ESCP
BUSINESS SCHOOL

IT ALL STARTS HERE

BERLIN | LONDON | MADRID | PARIS | TURIN | WARSAW

© TechNova Graphics / AdobeStock

LIGHTS - Innovation/Entrepreneurship

From ABACUS harms to ROBOTS benefits: A pragmatic framework for implementing and improving ethical and responsible AI

ESCP Impact Paper No.2026-16-ENG

Urs MUELLER - SDA BOCCONI

Martin KUPP - ESCP Business School

ESCP RESEARCH INSTITUTE OF MANAGEMENT (ERIM)

ESCP Impact Paper No.2026-17-ENG

From ABACUS harms to ROBOTS benefits: A pragmatic framework for implementing and improving ethical and responsible AI

Urs Mueller*
Martin Kupp**

ESCP Business School

Abstract

A responsible use of AI goes beyond a simple yes or no to deployment and cannot be ensured through technical fixes alone. A manager's task is to continuously make any use of AI, large or small, simultaneously less harmful and more beneficial. This requires not simply reviewing the AI system as a whole but reviewing every dimension for the harms (which we group as ABACUS) and benefits (ROBOTS) it creates separately and with human judgment. Plotted together, they form a practical decision grid: scale, safeguard, improve or park, deprioritize, or stop as currently designed. The framework does not replace judgment with a score; it helps leaders exercise practical wisdom in AI implementation.

Keywords: AI ethics, responsible AI, managerial decision-making, harm and benefit, practical wisdom; ABACUS harms and ROBOTS benefits

*Associate Professor of Practice, SDA Bocconi
* Professor, ESCP Business School

From ABACUS harms to ROBOTS benefits: A pragmatic framework for implementing and improving ethical and responsible AI

A decision on the Underground

In July 2025, Transport for London (TfL) faced an AI-related decision like many organizations are dealing with right now: whether to introduce or expand a system that promised benefits, despite prior errors and ethical concerns. Siwan Hayward, TfL's director of security, policing and enforcement, told the London Assembly that TfL was actively considering how more artificial intelligence could be used on the network (Lydall, 2025). Asked about a live facial recognition tool that scans faces in a crowd and matches them against a watchlist, she said she was "really alert" to the opportunity and described it as one element of TfL's thinking for the next five years. (Mayor of London, 2025)

Two years earlier, TfL had run a year-long AI trial at Willesden Green station that applied computer vision to the station's existing camera feeds. Documents later released under freedom of information showed the system doing far more than its public presentation initially suggested (O'Malley, 2024). It generated tens of thousands of alerts across a wide range of situations, from people pushing through ticket barriers to passengers standing too close to the platform edge. But at the same time, the system made mistakes that point to the need for continuous practical judgment about the benefits and harms: Unable to gather enough data to recognize aggression directly, the system was trained to flag anyone raising both arms and incorrectly flagged a man hanging a poster as a potential threat incident. It learned to spot children at the barriers only crudely, by disregarding anyone shorter than the gates. And similar errors could easily cause harm to individuals.

So here is a leader defending a contested AI system, and arguing for a more extensive use of AI, despite records about earlier errors. Whether one agrees with Hayward's argument or not, she faced the managerial decision about what to keep, what to cut, what to add, and what to watch given the technological capabilities arising from AI.

The same decision, on most managers' desks

Hayward's situation might look very specific because it involves a public transport network and a police watchlist. But despite the specific setting, the underlying structure is common to all organizations considering the adoption of AI for any process or product: AI offers substantial benefits, but at the same time there are often real harms to a broad set of stakeholders. Organizations have power over the people the system touches. And it is certain that all systems will sometimes be wrong about someone or something or cause other types of harm.

Consider the decisions being weighed in ordinary companies: Whether to monitor employees with AI, through productivity scoring, message analysis, or data collected from badges, wearables or cameras. Whether to screen job applicants with an algorithm. Whether to let an AI model decide how much bonus to pay, who is promoted and who is selected for a lay-off. Whether to read customer behavior and emotions in a shop, during a call, or through their behavior in an app. Whether to score fraud or credit risk automatically. Whether to give an AI agent authority to act, spend, or send messages on

the company's behalf. All of these are choices that are structurally similar to the choice Hayward faced: whether and how you should deploy AI for a real benefit even though you know that it might (at least occasionally) cause harm at the same time.

This is what makes these decisions ethical rather than merely technical or commercial: wherever an AI use harms or benefits people, ethical questions are unavoidable. We take this as a starting point: whom could a system harm and benefit and how? Our framework (see below) breaks these questions down into the 2x6 dimensions along which AI use can harm or do good to the people it touches.

The traditional conversations around "responsible AI" help to set a target, but not to plan managerial action: Fairness, transparency, accountability etc. are all correct, but they are too abstract to tell a manager what to do. They name principles (i.e., what to value) instead of guiding action (i.e. what and how to change). Part of the problem is that there are simply too many sets of responsible AI guidelines: surveys of ethics guidelines have found dozens of overlapping sets of principles with little agreement on how to apply any of them (Jobin, Ienca & Vayena, 2019). But managers do not need sets of principles, they need a way to improve ideas for AI use by investigating, dimension by dimension, where the harm and benefit lies, so that the harm can be reduced and the benefit be increased.

And to do so, managers need frameworks to guide their judgment and action. In this sense, an ethical consideration should not be understood as a fixed rule that is applied once. It is closer to what Aristotle called *phronesis*, practical wisdom: the capacity to judge well in a particular situation, under uncertainty, and to keep judging as the situation moves and as we learn more. Two things make a single verdict insufficient: (1) The context around a system changes, as new users, uses, and data arrive. (2) And the manager's own reading of the harms and the benefits sharpens with experience. Making responsible use of AI is therefore an ongoing managerial responsibility, not a one-time product innovation gate. The two lenses we propose below, ABACUS for harms and ROBOTS for benefits, are meant to inform such practical judgment.

Two lenses: ABACUS and ROBOTS

Most ethical analyses of a system ask a single, global question: is this responsible, good, or fair? That question is usually already hard to answer and even harder to act on. Our ABACUS and ROBOTS frameworks break harms and benefits from AI into parts. Each includes six dimensions, and the two sets mirror each other. The aim is that a manager can walk any AI use through twelve dimensions and come out with a list of things to improve.

ABACUS covers six families of the main AI harms:

- **Agency**, responsibility and control deficits: concerns the loss of human control and the blurring of who is responsible when a system acts; delegating a decision to a machine can also quietly change the behavior of the people who delegate (Köbis et al., 2025).
- **Bias** and errors of AI and digital systems: covers the data and design errors that make a system unfair or simply wrong, of the kind exposed when commercial face classifiers performed far worse on darker-skinned and female faces (Buolamwini & Gebru, 2018).
- **Abuse**, controversial use and overuse: covers malicious use, dual use, and the slow expansion of a tool beyond its original purpose (often without a new ethical assessment).

- **Copyright** and intellectual property violations: covers the unlicensed use of others' work as training data and the unsettled ownership of what a model produces.
- **Unemployment**, exploitation and societal divide: covers job displacement, deskilling, and the widening of economic divides (Brynjolfsson, Chandar & Chen, 2025).
- **Surveillance** and lack of control over data: covers intrusive collection, predictive profiling, and the conversion of data about people into power over them (Zuboff, 2019).

ROBOTS covers the matching families of AI benefits:

- **Responsibility**, human-centric design and empowerment: covers design that keeps humans meaningfully in control rather than removing them.
- **Objectivity**, inclusiveness and precision of analysis: covers the conditional gain in consistency, inclusiveness and precision that a well-built system might offer; even though biases might persist as "neutral" through alleged algorithmic "objectivity".
- **Beneficence** and positive use of digital technologies: covers the concrete good a system does for individuals, communities, and the wider world.
- **Open access** and intellectual empowerment: covers the widening of access to knowledge, tools, and opportunity.
- **Task** and job **creation**, equality and inclusion: covers the redesign of work so that capability is amplified and new tasks and roles are created, not only eliminated.
- **Security**, privacy and autonomy over data: covers the strengthening of privacy, data control, and resilience through AI.

These two lists are not meant to be mutually exclusive or exhaustive. They are managerial heuristics for reflection. Beneath each of the twelve headings sits a finer set of sub-dimensions, three to each, which an organization can use when a particular use needs closer work (see exhibit 1).

The two sets of harms and benefits are meant to correspond to each other without being zero-sum: Bias as a harm is contrasted with Objectivity that could (at the same time) be gained by AI. Similarly, Surveillance sits opposite Security; Unemployment opposite Task creation, and so on. An AI tool might very well create a harm on one of the six dimensions and simultaneously create a corresponding benefit, which is exactly why managers should look at both at the same time, but independently.

The decision grid

Plotting the two lenses of harms and benefits from a given AI application against each other turns them into a decision tool: Put the ABACUS harms on a vertical axis, and the ROBOTS benefits along a horizontal one. A system can be high on both at once, which is why only asking "how ethically problematic is it?" is insufficient.

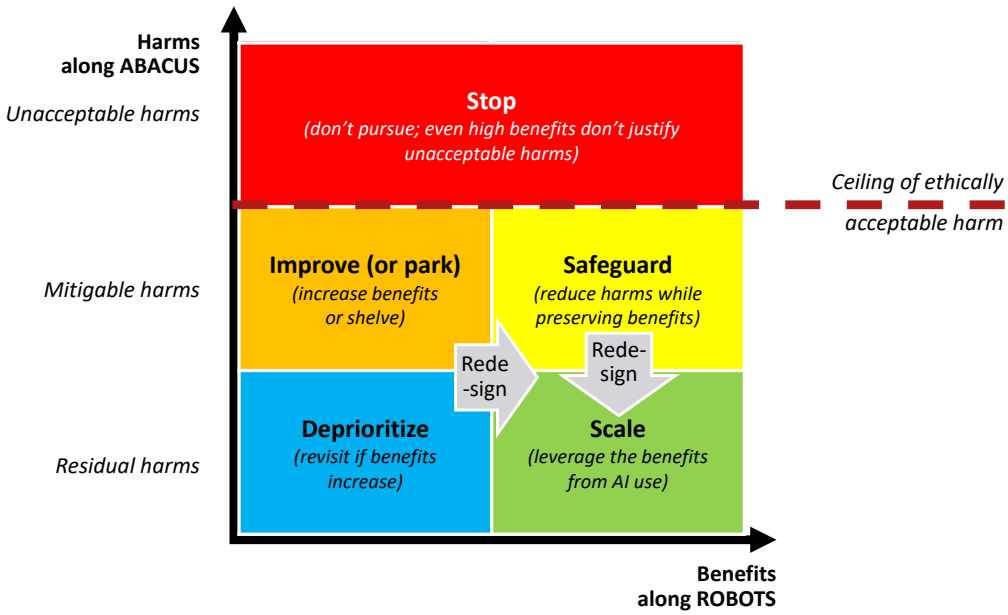

Figure 1: The ABACUS–ROBOTS decision grid

The vertical axis has three bands rather than two, because not all harm is similar in nature. At the bottom are residual harms, where harms are low or already controlled. In the middle are mitigable harms, where harms are real, but safeguards can bring them within the range of tolerance. At the top are unacceptable harms, where the harms cannot be brought within ethical tolerance as the system is currently designed. The line between mitigable and unacceptable harms might seem partially drawn by regulation such as the EU AI Act. But managers need to develop the practical judgment to define a line of ethically acceptable harm, which might well be stricter than regulation. In practice this threshold should not be set by an individual or the project team alone, but should include the perspectives of a broader set of stakeholders (senior management, legal and compliance, users, representatives of affected parties etc.) wherever possible. Cases of (i) severe and/or (ii) irreversible harm to individuals or groups (iii) who have not consented and (iv) who might be vulnerable are particularly likely to be above the threshold. And because the generation of harms and benefits can change, the ceiling of ethically acceptable harm is not permanent for a given system. A use that looks unacceptable when first designed can often be re-engineered until it sits below the line or suddenly creates new and unacceptable types of harm, which is why "stop" or “go” are rarely final answers. Instead, organizations need to identify owners per deployment who should set a review triggers. This should include fixed intervals, but also event triggers, such as any material modification of the systems, new data sources and cases of actual or near-miss harm. Changes in ABACUS and ROBOTS scores over time should be logged and stored to capture trends and developments.

Below the ceiling, six positions describe where a use can sit. With only residual harm and high benefit, the appropriate move is to **scale** and capture the benefit. With residual harm and low benefit, the move is to **deprioritize**, leaving a reasonably safe but unrewarding use for later. With mitigable harm and high benefit, the use might be worth keeping and the move is to **safeguard**, driving the harm down through controls or changes to the substance of the AI system. With mitigable harm and low benefit, the move is to **improve or park**: grow the benefits, or shelve it, since heavy safeguards are rarely worth it for weak or only speculative ethical benefits. Above the ceiling of ethically acceptable harm, at any level of benefit, the move is to **stop**, as the system currently stands.

The most important element is not any single box but the move between them. We call it **redesign**, and it is the general posture the grid encourages rather than one action. To redesign is to reduce ABACUS harms and raise ROBOTS benefits at the same time. On the grid, redesign points down and to the right: downward by reducing ABACUS harms, and rightward by increasing ROBOTS benefits. For almost any use below the ceiling, the best direction of travel is that diagonal. For a use sitting above the ceiling, redesign is what can bring it below. But it might be important to note that managers should not consider an upward pointing redesign to be a valid option: increasing a system's benefits will not automatically permit it to cause greater harm.

And redesign is never finished. A use does not necessarily stay in the same place on the decision grid. Its position drifts because the context drifts, technology evolves, new users, uses and data appear, and because the manager's own assessment of the harms and benefits sharpens over time. A use that sat safely in the scale position can rise into safeguard, or above the ceiling, without anyone changing a line of code. A use stuck at stop can become deployable once it is redesigned. The grid is therefore a position to re-check regularly, not a clearance granted once and forever. The question is never only "can we deploy this?" but, dimension by dimension and repeatedly, "how do we push each ABACUS harm down and each ROBOTS benefit up?"

Walking TfL through the grid

The transport case shows the grid doing real work. Read along ROBOTS, the deployment of AI in public transport can have genuine value. Detecting someone at the platform edge serves Beneficence directly. Helping passengers with baby strollers or wheelchairs serves accessibility. Recovering revenue lost to fare evasion, which runs into the tens of millions of pounds a year, funds a network that aims at serving everyone.

Read along ABACUS, the harm is equally real, and the finer sub-dimensions show why. Under Surveillance, the system risks moving from watching a space to profiling the people in it. Under Bias, a misclassification of individuals as fare-dodgers is a clear case of being confidently wrong about who has done what. Under Abuse, the initial trial illustrated function creep precisely: a capability introduced to flag safety incidents came to serve fare enforcement and behavioral monitoring, an expansion of purpose that was not clearly visible to the public. Under Agency, the question is who answers when the machine is wrong, and whether a human reviews each alert before anyone acts on it.

Our framework therefore points towards redesign, i.e., the goal of reducing the ABACUS harms and raising the ROBOTS benefits together: (1) Cutting the harms by not tracking individuals and their demographics, keeping a human in the loop before any action, strictly narrowing the permitted uses, publishing where and how the system runs, and deleting non-matches at once. (2) And raising the benefits by pointing the AI capabilities toward the uses where the payoff is high and the intrusion low, platform-edge safety and accessibility, rather than toward blanket enforcement. Done well, the use moves downward and to the right: toward Scale on use cases related to safety, and toward strict Safeguard for enforcement use cases. And like this the grid also prevents a common mistake: treating "the AI system" as one indivisible object. Platform-edge detection, accessibility support, fare-evasion monitoring and live facial recognition do not necessarily occupy the same cell.

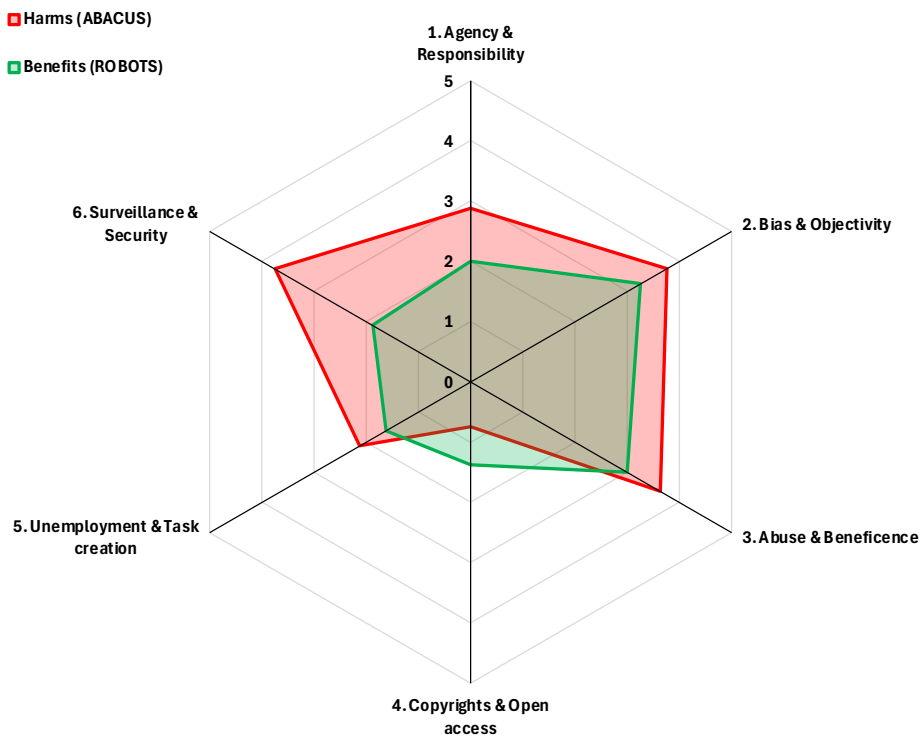

Figure 2: Illustrative ABACUS–ROBOTS diagnostic profile for TfL's AI-deployment

Notes:
Each axis carries two independent scores: the ABACUS harm (red) and the ROBOTS benefit (green).
The scores are an outside-in judgment by the authors based on publicly available information.
They are not meant to be a definite judgment, but rather an illustrative example.

The case also shows why redesign cannot stop at launch. The position did not hold still. The 2022 trial was scoped one way; by 2025 the same organization was weighing live facial recognition, which might push the system back up the ABACUS axes and demand a fresh redesign pass. The grid had to be run again, because the use, the technology and the leader's reading of the stakes had all moved. The verdict is not yes or no, and it is never final. It is the leader, not the algorithm, who owns the decision about where the ceiling sits, which way to push, and when to look again.

Responsible AI as a standing task

AI can automate a great deal. It can predict, classify, generate and recommend. It cannot automate responsibility, and it cannot do a manager's judging for them. The more capable these systems become, the more that judgment matters, because the harms and the benefits might both increase. Future research should first aim at ensuring inter-rater reliability and construct validity. Then a structured data collection along the ABACUS and ROBOTS frameworks should help to improve predictive validity by comparing pre-launch ratings to measurable outcomes, such as error rates, complaints, unequal impact on different groups, service quality, customer satisfaction etc. In this way, it will be possible to assess to what extent factoring in ethical considerations along ABACUS and ROBOTS will help to move an AI use case to the lower right quadrant of the decision grid.

Instead of removing the need for judgment, the ABACUS and ROBOTS frameworks and the decision grid aim at providing structure for practical judgment (*phronesis*) for managers. The twelve dimensions give people in any type of organization a shared language for asking where a use is harmful and where it is beneficial, and the grid turns those answers into a direction of travel. Three questions carry most of the work: (1) What is

the most serious ABACUS harm this use creates for the people it touches, and is it below or above our ceiling? (2) Which ROBOTS benefits are concrete rather than merely claimed? (3) And who remains answerable, and following what schedule do we examine the question again, when the system is wrong or the context shifts?

Responsible AI, in this view, is not achieved through technical solutions alone, but requires managers across any organization to employ practical judgment that takes into account the fact that business ultimately is a shared social activity for human benefit.

Exhibit 1: The complete ABACUS and ROBOTS taxonomy

An overview of all ABACUS/ROBOTS categories and sub-categories

ABACUS 1: Agency, Responsibility and Control Deficits
1. Loss of Human Agency
2. Responsibility and Control Deficits
3. Deceptive Anthropomorphism

ABACUS 2: Biases and Errors of AI and Digital Systems
1. Data-Related Biases and Contextual Errors
2. Algorithmic Biases and Technical Failures
3. Human-AI Interaction Failures

ABACUS 3: Abuse, Controversial Use and Overuse
1. Abuse and Malicious Use
2. Controversial and Dual Use
3. Overuse and Function Creep

ABACUS 4: Copyright and Intellectual Property Violations
1. Unauthorized Input Use
2. Ambiguous Output Ownership
3. Creator Protection and Attribution Issues

ABACUS 5: Unemployment, Exploitation and Societal Divide
1. Job Displacement and Skill Obsolescence
2. Exploitation in AI Development
3. Economic Dependencies and Inequality

ABACUS 6: Surveillance and Lack of Control over Data
1. Intrusive Data Collection and Consent Failures
2. Surveillance-Based Data Processing and Predictive Profiling
3. Exploitative Targeting and Behavioral Manipulation

ROBOTS 1: Responsibility, Human-Centric Design and Empowerment
1. Empowerment, Accessibility and Inclusion
2. Human-Centric Design and User Understanding
3. Control and Customization

ROBOTS 2: Objectivity, Inclusiveness and Precision of Analysis
1. Input Advantages
2. Processing Advantages
3. Output Accuracy and Stability

ROBOTS 3: Beneficence and Positive Use of Digital Technologies
1. Individual Welfare
2. Community and Institutional Welfare
3. Planetary Welfare

ROBOTS 4: Open Access and Intellectual Empowerment
1. Knowledge Access and Dissemination
2. Knowledge Creation
3. Knowledge Preservation

ROBOTS 5: Task and Job Creation, Equality and Inclusion
1. Professional Innovation
2. Capability Amplification
3. Access Democratization

ROBOTS 6: Security, Privacy and Autonomy over Data
1. Cybersecurity Protection
2. Privacy-Enhancing Innovation
3. Digital Rights and Data Control

References

Brynjolfsson, E., Chandar, B., and Chen, R. (2025). Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence. Stanford Digital Economy Lab, Working Paper.

Buolamwini, J., and Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the Conference on Fairness, Accountability and Transparency*, PMLR, 77–91.

Jobin, A., Ienca, M., and Vayena, E. (2019). The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence*, 1, 389–399.

Köbis, N., Rahwan, Z., Rilla, R., et al. (2025). Delegation to Artificial Intelligence Can Increase Dishonest Behaviour. *Nature*, 646, 126–134. https://doi.org/10.1038/s41586-025-09505-x

Lydall, R. (2025). Tube fare dodging: live facial recognition cameras could be used to catch most prolific evaders. *Evening Standard*, 9 July. Available at: https://ca.news.yahoo.com/facial-recognition-cameras-could-introduced-145113099.html

Mayor of London (2025). Agenda and minutes Transport Committee – Tuesday 8 July 2025 10.00 am. Appendix 2. Available at: https://www.london.gov.uk/about-us/londonassembly/meetings/ieListDocuments.aspx?CId=173&MId=7910

O'Malley, J. (2024). TfL's AI Tube Station experiment is amazing and slightly terrifying. *Odds and Ends of History* (Substack), 12 February. Available at: https://takes.jamesomalley.co.uk/p/tfls-ai-tube-station-experiment-is

Zuboff, S. (2019). "The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power". New York: PublicAffairs.