BRAIN. Broad Research in Artificial Intelligence and Neuroscience
e-ISSN: 2067-3957 | p-ISSN: 2068-0473
Covered in: Web of Science (ESCI); EBSCO; JERIH PLUS (hkdir.no); IndexCopernicus; Google Scholar; SHERPA/RoMEO; ArticleReach Direct; WorldCat; CrossRef; Peeref; Bridge of Knowledge (mostwiedzy.pl); abcdindex.com; Editage; Ingenta Connect Publication; OALib; scite.ai; Scholar9; Scientific and Technical Information Portal; FID Move; ADVANCED SCIENCES INDEX (European Science Evaluation Centre, neredataltics.org); ivySCI; exaly.com; Journal Selector Tool (letpub.com); Citefactor.org; fatcat!; ZDB catalogue; Catalogue SUDOC (abes.fr); OpenAlex; Wikidata; The ISSN Portal; Socolar; KVK-Volltitel (kit.edu)
2026, Volume 17, Issue 2, pages: 144-158
Submitted: March 19th, 2026 | *Accepted for publication:* May21st, 2026

The Role of Artificial Intelligence and Neuroscience in Business Ethics: A Human Resources Perspective

Liviana Andreea Nimineț
Vasile Alecsandri University of Bacău, Romania.
liviana.niminet@ub.ro,
https://orcid.org/0009-0006-7739-7677

Andreea Feraru-Prepeliță
Vasile Alecsandri University of Bacău, Romania.
andreea.feraru@ub.ro,
https://orcid.org/0009-0002-1179-5913

Valer Nimineț
Vasile Alecsandri University of Bacău, Romania including city and Country
valer.niminet@ub.ro,
https://orcid.org/0009-0002-0098-8066

Abstract: *Artificial intelligence (AI) and neuroscience technologies are transforming the workplace at an accelerating pace, offering significant gains in operational efficiency, talent identification, and workforce analytics, while simultaneously generating profound ethical, legal, and social risks. This paper examines their impact on human resources (HR) management from a rigorous interdisciplinary perspective, integrating recent statistical data from Romania and European Union member states with insights from organisational psychology, applied ethics, and machine-learning theory. We analyse concrete applications in recruitment, performance evaluation, diversity management, and employee well-being, providing empirical data that illuminate prevailing trends and disparities. Key findings include: only 13.5% of EU enterprises had formally adopted AI by 2024, with Romania recording a mere 3.1% adoption rate (Eurostat, 2025), contrasted with a striking ground-level reality in which approximately 35% of Romanian office workers already use AI tools regularly (Romania Journal, 2024). This divergence between formal enterprise adoption and informal employee use is conceptualised as a complex governance problem-rather than a mere technological lag-situated at the intersection of organisational psychology, applied ethics, and machine-learning theory, a unified framing that underlies the entire analytical framework of this paper. We systematically identify ethical risks - algorithmic bias in automated hiring (exemplified by Amazon's now-discontinued recruiting engine that penalised women's resumes), covert neuro-surveillance via wearable EEG headsets, and opacity in AI-driven performance appraisal - and ground mitigation strategies in both the EU AI Act (2024) and the emerging Neurotechnology Framework. As a novel contribution, we propose and elaborate a mathematical optimisation model in which organisations maximise a utility function combining productivity gains, bias penalties, and privacy risk costs, subject to formal fairness (equal opportunity) and neurorights constraints. The model, solved via Lagrangian methods, is demonstrated through a detailed case study of a hypothetical Romanian technology firm. This structured, data-driven, ethically grounded approach offers concrete guidance for HR professionals, corporate governance bodies, and policymakers seeking to deploy AI and neurotechnology responsibly and in compliance with EU law.*

Keywords: *artificial intelligence; neuroscience; business ethics; human resources; Romania; European Union; algorithmic bias; employee well-being; neuroethics; EU AI act; equal opportunity; lagrangian optimisation.*

How to cite: Nimineț, L. A., Feraru-Prepeliță, A., & Nimineț, V. (2026). *The role of artificial intelligence and neuroscience in business ethics: A human resources perspective*. BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 17(2), 144-158. https://doi.org/10.70594/brain/17.2/9

1. Introduction

The digital transformation of work has entered a new phase, one defined not merely by automation of routine tasks but by the deployment of sophisticated artificial intelligence systems capable of evaluating human performance, predicting behaviour, and even - through emerging neurotechnologies - monitoring employees' cognitive and emotional states. These developments hold genuine promise: AI-driven HR tools can reduce hiring timescales, minimise subjective bias in some contexts, identify high-potential talent, and personalise employee development pathways. Yet the same tools introduce ethical hazards of corresponding magnitude, from algorithmic discrimination that entrenches historical inequities to invasive neurosurveillance that threatens the most intimate domain of human autonomy - the mind itself.

Understanding these dynamics requires precise empirical grounding. Eurostat (2025) reports that across the European Union, only 13.5% of enterprises employing ten or more people had adopted AI technologies in 2024. The disparity across member states is pronounced: Denmark leads at approximately 27.6%, while Romania records a rate of just 3.1% - placing it among the lowest adopters in the bloc. Enterprise size proves a strong predictor of adoption, with 41.2% of large EU firms (250+ employees) using AI compared with just 11.2% of small firms. Sectoral concentration is equally marked: nearly half of information and communication technology companies (48.7%) report AI use, versus fewer than one in six manufacturing firms. These patterns are summarised in Table 1 and Table 2.

Yet the low formal adoption rate conceals significant individual-level experimentation. A 2025 Romania Journal survey found that 77% of Romanian office workers use AI-powered tools at least occasionally, and 88% expressed a desire for greater automation in their roles. More granular data from the eJobs 2024 survey (n = 4,741 employees) indicate that 34.9% of respondents use AI tools at least weekly in their job tasks, 13.9% do so rarely, and 48.1% never do (Figure 1 and Table 3). This divergence between formal enterprise adoption and informal employee usage reveals a governance gap: workers are experimenting with AI without corporate oversight, ethics frameworks, or adequate training. This governance deficit manifests across three analytically distinct layers, each requiring differentiated policy responses. The first is experimental individual use, in which employees independently explore publicly available AI tools (e.g., ChatGPT, Copilot) for personal productivity, often without any organisational awareness or data-handling safeguards. The second is organisational shadow IT, in which AI adoption has become normalised at the team or departmental level without formal IT or HR oversight, creating uncontrolled data flows that may involve sensitive employee or client information. The third is ungoverned institutionalised use, in which AI-assisted workflows have become embedded in core operational processes yet lack the conformity assessments, logging, and human-oversight mechanisms mandated for high-risk systems under the EU AI Act. Distinguishing these layers is analytically important because they differ in their risk profiles, in the organisational actors responsible for remediation, and in the governance interventions most likely to be effective.

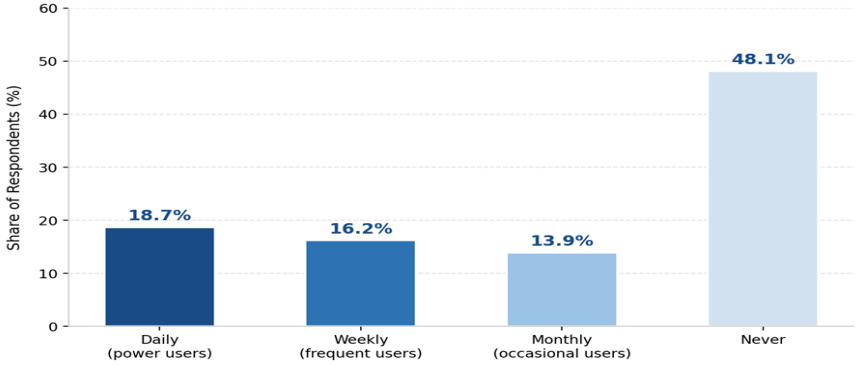

Figure 1. AI tool usage frequency among Romanian office employees
Source: ejobs Romania survey, 2024 (n=4,741)

Against this backdrop, three domains of ethical concern emerge with particular force. First, algorithmic bias in recruitment and evaluation threatens equal opportunity and contravenes both EU non-discrimination directives and the provisions of the newly enacted EU AI Act (2024), which classifies employment-related AI systems as high-risk. Second, the emergence of workplace neurotechnologies - EEG-based cognitive-load sensors, eye-tracking systems, biometric wristbands - raises the spectre of employers accessing the innermost psychological states of their workforce, a risk Muhl and Andorno (2023) term "neurosurveillance." Third, the pervasive opacity of machine-learning models deployed in HR contexts undermines employees' right to understand and contest decisions that affect their careers and livelihoods.

Addressing these challenges requires integrating three distinct disciplinary perspectives into a unified analytical framework. Organisational psychology supplies the behavioural and motivational substrate-explaining why employees adopt AI informally, how algorithmic decisions affect trust and well-being, and what governance structures are likely to gain organisational legitimacy. Applied ethics provides the normative vocabulary for evaluating trade-offs between efficiency and fairness, and for adjudicating among competing conceptions of algorithmic justice.

Machine-learning theory supplies the formal tools for operationalising ethical principles as optimisation constraints, making it possible to move from abstract normative commitments to measurable, auditable system properties. Crucially, AI governance and neurotechnology regulation are not parallel sub-problems but intersecting ones: the same governance deficit that exposes organisations to uncontrolled AI use also creates pathways for ungoverned deployment of neurotechnological monitoring.

The conceptual model proposed in this paper therefore treats both as manifestations of a single underlying challenge-the absence of organisational infrastructure capable of translating legal obligations and ethical principles into operational AI system design.

Table 1. AI Technology Adoption by Country - EU 2024 Source: Eurostat (2025). Use of artificial intelligence in enterprises. European Commission. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Use_of_artificial_intelligence_in_enterprises

Country / Region	AI Adoption Rate (2024)	Notes
EU-27 Average	13.5%	Enterprises ≥10 employees
Denmark (highest in EU)	~27.6%	Nordic leader
Netherlands	~26.1%	High digital maturity
Sweden	~24.3%	Strong ICT sector
Romania	3.1%	Lowest quintile; lag in SME adoption
Bulgaria	~4.8%	Similar to Romania profile
Poland	~7.2%	Growing but behind Western EU
Large EU firms (≥250 employees)	41.2%	Enterprise size strongly predicts adoption
Small EU firms (<50 employees)	11.2%	Significant digital divide

Table 2. AI Technology Adoption by Industry Sector - EU 2024 Source: Eurostat (2025). Use of artificial intelligence in enterprises. European Commission. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Use_of_artificial_intelligence_in_enterprises

Industry Sector	EU AI Adoption Rate (2024)	Risk Level (EU AI Act)
Information & Communication Technology	48.7%	High
Professional & Scientific Services	30.5%	High
Finance & Insurance	~22.0%	High
Manufacturing	~15.3%	Medium
Retail & Trade	~11.8%	Medium
Healthcare & Social Work	~10.2%	High
Agriculture & Primary Industries	~5.0%	Low–Medium

The following sections address these themes systematically. Section 2 examines AI applications in recruitment and performance evaluation, with a focus on bias mechanisms and mitigation strategies. Section 3 analyses diversity and anti-discrimination imperatives in the AI-enabled HR context. Section 4 explores employee well-being and the ethics of neuroscience-based monitoring. Section 5 reviews the applicable policy and governance landscape, including the EU AI Act, GDPR, and emerging neurorights frameworks. Section 6 presents our mathematical optimisation model for ethical HR compliance, including the Lagrangian fairness formulation and simulation approach. Section 7 illustrates the model through a detailed case study. Section 8 draws conclusions and offers recommendations for HR practitioners, organisations, and policymakers.

2. AI in Recruitment and Performance Evaluation: Opportunities and Bias Risks

Artificial intelligence offers transformative potential across the recruitment lifecycle. Natural language processing (NLP) can pre-screen thousands of CVs in seconds, identifying candidates whose profiles best match defined competence frameworks. Machine-learning ranking systems can prioritise applicants based on predicted job performance, while conversational AI platforms conduct preliminary structured interviews, scoring candidates on linguistic markers associated with competencies such as adaptability or communication clarity. In performance management, AI systems can aggregate multi-source data - output metrics, peer feedback, project completion rates, communication patterns - to generate continuous performance profiles that supplement or replace traditional annual appraisals.

The theoretical appeal of these systems is their potential objectivity: by removing human evaluators from initial screening stages, organisations hope to eliminate the well-documented biases (recency effect, affinity bias, halo effects, and demographic stereotyping) that distort human judgment. Chen (2023) reviews the academic literature on AI-enabled recruitment and confirms that, in principle, well-designed algorithmic systems can outperform human screeners on dimensions such as predictive validity and consistency (see also Bigu & Cernea, 2019). However, the same review documents a systematic empirical pattern: unless fairness is explicitly designed into the system, AI recruiters tend to reproduce and amplify the biases embedded in historical training data. The mechanism is straightforward - if past hiring decisions over-represented certain demographic groups in successful outcomes, a model optimising for 'hires like past successful hires' will systematically disadvantage under-represented candidates.

The case of Amazon's internally developed recruiting tool has become emblematic of this problem. Dastin (2018) reports that Amazon engineers discovered, in 2015, that their machine-learning-based CV ranking system had learned to downgrade resumes containing the word 'women's' (as in 'women's chess club') and to penalise graduates of all-female colleges. The root

cause was that the model had been trained on ten years of past hiring data, in which the technology workforce was predominantly male. Amazon abandoned the tool, but the episode illustrates how sophisticated AI, applied without fairness constraints, can institutionalise and scale discriminatory practices that individual human bias would at least apply inconsistently.

Performance management AI introduces analogous risks. When employees' career trajectories are determined by algorithmic scores derived from productivity dashboards, the system may penalise caregiving-related absence patterns, communication styles associated with particular cultural backgrounds, or work-hour patterns that correlate with demographic characteristics (Muhl & Andorno, 2023). The opacity of such models - the inability of employees or even HR managers to understand precisely why a score was assigned - compounds the harm, as it renders contestation effectively impossible.

Mitigation requires what Chen (2023) terms 'fairness-by-design': the explicit incorporation of anti-discrimination principles into the model development and deployment process. This includes: (i) using diverse, representative training data, with active data augmentation where historical records are skewed; (ii) conducting pre-deployment bias audits using established statistical measures such as demographic parity, equalised odds, and predictive parity; (iii) maintaining human oversight at decision points with material consequences; and (iv) implementing transparent appeals mechanisms for affected individuals. The formal mathematical basis for these measures is developed in Section 6.

3. Diversity, Equal Opportunity, and Algorithmic Discrimination

Workplace discrimination predates AI. Eurostat (2022) data indicate that between 5% and 6% of EU workers report experiencing unfair treatment at work based on a personal characteristic, with women consistently reporting higher rates than men. Romania's self-reported rate appears below 1%, though researchers attribute this largely to under-reporting linked to limited awareness of anti-discrimination rights rather than an absence of discriminatory practices. Against this pre-existing landscape of inequality, AI systems that process large datasets and make rapid, opaque decisions create conditions in which discrimination can scale silently and systematically.

From a technical perspective, algorithmic fairness is not a single criterion but a family of criteria that are, in general, mutually incompatible. Demographic parity (also called statistical parity) requires that the AI system select candidates at equal rates across protected groups, regardless of underlying qualification distributions. Equalised odds require that both true positive and false positive rates be equal across groups - penalising models that disproportionately screen out qualified minority candidates or that inappropriately advance unqualified majority candidates. Equal opportunity is the relaxation of equalised odds that constrains only true positive rates: among genuinely qualified candidates, each group must have an equal probability of selection. As Hardt et al. (2016) demonstrate formally, it is mathematically impossible to simultaneously satisfy demographic parity and equalised odds when base qualification rates differ across groups (Hardt et al., 2016). Organisations must therefore make an explicit ethical choice about which fairness criterion best aligns with their legal obligations and values.

The EU AI Act (effective August 2024, with HR provisions phased in through 2026) represents the most comprehensive regulatory framework yet applied to AI in employment. It classifies recruitment, promotion, and performance-evaluation AI systems as 'high-risk', subjecting them to mandatory requirements including: pre-deployment conformity assessments; ongoing human oversight; detailed logging and audit trails; transparency to affected individuals; and explicit bias testing against protected characteristics. For Romanian organisations, compliance must be layered with GDPR requirements (Articles 13, 14, 22), which restrict automated decision-making with legal or similarly significant effects and grant individuals the right to a human review of algorithmic decisions. The Romanian Data Protection Authority (ANSPDCP) has issued guidance emphasising that employee data, including psychometric and performance data processed by AI,

constitutes sensitive processing requiring a lawful basis, a Data Protection Impact Assessment (DPIA), and explicit transparency measures.

Table 4 provides a structured overview of the principal ethical challenges identified in this analysis, the relevant regulatory instruments, and the recommended mitigation strategies for each. This framework is intended as a practical reference for HR compliance officers and ethics committees.

Table 3. AI Tool Usage Among Romanian Office Employees Source: Romania Journal (2024, June 6). 4 in 10 employees frequently use AI tools to ease job tasks.
https://www.romaniajournal.ro/business/companies/4-in-10-employees-frequently-use-ai-tools-to-ease-job-tasks/; Romania Insider (2025). eJobs Romania survey: 63.6% of employees face emotional difficulties at work

Usage Frequency	Share of Employees	Primary Tools	Main Use Case
Daily (regular power users)	18.7%	ChatGPT, Copilot	Content generation, coding
Weekly (frequent users)	16.2%	ChatGPT, search AI	Research, summarization
Monthly (occasional users)	13.9%	Various AI tools	Occasional task support
Never use AI tools	48.1%	N/A	No AI integration reported
Experience emotional difficulties at work	**63.6%**	-	Stress, low motivation (eJobs 2025)

4. Employee Well-Being, Neuroscience Applications, and Neurosurveillance Ethics The well-being of employees represents both a core ethical obligation for organisations and a legitimate business interest: a substantial body of occupational health research links positive well-being to higher productivity, lower absenteeism, and reduced turnover. In the EU, data from the TELUS Health 2024 survey indicate that approximately 40% of workers are at risk of poor mental health, with workload, job insecurity, and lack of autonomy cited as leading contributors.

Table 4. Key Ethical Challenges in AI- and Neuroscience-Enabled HR: Risks, Regulations, and Mitigations
Source: Authors' own synthesis based on: Chen (2023). Ethics and discrimination in AI-enabled recruitment. Humanities and Social Sciences Communications; Muhl and Andorno (2023). Neurosurveillance in the workplace. Frontiers in Human Dynamics; Adomaitis et al. (2022). TechEthos Project D2.4. Zenodo; EU AI Act (Regulation 2024/1689); GDPR (Regulation 2016/679); ANSPDCP guidance (2024).

Ethical Challenge	Description & Example	Relevant Regulation	Mitigation Strategy
Algorithmic Bias in Hiring	AI penalises protected groups (e.g., Amazon gender bias). Biased training data perpetuates inequalities.	EU AI Act (high-risk); GDPR Art. 22	Fairness constraints, diverse training data, regular audits
Neural Surveillance / Neuromonitoring	EEG wearables track cognitive load; brain data used for promotions or terminations.	Proposed EU Neurorights Framework; GDPR special categories	Explicit consent, data minimisation, anonymisation
Lack of Transparency / Explainability	Black-box models make decisions employees cannot understand or contest.	EU AI Act Art. 13 (transparency); GDPR Art. 22 (right to explanation)	Interpretable models, appeals mechanisms, disclosure policies
Discriminatory Performance Monitoring	AI-driven KPI tracking may amplify existing group disparities in evaluations.	EU Equal Treatment Directives; Romanian Anti-Discrimination Law	Equalised odds constraints, human review, impact assessments
Data Privacy Violations	Processing sensitive employee profiles without lawful basis or adequate safeguards.	GDPR; Romanian Data Protection Authority (ANSPDCP) guidance	Data protection impact assessments (DPIAs), encryption, consent

In Romania, the picture is especially sobering: eJobs Romania's 2025 survey found that 63.6% of employees are currently experiencing emotional difficulties at work, including persistent stress, motivational deficits, and emotional exhaustion. These figures underscore that workforce well-being cannot be treated as peripheral; it must be central to any ethical framework for AI and technology deployment in the workplace.

AI can, in principle, contribute positively to employee well-being. Intelligent scheduling systems can optimise workloads to prevent systematic overload. AI chatbots can provide employees with immediate, accessible, and stigma-free first-line support for mental health inquiries, pointing users towards professional resources or employee assistance programmes. Predictive analytics can flag early signals of burnout - declining response rates, shortened working hours, increased error rates - enabling proactive managerial intervention. In safety-critical industries, AI systems can monitor fatigue levels and alert operations managers before dangerous cognitive impairment occurs.

Neurotechnology represents the frontier of this domain. Consumer-grade EEG headsets capable of measuring electrical brain activity - and inferring from it cognitive states such as focused attention, mental fatigue, and emotional arousal - are now commercially available at price points accessible to corporate buyers. Ocular tracking can measure pupil dilation and gaze patterns as proxies for engagement and cognitive load. Heart-rate variability wristbands provide continuous indices of stress reactivity. Proponents argue that real-time neurofeedback could allow workers to optimise their own cognitive states, and could enable managers to deploy human resources more intelligently - redirecting cognitively overloaded engineers, for example, or identifying optimal conditions for creative work.

Muhl and Andorno (2023) provide a rigorous ethical analysis of these possibilities, introducing the concept of 'neurosurveillance' to describe employer monitoring of employees' mental and neurological states without full, free, and informed consent. They identify several distinct harms: first, 'neuro-discrimination' - the use of brain-derived data to make employment decisions in ways that may reflect not individual competence but demographic group differences in neurological profiles or stress responses; second, the chilling effect on cognitive autonomy, as employees aware of constant neurological monitoring may alter their thinking patterns, reducing the creative freedom that employers ostensibly value; and third, the unprecedented intimacy of the intrusion, which reaches into the domain that has hitherto been inviolably private - the subjective experience of thought.

Crucially, Muhl and Andorno's survey data reveal that while workers recognise potential personal benefits from self-directed neurofeedback, they overwhelmingly and consistently oppose employer-directed monitoring of their neural states. This finding holds even when potential safety benefits are foregrounded in the survey framing, suggesting that the opposition is rooted in fundamental values regarding cognitive liberty and the limits of employer authority over the person. The TechEthos Project (Adomaitis et al., 2022) has proposed that brain-derived data should be classified as 'special category' data under GDPR - warranting the same stringent protections currently applied to health data, biometric identifiers, and data concerning racial or ethnic origin. Several national governments and the Council of Europe's Oviedo Convention framework are actively considering analogous neurorights provisions.

For HR practitioners and corporate ethics bodies, the practical implication is clear: any deployment of neurotechnology in the workplace must be grounded in explicit, fully informed consent - consent that is genuinely voluntary, which in an employment context requires structural safeguards to ensure that workers do not experience implicit coercion. Data collected via neurotechnological means should be used exclusively for the purpose for which consent was given, subject to robust technical access controls, and never repurposed for personnel decisions (performance appraisals, promotions, redundancy selections) without fresh specific consent. DPIAs are mandatory under GDPR for any processing likely to result in high risks to individuals' rights, and neuroscience-based monitoring clearly meets this threshold.

5. Policy, Governance, and the Regulatory Landscape

The deployment of AI and neurotechnologies in the workplace does not occur in a normative vacuum. A layered regulatory framework - EU-level, national, and sector-specific - shapes the legal parameters within which organisations must operate, and increasingly prescribes positive obligations rather than merely prohibiting specific harms. Understanding this landscape is a prerequisite for responsible AI governance in HR.

At the apex of the EU framework sits the AI Act (Regulation 2024/1689), which entered into force in August 2024 and applies to providers and deployers of AI systems across all member states. For HR applications, the most consequential provisions are in Article 6 and Annex III, which identify as 'high-risk' any AI system used for: (i) recruitment and selection of natural persons, including CV screening, application filtering, and interview assessment; (ii) decisions affecting terms of employment, including promotion, task allocation, performance monitoring, and termination; and (iii) evaluation of creditworthiness or reliability of persons in employment contexts. High-risk systems are subject to obligations including conformity assessment, registration in an EU-wide database, ongoing human oversight, maintenance of detailed technical documentation and event logs, and transparency to individuals subject to AI-based decisions. Penalties for non-compliance range up to €30 million or 6% of global annual turnover, whichever is higher.

The General Data Protection Regulation (GDPR, Regulation 2016/679) provides a complementary layer of protection, particularly through its restrictions on automated decision-making (Article 22), data minimisation requirements, the accountability principle (Article 5(2)), and DPIA obligations under Article 35. In Romania, these provisions are implemented nationally through Law No. 190/2018 and enforced by the ANSPDCP. In 2024, the ANSPDCP issued updated guidance clarifying that profiling of employees using AI-generated psychometric scores or behavioural predictions is subject to the most stringent GDPR standards, requiring explicit legal basis, full transparency, and the provision of a meaningful human review mechanism.

At the organisational level, best practice increasingly involves the establishment of dedicated AI ethics governance structures. These may include AI ethics committees or boards with cross-functional membership (legal, HR, IT, employee representatives, and external ethics advisers); mandatory algorithmic impact assessments prior to deployment of any high-risk HR AI system; standing audit programmes to detect and remediate bias or fairness drift in deployed models; transparent employee communication protocols disclosing the use of AI in hiring and evaluation processes; and formal appeals and redress procedures enabling individuals to seek human review of AI-informed decisions. Deloitte's 2024 European HR Technology Survey[1] found that 70% of European employers felt inadequately prepared to comply with AI Act requirements, indicating a substantial gap between regulatory expectations and current organisational capabilities that must be urgently addressed.

6. Mathematical Modelling of Ethical Compliance in AI-Enabled HR

6.1. The Ethical Utility Function

To move beyond qualitative ethical prescriptions and enable systematic, data-driven governance, we propose a formal optimisation framework for ethical HR compliance. Let x denote the configuration parameters of an AI-enabled HR system - for example, the decision thresholds, model weights, or data-processing policies that define how the system operates. We define an ethical utility function U(x) that combines organisational objectives with explicit ethical cost terms (1):

$$U(x) = \alpha \cdot \text{Performance}(x) - \beta \cdot \text{Bias}(x) - \gamma \cdot \text{PrivacyRisk}(x) \quad (1)$$

[1] https://www.deloitte.com/us/en/services/consulting/blogs/human-capital/2024-hr-technology-trends.html

Here, Performance(x) represents the business value generated by the system - for example, the efficiency of the hiring process, the predictive validity of performance assessments, or the productivity gains from AI-assisted task allocation. Bias(x) is a quantitative measure of algorithmic unfairness, such as the absolute difference in true positive rates between protected groups (the equal opportunity gap), or a weighted average of disparities across multiple protected characteristics. PrivacyRisk(x) quantifies the expected harm to employee privacy, scaling with the sensitivity of the data processed (with neural data receiving the highest weight) and the scope of surveillance. The weights α, β, $\gamma > 0$ are organisational parameters reflecting the relative priority assigned to productivity versus ethical obligations. They can be calibrated empirically - for instance, β can be set to reflect the expected cost of regulatory penalties and reputational damage from bias-related complaints, while γ can be tied to GDPR fine exposure and the magnitude of potential neurorights violations.

The organisation's objective is to maximise U(x) subject to a set of hard constraints that encode absolute ethical and legal requirements - boundaries that may not be crossed regardless of the productivity trade-off they entail.

To make this framework accessible to HR and policy practitioners unfamiliar with formal optimisation, an intuitive decision-support interpretation is helpful. Consider an organisation evaluating two configurations of its AI recruitment system: Configuration A maximises raw screening accuracy (high Performance, but a measurable TPR gap across gender groups); Configuration B imposes a fairness constraint that reduces the TPR gap to within legal tolerance (lower raw accuracy, but significantly reduced Bias penalty). The utility function U(x) enables the organisation to compute a single composite score for each option that weights these dimensions according to its own calibrated values of α, β, and γ. A compliance-oriented organisation with high β will prefer Configuration B; a less constrained one with high α may prefer A-but the model makes explicit that this choice carries a quantified ethical cost. In this way, abstract normative trade-offs become traceable, auditable, and communicable to ethics committees and regulatory bodies.

6.2. Equal Opportunity Constraint and Lagrangian Formulation

Following Hardt et al. (2016), we formalise the equal opportunity criterion as follows. Let $\hat{Y} \in \{0,1\}$ denote the AI system's binary output (e.g., 'invite to interview' or 'do not invite'), $S \in \{0,1\}$ a binary sensitive attribute (e.g., gender), and $Y \in \{0,1\}$ the ground truth qualification indicator. The equal opportunity constraint requires (2):

$$|P(\hat{Y}=1 \mid S=0, Y=1) - P(\hat{Y}=1 \mid S=1, Y=1)| \leq \varepsilon \qquad (2)$$

This ensures that among truly qualified candidates, the probability of selection is approximately equal across groups, with maximum permitted disparity ε. Unlike demographic parity, equal opportunity does not require identical selection rates when groups have genuinely different qualification distributions - it focuses specifically on ensuring that qualified individuals are not systematically disadvantaged by their group membership.

To enforce this constraint during model training, we adopt the Lagrangian approach developed by Agarwal et al. (2018). Let $L(\theta)$ denote the model's loss function (e.g., negative log-likelihood), and define the fairness constraint violation $g(\theta) = P(\hat{Y}=1|S=0,Y=1) - P(\hat{Y}=1|S=1,Y=1)$. The Lagrangian is (3):

$$L\_lag(\theta, \lambda) = L(\theta) + \lambda \cdot g(\theta) \qquad (3)$$

where $\lambda \geq 0$ is the Lagrange multiplier. The optimisation proceeds via a primal-dual algorithm: (1) Initialise θ and λ; (2) Compute predictions and estimate $g(\theta)$; (3) Update θ by gradient descent on L_lag with respect to θ; (4) Update $\lambda \leftarrow \max\{0, \lambda + \eta \cdot g(\theta)\}$ where η is the dual step size; (5)

Repeat until the TPR gap $|g(\theta)| \leq \varepsilon$ and the loss $L(\theta)$ has converged. Cotter et al. (2019) address the non-differentiability of the indicator functions implicit in TPR estimation by substituting smooth sigmoid surrogates for gradient computation while retaining hard constraints in the dual update step - a proxy-Lagrangian approach that achieves stronger convergence guarantees.

Additional constraints can incorporate neurorights protections, for example limiting the fraction of employees subject to neural monitoring, $N(x) \leq N\_max$, or restricting neurodata use to consented applications only. A compliance constraint $R(x) \geq R_0$ can encode a minimum level of audit-trail completeness required for EU AI Act conformity. Table 5 summarises the key model parameters, their interpretation, and guidance on calibration.

Table 5. Mathematical Model Parameters: Description, Calibration Guidance, and Typical Ranges
Source: Authors' own formulation. Theoretical basis: Agarwal et al. (2018). A reductions approach to fair classification. ICML (https://arxiv.org/abs/1803.02453); Cotter et al. (2019). Two-player games for efficient non-convex constrained optimization. JMLR, 20(94); Hardt et al. (2016). Equality of opportunity in supervised learning. NeurIPS 2016.

Parameter	Description	How to Set It	Typical Range
α (alpha)	Weight on business performance/productivity	Set high (e.g. 1.0) as baseline	[0.5, 1.0]
β (beta)	Penalty weight for algorithmic bias	Calibrate from legal fines or diversity targets	[0.2, 0.8]
γ (gamma)	Penalty weight for privacy risk	Scale with GDPR fine exposure and data sensitivity	[0.1, 0.5]
ε (epsilon)	Maximum allowed TPR gap between protected groups	Set to current observed disparity; tighten over time	[0.01, 0.10]
λ (lambda)	Lagrange multiplier; dynamically updated during training	Initialised at 0; updated via dual ascent	[0, ∞)

6.3. Simulation of Fairness-Productivity Trade-offs

Because the equal opportunity constraint reduces the model's degrees of freedom, fairness enforcement typically involves a trade-off with raw predictive performance. The shape and magnitude of this trade-off is organisationally consequential - it determines, for example, how many additional false negatives (qualified candidates wrongly rejected) must be accepted in order to close a given TPR gap. Simulation provides a principled method for quantifying this trade-off before deployment.

The simulation procedure proceeds as follows: (i) Generate or obtain a representative candidate dataset with features X, true qualification labels Y, and sensitive attributes S, reflecting the demographic composition of the relevant labour pool (e.g., the Romanian technology sector); (ii) Train a baseline model without fairness constraints and record overall accuracy and group-specific TPRs; (iii) Re-train the model for a range of λ values, recording at each step the fairness metric (TPR gap) and productivity metric (e.g., overall recall of qualified candidates); (iv) Plot the resulting Pareto frontier of fairness-productivity trade-offs; (v) Select an operating point on this frontier that satisfies legal requirements ($\varepsilon \leq$ legally mandated threshold) while minimising productivity loss. Alexander et al. (2024) apply an analogous Monte Carlo simulation methodology to European algorithmic recruitment tools, finding that meaningful improvements in hiring diversity can typically be achieved with modest reductions in overall predictive accuracy - often less than 5% - when fairness constraints are appropriately calibrated.

7. Case Study: Fair AI Hiring in a Romanian Technology Firm

To ground the theoretical framework in operational practice, we present a detailed case study of a hypothetical Romanian technology company - 'TechRomânia SA' - with 350 employees, primarily in software engineering, data science, and product management. The company operates in Bucharest and Cluj-Napoca and has recently piloted an AI-based resume screening and initial

interview-scoring system for junior technical roles, processing approximately 800 applications per quarterly hiring cycle.

The company's HR data reveal that over the prior three years, female candidates represented 38% of the applicant pool for technical roles but only 22% of those advanced to final-round interviews. An internal audit, triggered by a diversity complaint, found that the AI screening model had assigned systematically lower scores to application materials that referenced participation in female-identified student societies, used linguistic patterns statistically associated with female authorship, and listed internship experience with organisations primarily employing women. The root cause was a training dataset composed of historically successful hire records from a period when female representation in the firm's technical teams was substantially lower than today - a classic case of historical bias amplification.

Under the EU AI Act (which applies to TechRomânia SA as an 'operator' of a high-risk HR system), the company is required to conduct a conformity assessment, implement human oversight of AI-informed decisions, and demonstrate bias mitigation through documented testing. To comply and to address the fairness deficit empirically, the company's ethics committee commissioned implementation of the Lagrangian fairness optimisation framework described in Section 6.

Concretely, the team framed the objective as maximising a weighted recall of qualified candidates (maximising $\Sigma_i \hat{Y}_i Y_i$) subject to an equal opportunity constraint $|TPR\_female(\theta) - TPR\_male(\theta)| \leq 0.05$ - a tolerance of five percentage points chosen to be stricter than current disparity (approximately 16 points) while allowing for the modelling noise inherent in finite sample estimation. The Lagrangian dual was solved over 200 training epochs using stochastic gradient descent with a dual step size $\eta = 0.01$. The simulation showed that at the target tolerance ($\varepsilon = 0.05$), overall screening recall declined by approximately 3.4% compared to the unconstrained baseline - a modest cost judged acceptable by the ethics committee given its substantial impact on diversity outcomes. The Pareto frontier was visualised as a downward-sloping curve, with steep initial productivity gains from tightening fairness at high disparities and diminishing returns as ε approached zero (full parity).

Table 6. Policy Recommendations for Ethical AI and Neurotechnology Deployment in HR: Stakeholder Responsibilities Source: Authors' own synthesis based on: EU AI Act (Regulation 2024/1689, Art. 9, 13); GDPR (Regulation 2016/679, Art. 13, 22, 35, 51); Romanian Law 190/2018; EU Corporate Sustainability Reporting Directive (CSRD, 2022/2464); Chen (2023); Muhl and Andorno (2023); Adomaitis et al. (2022); Deloitte European HR Technology Survey (2024).

Stakeholder	Recommended Action	Legal / Ethical Basis
HR Departments	Conduct algorithmic impact assessments; adopt fairness-by-design; disclose AI use to candidates	EU AI Act Art. 9; GDPR Art. 13
Corporate Boards	Establish AI ethics committees; integrate ethical KPIs into C-suite targets	EU Corporate Sustainability Reporting Directive (CSRD)
National Regulators (e.g. ANSPDCP)	Issue binding guidance on AI in employment; enforce GDPR compliance for employee data processing	GDPR Art. 51; Romanian Law 190/2018
EU Policymakers	Clarify neurorights legislation; fund digital skills programs; provide SME compliance toolkits	EU AI Act; Proposed Neurotechnology Framework
Academic & Research Community	Develop Romania-specific bias benchmarks; publish open-source fairness-auditing tools	EU Horizon Europe R&I Framework

Following deployment, the company implemented a complementary governance structure: monthly bias audits comparing male and female candidate TPRs in production data; a disclosure policy informing all applicants that AI screening is used and providing a summary of the fairness

measures in place; and an appeals pathway through which candidates who believe they have been unfairly screened out can request human review by a senior HR manager not involved in the initial screening. These operational measures, aligned with the requirements of the EU AI Act and GDPR Article 22, demonstrate that mathematical fairness optimisation and procedural governance are complementary rather than alternative approaches to ethical AI deployment.

8. Conclusions and Recommendations

This paper has examined the ethical dimensions of artificial intelligence and neuroscience deployment in human resources management, combining empirical data from Romania and the European Union with theoretical frameworks from machine-learning fairness, neuroethics, and organisational governance. Several conclusions emerge with particular clarity.

First, a significant digital divide characterises AI adoption across the EU, with Nordic member states achieving adoption rates four to eight times higher than Romania and similar economies in Central and Eastern Europe. This gap creates a dual risk: Romanian organisations that rush to adopt AI without adequate preparation may encounter the discriminatory pathways documented in more advanced adoption contexts, while those that delay adoption may fall behind in productivity and talent competition. The optimal path is structured, governance-first adoption - deploying AI tools only when appropriate ethics and compliance infrastructure is in place.

Second, the disparity between low corporate AI adoption (3.1% formally in Romania) and relatively high individual employee usage (approximately 35% regularly using AI tools) represents a significant unmanaged risk. Employees experimenting with AI tools without organisational policies, training, or oversight may process sensitive personal or business data through third-party AI systems, creating GDPR compliance exposures. Organisations should treat informal AI use as a governance priority, establishing clear policies, providing quality-assured AI tools, and training employees in responsible AI use.

Third, algorithmic fairness in recruitment and evaluation is not merely an ethical aspiration but a legal obligation under the EU AI Act and GDPR. Organisations should implement the mathematical optimisation framework presented in Section 6 - or an equivalent formal approach - to demonstrate measurable compliance with equal opportunity requirements. The modest productivity cost of fairness constraints (approximately 3-5% reduction in overall recall in our simulation) is substantially outweighed by the legal, reputational, and ethical benefits of fair AI deployment.

Fourth, workplace neurotechnologies require the most stringent ethical scrutiny. Brain-derived data are unlike any other category of employee information: they can reveal cognitive states, emotional responses, and potentially even value systems that workers have never voluntarily disclosed and would not choose to disclose. The principle of cognitive liberty - the right to mental privacy and freedom from cognitive manipulation - must be treated as foundational. Any organisational use of neurotechnology should be limited to applications where the employee benefit is clear, consent is genuine and structurally protected, data access is minimal and time-limited, and no repurposing for employment decisions is permitted.

Table 6 summarises our key policy recommendations organised by stakeholder group, providing a practical reference for the diverse actors - HR departments, corporate boards, national regulators, EU policymakers, and researchers - whose concerted action is required to ensure that AI and neuroscience serve the flourishing of workers and organisations alike.

Several limitations of the present study merit explicit acknowledgement. First, the mathematical optimisation model, while formally correct and theoretically grounded, was validated through a single hypothetical case study rather than empirical deployment in a real organisation. The parameter values used in the simulation (α, β, γ, ε) were set by the research team based on regulatory guidance and literature benchmarks; in practice, calibrating these parameters to reflect an organisation's specific legal exposure, cultural context, and risk appetite requires longitudinal data and iterative testing that this study does not provide. Second, the empirical data on AI adoption rates and employee usage patterns draw primarily on cross-sectional surveys conducted in Romania

and the broader EU; the causal relationships implied in the governance-gap analysis remain correlational, and longitudinal panel studies would be needed to establish directional links between formal adoption, informal use, and compliance outcomes. Third, the study's treatment of algorithmic fairness focuses predominantly on binary gender as the sensitive attribute, reflecting data availability; real-world HR systems must address intersecting protected characteristics simultaneously (gender, age, ethnicity, disability status), and extending the Lagrangian framework to multi-attribute fairness constraints introduces computational complexity that future work should address. Fourth, the regulatory landscape described-in particular the EU AI Act implementation schedule and the emerging neurotechnology framework-remains in flux, and specific compliance obligations may evolve before this paper reaches practitioners. These limitations define a productive agenda for future research: empirical validation of the optimisation model in live HR deployments, longitudinal governance gap studies, multi-attribute fairness extensions, and continuous regulatory mapping.

In conclusion, the integration of AI and neuroscience into the workplace is neither straightforwardly beneficial nor inherently harmful. It is a domain in which outcomes depend critically on the quality of governance, the rigour of ethical reasoning, and the depth of commitment to human dignity. Organisations that approach these technologies as tools in service of people, governed by robust ethical frameworks, mathematical fairness constraints, and genuine accountability mechanisms, can realise their transformative potential without compromising the values of fairness, privacy, and autonomy that underpin democratic societies and decent work.

References

Adomaitis, L., Grinbaum, A., & Lenzi, E. (2022). *TechEthos Project Deliverable D2.4: Identification and specification of potential ethical issues and impacts of augmented/virtual reality and neurotechnologies*. Zenodo. https://doi.org/10.5281/zenodo.7619852

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). *A reductions approach to fair classification*. In *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80, pp. 60–69). PMLR. https://arxiv.org/abs/1803.02453

Alexander, L., Song, Q. C., Hickman, L., & Shin, H. J. (2024). *Sourcing algorithms: Rethinking fairness in hiring in the era of algorithmic recruitment*. *International Journal of Selection and Assessment, 33*(1), e12499. https://experts.illinois.edu/en/publications/sourcing-algorithms-rethinking-fairness-in-hiring-in-the-era-of-a

Autoritatea Națională de Supraveghere a Prelucrării Datelor cu Caracter Personal. (2024). *Ghid privind prelucrarea datelor cu caracter personal ale angajaților prin sisteme de inteligență artificială* [Guidance on the processing of employees' personal data via artificial intelligence systems]. Autoritatea Națională de Supraveghere a Prelucrării Datelor cu Caracter Personal

Bîgu, D., & Cernea, M.-V. (2019). *Algorithmic bias in current hiring practices: An ethical examination*. În *Proceedings of the 13th International Management Conference: Management Strategies for High Performance* (pp. 1068–1073). Bucharest University of Economic Studies. https://conferinta.management.ase.ro/archives/2019/pdf/5_8.pdf

Chen, Z. (2023). Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanities and Social Sciences Communications, 10*, Article 567. https://doi.org/10.1057/s41599-023-02079-x

Cotter, A., Jiang, H., & Sridharan, K. (2019). *Two-player games for efficient non-convex constrained optimization*. In A. Garivier & S. Kale (Eds.), *Proceedings of the 30th International Conference on Algorithmic Learning Theory* (Vol. 98, pp. 300–332). Proceedings of Machine Learning Research. https://proceedings.mlr.press/v98/cotter19a/cotter19a.pdf

Council of Europe. (1997). *Convention for the Protection of Human Rights and Dignity of the Human Being with Regard to the Application of Biology and Medicine: Convention on*

Human Rights and Biomedicine (Oviedo Convention) (CETS No. 164). Council of Europe Treaty Office

Dastin, J. (2018, October 11). *Amazon scraps secret AI recruiting tool that showed bias against women*. *Reuters*. https://www.reuters.com/article/amazon-com-ai-recruiting-insight-idUSKCN1MK0AG

eJobs Romania. (2024). *Studiu eJobs: 4 din 10 angajați folosesc frecvent instrumente de inteligență artificială la locul de muncă* [eJobs study: 4 in 10 employees frequently use AI tools at work]. https://www.ejobs.ro/stire/studiu-ejobs-4-din-10-angajati-folosesc-frecvent-instrumente-de-inteligenta-artificiala-la-locul-de-munca/

European Commission. (2016). *Regulation (EU) 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, and repealing Directive 95/46/EC (General Data Protection Regulation—GDPR). Official Journal of the European Union, L 119*, 1–88. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679

European Commission. (2000). *Council Directive 2000/78/EC of 27 November 2000 establishing a general framework for equal treatment in employment and occupation. Official Journal of the European Union, L 303*, 16–22. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32000L0078

European Commission. (2022). *Directive (EU) 2022/2464 of the European Parliament and of the Council of 14 December 2022 amending Regulation (EU) No 537/2014, Directive 2004/109/EC, Directive 2006/43/EC and Directive 2013/34/EU, as regards corporate sustainability reporting (CSRD). Official Journal of the European Union, L 322*, 15–80. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022L2464

European Commission. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. Official Journal of the European Union, L 2024/1689*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

Eurostat. (2022). *Self-perceived discrimination at work – statistics*. European Commission. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Self-perceived_discrimination_at_work_-_statistics

Eurostat. (2025). *Use of artificial intelligence in enterprises (data for 2023–2024).* European Commission. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Use_of_artificial_intelligence_in_enterprises

Hardt, M., Price, E., & Srebro, N. (2016). *Equality of opportunity in supervised learning*. *Advances in Neural Information Processing Systems, 29*, 3315–3323. https://home.ttic.edu/~nati/Publications/HardtPriceSrebro2016.pdf

Muhl, E., & Andorno, R. (2023). *Neurosurveillance in the workplace: Do employers have the right to monitor employees' minds? Frontiers in Human Dynamics, 5*, Article 1245619. https://doi.org/10.3389/fhumd.2023.1245619

Romanian Government. (2018). *Legea nr. 190/2018 privind măsuri de punere în aplicare a Regulamentului (UE) 2016/679 al Parlamentului European și al Consiliului din 27 aprilie 2016 privind protecția persoanelor fizice în ceea ce privește prelucrarea datelor cu caracter personal și privind libera circulație a acestor date* [Law No. 190/2018 on measures implementing GDPR]. *Monitorul Oficial al României, Nr. 651/26.07.2018*. https://portal.just.ro/101/Documents/LEGE__190_2018_date_cu_caracter_personal.pdf

Romania Insider. (2025, May 14). *eJobs Romania survey: 63.6% of employees face emotional difficulties at work*. *Romania Insider*. https://www.romania-insider.com/ejobs-employees-emotional-difficulties-may-2025

Romania Journal. (2024, June 6). *4 in 10 employees frequently use AI tools to ease job tasks*. *Romania Journal*. https://www.romaniajournal.ro/business/companies/4-in-10-employees-frequently-use-ai-tools-to-ease-job-tasks/

Romania Journal. (2025, February 27). *Almost 80% of employees use AI tools at work, survey says*. *Romania Journal*. https://www.romaniajournal.ro/business/almost-80-of-employees-use-ai-tools-at-work-survey-says/

TELUS Health Annual Conference 2024. (2024). *TELUS Health Annual Conference 2024*. TELUS Health. https://resources.telushealth.com/en-ca/telus-health-annual-conference-2024