going beyond the topical cases of discrimination, **Balendra** (*Meta's moderation and Free Speech: Ongoing Challenges in the Global South*) gives an insightful and compelling analysis on how Meta's AI-driven content moderation undermines the *freedom of expression* in the Global South, resulting in little regard – from Meta as well as public authorities – for hate speech and disinformation campaigns, such as that targeting the Rohingya minority in Myanmar.

Though substantive law can certainly step in to address the various forms of inequality mentioned in the cited contributions, it would presumably remain ineffective without its procedural counterpart. Indeed, in standard legal scholarship, fair outcomes are typically associated with procedures, particularly those that enable judicial redress. According to the HLEG's *Guidelines*, procedural fairness translates to the ability to seek effective redress against decisions made by AI systems or the humans operating them (Laux et al., 2023). To that end – the HLEG (2019) argues – the entity accountable for the decision should be identifiable, and the decision-making process, explainable. In these succinct observations, the HLEG, in fact, highlighted the content of the procedural safeguards which we will, for convenience, classify in two categories: informational (or epistemic) and remedial.

The informational – or epistemic – dimension of the procedural safeguards aims to promote fairness by addressing informational asymmetries between end-users and AI providers or deployers. These safeguards seek to ensure that users can adequately understand a given AI output and the extent to which it influenced a human decision. The *right to an explanation* enshrined in Article 86 AIA "transforms" the principle of transparency into an entitlement, in instances where high-risk systems are deployed. That principle is, in addition, expressed in the right to request access to relevant evidence, enshrined in the revamped Product Liability Directive (PLD)[10] and the recently axed AI Liability Directive (AILD) Proposal.[11] These entitlements are meant to provide individuals with relevant information (explanations and evidence), in cases where clarity is needed on the (possibly unfair?) outcomes of specific AI deployment. Importantly, the duty to provide such clarity applies to all types of AI deployment, including in the performance of public functions. For instance, **Hendrickx** (*Rethinking the judicial duty to state reasons in the age of automation?*) offers an interesting analysis on the rational and argumentative requirements when courts, having used or relied on AI systems, perform their duty to state reasons.

In their remedial dimension, the procedural safeguards considered in this issue give a specific expression, in the field of AI, to the constitutional safeguards enshrined in Article 47 CFR. These safeguards are intended to guarantee individuals' access to remedies and the fair unfolding of proceedings

[10]Directive 2024/2853 on liability for defective products (PLD) [2024] OJ L 2853.

[11]Proposal for a Directive on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive – AILD), COM (2022) 496 final.

through principles like the equality of arms, the contradictory debate and the impartiality of courts. **Pi and Proctor** (*Towards Empowering AI Governance with Redress Mechanisms*) focus on the need to complement substantive AI regulation with effective redress mechanisms designed to empower individuals. Observing an alarming lack of procedural means that would enable end-users to seek effective judicial redress, the authors make important regulatory recommendations by suggesting a combination of *ex-ante* and *ex-post* procedural measures.

3. Closing the circle: Individual safeguards and fairness (of algorithms, regulation and procedures)

Let us return to the question raised at the outset of this editorial and explored in various ways by the contributions to this issue: what is the relationship between individual safeguards and fairness? Our aim is not to offer a definitive answer but to highlight two key dimensions. The first is *instrumental*: fairness functions as a means to achieve individual protection. Whether the notion of a "safeguard" is understood as a regulatory objective (e.g. legislation designed to safeguard) or as a personal entitlement (e.g. an individual's right to a safeguard), fairness appears to have played a central role in shaping the design of the AIA. This is apparent in its provisions that seek to protect individuals' reasoned and uncoerced agency (e.g. provisions on manipulation and vulnerability exploitation), as well as their equal access to, and enjoyment of, various rights and benefits (e.g. the right to an explanation, the benefits in the so-called high-risk sectors).

The second dimension of the safeguard/fairness interrelationship is *consequentialist*: when effectively exercised, substantive and procedural individual safeguards (understood as entitlements) aim to achieve fair outcomes. Indeed, substantive rights (data protection, privacy, non-discrimination, etc.) and procedural rights (access to remedies and courts, effective judicial redress) offer guarantees and protective mechanisms that individuals can rely on to either prevent an unfair outcome from occurring (e.g. the violation of a fundamental right) or rectify it, if it occurs (e.g. compensate a harm suffered).

If these two dimensions of the safeguard/fairness relationship reveal anything, it is that the concepts of "individual protection" and "fairness" are interconnected, each acquiring meaning in relation to the other. The unprecedented challenges posed by AI technologies heighten the need for stronger individual protection, encouraging regulators to embed fairness in the legal and regulatory frameworks designed to offer that protection. Hence, the central theme of this special issue: what do individual safeguards imply in the age of AI? Fair algorithms, fair regulation, fair procedures.

Funding statement. The authors declare none.

Competing interests. The authors declare none.

References

Agarwal, A., Agarwal, H., & Agarwal, N. (2023). Fairness Score and process standardisation: Framework for fairness certification in artificial intelligence systems. *AI & Ethics*, *3*(1), 267.

Aizenberg, E., & van den Hoven, J. (2020). Designing for human rights in AI. *Big Data and Society*, *7*(2), 1.

Anderson, E. (2015). The fundamental disagreement between luck egalitarians and relational egalitarians. In A. Kaufman (Ed.), *Distributive justice and access to advantage. G.A. Cohen's egalitarianism* (pp. 21). Cambridge University Press.

Aristotle (1956). *The Nicomachean ethics*. Harvard University Press.

Bardin, A. (2015). *Epistemology and political philosophy in Gilbert Simondon. Individuation, technics, social systems*. Springer.

Bentham, J. (2000). *An introduction to the principles of morals and legislation*. Batoche Books - Kitchener.

Boddington, P. (2021). AI and moral thinking: How we can live well with machines to enhance our moral agency? *AI & Ethics*, *1*(2), 109.

Brey, P. (2010). Values in technology and disclosive computer ethics. In L. Floridi (Ed.), *The Cambridge handbook of information and computer ethics* (pp. 41). Cambridge University Press.

Cavalcante Siebert L., Lupetti M. L., Aizenberg E., Beckers N., Zgonnikov A., Veluwenkamp H., Abbink D., Giaccardi E., Houben G.-J., Jonker C. M., van den Hoven J., Forster D., & Lagendijk R. L. (2023). Meaningful human control: Actionable properties for AI system development. *AI & Ethics*, *3*(1), 241.

Coeckelbergh, M. (2013). *Human being @ risk. Enhancement, technology, and the evaluation of vulnerability transformations*. Springer.

De Gregorio, G. (2022). *Digital constitutionalism in Europe: Reframing rights and powers in the algorithmic society*. Oxford University Press.

Deleuze, G. (1988). *Spinoza, practical philosophy*. City Lights Books.

Dworkin, R. (1981). What is equality? Part 1: Equality of welfare. *Philosophy and Public Affairs*, *10*(3), 185.

European Commission (2020). White paper on AI, COM (2020) 65 final.

Farina, M., Zhdanov P., Karimov A., & Lavazza A. (2024). AI and society: A virtue ethics approach. *AI and Society*, *39*(3), 1127.

Floridi, L. (2012). Ethics after the information revolution. In L. Floridi (Ed.), *The Cambridge handbook of information and computer ethics* (pp. 3). Cambridge University Press.

Gasper, D. (2014). Logos, pathos and ethos in Martha C. Nusbaum's capabilities approach to human development. In F. Comim, and M. C. Nussbaum (Eds.), *Capabilities, gender, equality. Towards fundamental entitlements* (pp. 96). Cambridge University Press.

Grosz, E. (2012). Identity and individuation: Some feminist reflections. In E. Grosz et al. (Ed.), *Gilbert Simondon. Being and technology* (pp. 37). Cambridge University Press.

Hassani, B. K. (2021). Societal bias reinforcement through machine learning: A credit scoring perspective. *AI & Ethics*, *1*(3), 239.

HLEG (2019). Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed 15 April, 2025).

Jones, M. (2024). Mind extended: Relational, spatial, and performative ontologies. *AI and Society*, *39*(1), 21.

Kant, I. (1997). *Groundwork of the metaphysics of morals*. Cambridge University Press.

Kasparov, G., & De Cremer, D. (2022). The ethics of technology innovation: A double-edged sword? *AI & Ethics*, *2*(3), 533.

Kazim, E., & Hanna, R. (2021). Philosophical foundations for digital ethics and AI ethics: A dignitarian approach. *AI & Ethics*, *1*(4), 405.

Kim, M.-S. (2023). Meta-narratives on otherness: Beyond anthropocentrism and exoticism. *AI and Society*, *38*(4), 1763.

Klijnsma, J. (2015). Contract law as fairness. *Ratio Juris*, *28*(1), 68.

Latour, B. (2014). Agency at the time of Anthropocene. *New Literary History*, *45*(1), 1.

Laux, J., Wachter, S., & Mittelstadt, B. (2023). Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, *18*(1), 3.

MacCormick, N. (2008). *Practical reason in law and morality*. Oxford University Press.

Malgieri, G. (2023). *Vulnerability and data protection law*. Oxford University Press.

Pierosara, S. (2024). Narrative autonomy and artificial storytelling. *AI and Society*, *39*(4), 1785.

Rakover, S. (2024). AI and consciousness. *AI and Society*, *39*(4), 2139.

Rawls, J. (1999). *A theory of justice*. Harvard University Press.

Smuha, N. (2024). *Algorithmic rule by law: How algorithmic regulation in the public sector erodes the rule of law*. Cambridge University Press.

Soeffner, J. (2024). Meaning-thinking-AI. *AI and Society*, *39*(5), 2213.

Stith, R. (1982). A critique of fairness. *Valparaiso University Law Review*, *16*(3), 459.

Taylor, R. R., O'Dell, B., & Murphy, J. W. (2024). Human-centric AI: Philosophical and community-centric considerations. *AI and Society*, *39*(5), 2417.

Cite this article: Grozdanovski L. and De Cooman J. (2025). Individual safeguards in the era of AI: Fair algorithms, fair regulation, fair procedures. *Cambridge Forum on AI: Law and Governance* **1**, e18, 1–9. https://doi.org/10.1017/cfl.2025.10