This paper was presented at the XXXVII ISPIM Innovation Conference, held in Granada, Spain on 07 June to 10 June 2026. ISBN 978-952-65771-2-8.

Operationalizing ethical principles for artificial intelligence in real-world innovation

Erich Prem

Vienna University of Technology, Karlsplatz 13, A-1040 Austria and
eutema GmbH, Lindengasse 43/13, A-1070 Vienna, Austria.
E-mail: prem@eutema.com

Abstract: This contribution examines the central innovation management challenge of translating abstract ethical principles for artificial intelligence into actionable design and governance practices. Drawing on seven real-world AI innovation cases from Austrian industries, we analyse how digital humanism—a framework advocating for human-centred and societally aligned digital technologies—can guide the practical management of AI-based innovation. We identify key obstacles that innovators encounter, including the difficulty of anticipating undesirable effects of AI, navigating an overwhelming range of design choices, hesitations around enhancing user autonomy, and the risk of workforce deskilling. We propose a digital-humanist approach to innovation management that emphasizes early-stage ethical integration, cross-disciplinary collaboration, and capability-building as foundational components for developing responsible and trustworthy AI systems.

Keywords: Artificial intelligence; AI; AI ethics; AI innovation; AI risks; AI practice, AI risk management.

1 Introduction

Designing ethically aligned AI systems

In recent years, the design and governance of digital innovation—particularly AI systems—has come under increasing scrutiny due to mounting concerns about their ethical, social, environmental, and organizational consequences. Digital humanism, a growing interdisciplinary movement, argues for placing human agency, societal values, and democratic principles at the centre of digital system design (Prem, 2024a). Digital humanism is a broad philosophical approach to the design of ethical systems with the intent of putting humans and democratic values first (Werthner et al. 2022, Werthner et al. 2024). It has developed from an academic initiative to a broad intellectual, political, and increasingly engineering movement. The European Digital Humanism Initiative (EUDHIT) demonstrates this shift with its dedicated aim to ground philosophical ideas in practical approaches, involve broader audiences from industry, the education sector, and policy makers. Although the movement has gained substantial academic momentum, its practical implications for innovation management remain insufficiently explored.

This paper therefore addresses the ethical challenges of managing AI innovations from both a practical and digital humanist perspective. Our analysis draws on seven real-world innovation cases from Austrian companies participating in the

collaborative research initiative FAIR-AI. The cases span a diverse set of industries, including e-government, engineering, waste management, human–computer interaction, renewable energy, energy management, and work planning. Each case is developed by engineering teams working in close collaboration with AI ethics researchers and legal scholars, ensuring that ethical and societal considerations are embedded throughout the innovation process.

This work contributes to a broader effort to translate abstract AI ethics principles into concrete, actionable design and management practices (Zhou & Chen, 2023). This translation—moving from high-level frameworks to operational guidance—has emerged as one of the most pressing challenges in contemporary AI innovation management (Prem, 2023). Overall, our findings highlight both the necessity and the complexity of embedding digital humanism within AI innovation processes, demonstrating that ethical management is not an add-on but a core capability for responsible, future-oriented innovation.

2 Methodology and approach

Work reported here researches the gap in managing societal risks in AI by focusing on the practical implementation of the EU AI Act and AI principles in real-world AI innovations (FAIR-AI). It identifies a range of obstacles: technical (such as data bias in machine learning), managerial (managing AI risks), and socio-technical (like societal implications of AI-assisted decisions). It emphasizes the importance of anticipating, detecting, monitoring, and managing risks throughout the AI lifecycle, i.e. data collection, data formatting, model training, model testing etc. (Prem, 2022; Prem, 2024b). Adopting a bottom-up methodology, the project studies concrete use cases that illustrate more general pitfalls, using them to study and disentangle different risks. It advances existing approaches by integrating risk analysis towards approaches that provide actionable guidance. FAIR-AI places strong emphasis on ethics and legal considerations developing workshops, tools, and training modules in collaboration with industry partners to reach responsible and compliant AI practices. This is a collaboration of industry and academic partners with a solid background in legal, societal, philosophical, and ethical aspects of AI. In practice, this means intense discussions with company representatives presenting and discussing their cases – both at the management level and with implementation and engineering experts. In some cases, marketing or legal experts were also involved. Our cases include a mix of small companies and large enterprises.

The AI-driven innovation cases of our industry partners cover a broad range of sectors. They range from construction engineering to energy, software, waste disposal, work organisation etc. The applications include AI systems for scheduling work, LLM-based chatbot applications, and AI for classification tasks. Early in the project, each industry innovation case was discussed from an ethical, legal, and societal perspective with the experts in the consortium. The ethics and legal experts raised issues with the companies, the development teams, and in internal debates. Consequently, the projects were adapted over the course of the project to manage the legal, ethical, or societal risks and issues. This includes engineering adaptations, changes suggested from the legal or AI ethics experts in the team, and more general changes driven from the side of the industry partners.

In the following, we report about the experiences from the cases without disclosing the companies, their precise innovation cases, and detailed problematic ethics issues as the focus in this paper is on the overall lessons learned and not on the single cases. It must therefore be noted that the evidence presented here remains necessarily anecdotal.

3 Ethics issues and observations

The challenge of predicting and discussing undesirable outcomes

The design of ethically aligned AI systems is subject of a broad academic and to some extent also the engineering literature. The topic has been subject of policy papers, legal and regulatory approaches, philosophical work, political scholarly work and work in other areas (e.g. in the Arts). However, it has only rarely been driven by an innovation- and risk-based management perspective (Prem, 2024b). Addressing critical or risky issues first requires the ability to *identify* ethical, legal, or societal risks of an AI application. Unfortunately, the academic literature today provides little guidance on how to perform such assessments. Existing efforts to systematically approach this problem are laudable (Ayling & Chapman, 2022, Tartaro et al., 2024, Nitta et al., 2022), however, they often – perhaps necessarily – remain at a very general level.

From our cases it became clear early in the debates that both engineers and management-level executives found it hugely challenging to explore the potential problems of their own proposed AI innovations. There was a tendency in several of the cases for the engineers to argue that their ideas either had no critical issues or that they were sufficiently considered or that they were negligible in practice. The AI ethics and legal risk experts on the other hand could quickly see potentially detrimental effects of the proposed AI innovations and identified a range of problems.

For example, in the case of an application in waste management, the legal team quickly identified significant potential legal and ethics issues. These ranged from individual privacy violations to questions of fairness, social and even political implications and issues at the level of local communities. Similarly, in the case of a work scheduling application, the ethics and legal experts quickly discussed a lack of explanation, potential lack of staff autonomy regarding their scheduling, data protection aspects, and responsibilities – especially for suboptimal schedules. For the chatbot case, issues of reliability of recommendations, precise design of the interaction pattern, and legal responsibility were quickly raised.

Although the innovators had considered many of the issues, they also showed clear limitations regarding the extent to which they would consider potentially detrimental effects. The discussions of the legal and ethical experts with the system designers and company representatives imply the following observations.

(O1) Ethics risk assessment requires *substantial background knowledge* about potential shortcomings and identified ethics and societal issues, e.g. from literature or previous cases. These competencies are substantially different from the domain expertise of the involved AI and IT system designers.

(O2) Ethics and legal experts approach the system from a type of *worst-case* or *devil's advocate* attitude. Both often start from analyses of what might go wrong or that there will be maximum evil intention when using the system. Designers and innovators do not usually use this thinking for their own systems despite well-established worst-case scenario technique (Yoe, 2011).

Trust and perspective in ethics debates

It was anticipated in our research that discussions about potential ethics and legal issues in a setting of experts from within and outside the companies needed a careful approach to ensure a trusted environment. The aim was to create a setting where experts could freely consider actual or potential shortcomings. This can be challenging given that the shortcomings need to be identified and discussed with people from outside the company. Moreover, this often happened in the presence of a team leader or management representative of the company. There is, hence, an understandable tendency to either downplay potential issues or argue that risks can be easily mitigated. The establishment of a sufficient level of trust for an open critical debate takes time.

In addition, there is a tendency of the designers to perceive the identification of ethics or legal shortcomings as a negative evaluation of their own work. It was made very clear in the project that this was not the point of the exercise. However, it can be hard for the system owners to switch perspective in this respect and become a devil's advocate of own work.

(O3) It can be *personally challenging for the system designers* to question their own design decisions. There is a tendency to defend design choices with respect to potential detrimental effects of the AI system. This situation can be worse where management-level executives partake in the discussion.

Risk avoidance versus risk mitigation

Originally, the project's hypothesis was that identified risks would be addressed with a range of existing tools and methods, e.g. statistical methods to debias data or data science approaches to improve explainability. To the surprise of the ethics and legal team, the companies seemed to often react to identified risks with a general risk avoidance attitude at the application level. This meant that they would prefer a change in the overall system design or application rather than only tackle a risk with engineering solutions (e.g. de-biasing algorithms) or systemic tools (e.g. contractual terms). Instead of directly addressing an identified shortcoming at the level of system design, e.g. by changing the implementation, advising users, or changing Terms of Reference, the whole application or original idea behind the innovation was questioned. This led to significantly redesigned innovations with changed functionality compared to the original idea.

(O4) Innovators may tend to be *risk averse at the application level* or at the basic innovation design. They will rather avoid risks at the highest abstraction level than target those risks during the implementation or with legal constraints such as terms and conditions.

Navigating design options

The use cases cover a broad range of types of applications and sectors. The main ethics issues are concentrated around *data governance, accountability, transparency, and human oversight*, with additional concerns about *privacy, fairness, deskilling, safety*, and *sustainability*. Across the cases, the ethical challenges are mostly about how AI changes decision-making, responsibility, and trust in real operational settings rather than about AI in the abstract. The ethical pressure points are the same ones that appear whenever AI affects work, rights, or decisions at scale—who is responsible, who understands the system, who can contest outcomes, and whose data or labour makes the system possible.

The current approaches to tackle the issues are similar across cases: they mostly rely on *human-in-the-loop review, role clarification, controlled deployment, communication measures, and further legal or governance clarification.* In other words, the proposed solutions are less about fully solving each ethical issue and more about containing risk through oversight, process design, and clearer responsibility. The proposed responses are mostly mitigation strategies rather than final fixes: *keep humans accountable, make the system understandable, restrict risky use, and clarify legal responsibilities* before broader deployment.

On the other hand, there is an enormous breadth of options to address ethics issues. Today, system designers can tap into a wide landscape of tools, methods, and broad approaches to mitigate ethics or level issues and risks (Prem, 2023). Methods, tools and approaches are constantly evolving but can overwhelming for designers. Often, system designers will excel only in a few of those but remain generally unaware of the different levels and options where these issues can be addressed.

For example, a privacy issue can be addressed at the data level, algorithmically, through access restrictions, computational limitations, consent rules and many more approaches. However, there is today no readily available guidance on which tool to choose or which level – from implementation to the application – to target.

(O5) Choosing the right tools and approaches for mitigating AI risks is a huge challenge given the enormously broad range of suggestions in the literature. Consequently, innovators may resort to more general, higher-level risk mitigation strategies, e.g. keeping humans in the loop etc.

High-level design choices – User autonomy versus prescription

A very important and indeed basic design choice is the level of user autonomy. This is a central principle of many of the existing AI ethics frameworks. Many scholars now suggest precisely five principles for AI ethics considerations: autonomy, beneficence, non-maleficence, fairness, and explainability (or variations of these principles). However, just autonomy is notoriously difficult as regards how much decision-making power users should have. Judging from our experience, industry partners seem to find it difficult to fully embrace approaches that maximize user autonomy compared to implementations that make systems "safe" or "unproblematic". Note that this runs somewhat against the previous observation (O5) to keep users in the loop.

For example, the degree of moderation for chatbots, e.g. how much should dialogue with users be thematically restricted, is debatable. From an autonomy perspective, there should be little interference with user intent and discourse should be quite free. From the business perspective, however, it seems better to limit dialogue

systems to text that is "unproblematic" and solely focused on anticipated use cases so as to minimize business risks (e.g. liability or negative publicity).

Addressing emerging skills gaps

Most, if not all the AI innovations studied here, suggest that some kind of de-skilling takes place as a result of using the developed innovations. This is relevant business-internally, as in the case of generative AI for computer programming, or with respect to the application, as in the case of using AI expertise for finding construction faults or material degradation. For the companies, de-skilling comes in two variants:

- deskilling of the labour force as AI systems increasingly take over tasks previously performed by human experts but require fewer human skills and hence provide fewer options to learn
- addressing risks emerging from increased automation, e.g. loss of control through automated generation of code, schedules, etc. which implies more and new skills that are needed, for example to move from code generation to code test and verification.

The companies in our set showed increasing awareness of these issues over the course of the discussions with some being relatively unaware about them at first. It was evident, however, that deskilling is a significant challenge given that there are insufficient guidelines or practical schemes to address the issue. Although the problem is being recognized as significant and structural (Ferdman, 2025), it remains an important open research problem (Crowston & Bolici, 2025).

(O6) AI poses significant challenges for maintaining and adapting the right skill set of users – including own staff. Where AI automates skilled work, there is a risk of de-skilling of the labour force. In addition, increased automation may require new and elaborate skills to monitor and evaluate AI outputs.

4 Conclusions and further work

Even for smaller AI-based innovations, the implementation of ethics principles proves difficult for a range of reasons and on several levels. Firstly, there are significant challenges to predict undesirable or emergent ethical outcomes in practice. It is necessary but can be hard to understand the consequences of digital tools for individuals or at the societal level. Secondly, the identified challenges need to be addressed practically. Different from only a few years ago, system designers now have a wide set of options and tools that can be used to address issues such as bias, fairness, explainability etc. The landscape of available tools and options is continuously growing and some aspects such as explainability have developed into whole research fields, e.g. explainable AI or XAI. Thirdly, systems need to be reassessed from a broader perspective including beyond their direct ethical assessment, e.g. regarding risks and business impact. At this level, business have options of changing an originally planned innovation, adapting terms of use, disclaimers etc. Industry is also facing difficult choices of a range of high-level options or policies, e.g. regarding designs that to maximize user autonomy or minimize potential risks (Prem, 2024 b). In addition, businesses must increasingly deal with concerns about

the effects of AI systems internally. This may include deskilling as AI systems increasingly take over tasks previously performed by human experts, cf. Table 1.

The challenges manifest across multiple organizational layers—from individual engineers making design decisions to leadership teams responsible for strategic direction, innovation management, and risk governance. Addressing these issues requires more than ethical awareness; it demands targeted training, domain-specific expertise, and organizational structures that integrate ethical reasoning into everyday innovation practices. While emerging support tools and methodological frameworks hold promise, that building the necessary competencies remains essential for ethically aligned management of AI-based digital innovations.

Table 1 Major design steps for ethical systems implementation

Design step	*Characterisation*	*Challenge*
Identification of ethics issues	Anticipation of potential ethical risks and issues	Ability to correctly predict potential flaws and ethics issues; availability of ethics experts; limitations of system designers to critically assess own design choices
Addressing ethics issues	Selection of tools and methods	Selecting from a set of large and growing options; following new developments with little guidance and recommendation
Overall approach and policies	Redesigning the overall business case; boundary conditions and use scenarios	Balancing risk management at the overall innovation with targeted tools at the implementation or realisation level
Skills and workforce	Monitoring and developing the AI innovation and its effects	Skills-gap identification, continuous training and shift of necessary skills

This work is part of a broader effort to translate abstract AI ethics principles into concrete, actionable designs and management practices. This translation—moving from high-level frameworks to operational guidance—has emerged as one of the most pressing challenges in contemporary AI innovation management. Through a comparative analysis of a range of real-world cases, we identified persistent obstacles to implementing digital-humanist principles in practice. These include:

- The difficulty of predicting undesirable or emergent outcomes due to lack of ethics-specific knowledge and psychological challenges to question own designs
- The challenge of navigating a wide landscape of design options and methodological choices
- The preference of avoiding ethics issues at the level of the innovation objectives
- Hesitation among industry partners to embrace approaches that maximize user autonomy
- Concerns about deskilling as AI systems increasingly take over tasks previously performed by human experts.

In response, we propose a digital-humanist approach to innovation management that emphasizes early-stage ethical integration, cross-disciplinary collaboration, and capability-building as foundational components for developing responsible and trustworthy AI systems. More research is needed to better guide AI system designers and innovators in understanding the potential effects of their systems at the ethical and societal level and to select and adopt tools described in an increasingly complex field targeting the design of ethical AI systems.

Until such results are widely available and standardized approaches become available, risk containment through oversight, humans-in-the-loop, clear responsibilities and other philosophies aligned with digital humanist principles are commendable practical approaches. This also includes ethical integration in the engineering design process, cross-disciplinary collaboration with ethics and legal experts to ensure proper identification of issues and capability-building.

Acknowledgements

This research was funded by the European Commission (Digital, Industry, and Space) as EU project EUDHIT - *The European Digital Humanism Initiative* – Grant Agreement ID 101212890 and by the Austrian Research Promotion Agency FFG as FAIR-AI, *Fostering Austria's Innovative Strength and Research Excellence in AI*, project nr. 904624.

References

Ayling, J., Chapman, A. (2022) Putting AI ethics to work: are the tools fit for purpose?. *AI Ethics* **2**, 405–429. https://doi.org/10.1007/s43681-021-00084-x

Crowston, K., & Bolici, F. (2025). Deskilling and upskilling with AI systems. *Information Research an international electronic journal*, *30*(iConf). https://doi.org/10.47989/ir30iConf47143

Ferdman, A. (2025). AI Deskilling is a structural problem. *AI & SOCIETY*, 1-13. https://link.springer.com/article/10.1007/s00146-025-02686-z

Nitta I., Ohashi K., Shiga S., and Onodera S. (2022) 'AI Ethics Impact Assessment based on Requirement Engineering', *2022 IEEE 30th International Requirements Engineering Conference Workshops (REW)*, Melbourne, Australia, 2022, pp. 152-161, doi: 10.1109/REW56159.2022.00037.

Prem E. (2022) A knowledge-based perspective of strategic AI innovation management. In: Tanev S., Blackbright H. (Eds.) *Artificial Intelligence and Innovation,* World Scientific, February 2022, https://www.worldscientific.com/doi/abs/10.1142/9781800611337_0003

Prem, E. (2023) 'From ethical AI frameworks to tools: a review of approaches', *AI and Ethics*, 3, pp. 699–716. https://link.springer.com/article/10.1007/s43681-023-00258-9

Prem, E. (2024 a) 'Principles of digital humanism: a critical post-humanist view', *Journal of Responsible Technologies*, 20 January, p. 100075. https://www.sciencedirect.com/science/article/pii/S2666659624000015

Prem, E. (2024 b) 'Managing entrepreneurial AI ethics risks', in Hoffmann, C.H. (ed.) *Artificial Intelligence, Entrepreneurship and Risk Management: Reflections and Positions at the Crossroads between Philosophy and Management*. Wiesbaden: Springer Fachmedien. https://link.springer.com/chapter/10.1007/978-3-658-45544-6_23

Tartaro, A., Panai, E. & Cocchiaro, M.Z. AI risk assessment using ethical dimensions. *AI Ethics* **4**, 105–112 (2024). https://doi.org/10.1007/s43681-023-00401-6

Werthner, H., Prem, E., Lee, E. A., & Ghezzi, C. (Eds.). (2022) *Perspectives on digital humanism*. Berlin: Springer. https://link.springer.com/book/10.1007/978-3-030-86144-5

Werthner, H., Ghezzi, C., Kramer, J., Nida-Rümelin, J., Nuseibeh, B., Prem, E., & Stanger, A. (2024). *Introduction to digital humanism: A textbook.* Cham: Springer Nature. https://link.springer.com/book/10.1007/978-3-031-45304-5

Yoe C. (2011) *Principles of Risk Analysis: Decision Making Under Uncertainty*. CRC press, p. 429-30. https://doi.org/10.1201/9780429021121

Zhou, J., & Chen, F. (2023) 'AI ethics: from principles to practice', *Ai & Society*, *38*(6), 2693-2703. https://link.springer.com/article/10.1007/s00146-022-01602-z