AI·PROFICIENT

Artificial intelligence
for improved production efficiency,
quality and maintenance

Deliverable 6.4

D6.4: AI-PROFICIENT ethical recommendations

WP 6: Use case evaluation and ethics considerations

T6.4: Instantiation of HLEG guidelines and ethical recommendations

Version: 1.0

Dissemination Level: PU

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957391.

Table of Contents

List of Figures

List of Tables

Disclaimer

This document contains description of the AI-PROFICIENT project work and findings.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any responsibility for actions that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of the AI-PROFICIENT consortium and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 28 Member States of the Union. It is based on the European Communities and the Member States cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors (http://europa.eu/).

AI-PROFICIENT has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957391.

Title: D6.4 AI-PROFICIENT ethical recommendations

Lead Beneficiary	UL
Due Date	31/10/2023
Submission Date	30/10/2023
Status	Final
Description	Instantiation of HLEG guidelines and ethical recommendations
Authors	Marc Anderson (UL), Karën Fort (UL)
Type	Report
Review Status	¨ ~~Draft~~ ¨ ~~Partners Accepted~~ ¨ ~~WP Leader Accepted~~ ¨ EEA Accepted
Action Requested	¨ ~~To be revised by partners~~ ¨ ~~For approval by the WP leader~~ ¨ ~~For approval by EEA~~ ¨ ~~For approval by PMT member~~ ¨ For acknowledgement by partners

VERSION	ACTION	OWNER	DATE
0.1	Adding text in template (MA)	KF and MA	18/07/2023
0.2	First draft	KF and MA	29/08/2023
0.3	Considering UL and partner feedback	KF and MA	20/09/2023
0.4	Validated by WP Leader	KF and MA	5/10/2023
0.5	Considering EEA feedback	KF and MA	25/10/2023
0.6	Validated by PMT member	KF and MA	27/10/2023

1.0	Final Version	KF and MA	30/10/2023

Executive Summary

The Deliverable D6.4 is a public document of AI-PROFICIENT project delivered in the context of WP6, (Use case evaluation and ethical considerations), with regard to the Task 6.4: Instantiation of HLEG guidelines and ethical recommendations. It serves as a description of the ethical component of the AI-PROFICIENT project and an evaluation of the success of the ethical approach in the project, in integration with the more general evaluation of the project.

Because it is a public document, project partners are not named specifically except in the general discussion of plant visits in section 2.1.6. In addition, all data are anonymized, including Use Case code, pilot plant, and task leader, with regard to examples of ethical recommendations, and implementing partner(s) are anonymized with regard to ethical recommendation implementation results. The anonymized data regarding ethical recommendation implementation results will be published separately in a somewhat modified format, in a peer reviewed journal article.

The report contains a detailed overview and discussion of the European Commission High-Level Expert Group Ethics Guidelines for Trustworthy AI both in general and in relation to the particular difficulties and issues regarding ethical AI encountered in the industrial manufacturing context in, and particularly in the AI-PROFICIENT context. It describes modifications and adjustments made to the Ethics Guidelines for Trustworthy AI to better fit them to the realities of the AI-PROFICIENT context.

A review of the ethical methods adopted in the project is made, accompanied by a description of ethical recommendations given, recommendation categorization, and research and dissemination results. This includes descriptions of meetings, plant visits, deliverable reviews, and generation and monitoring of ethical recommendations in context.

Anonymized examples of ethical recommendations are provided, along with the approach to recommendation categorization. Other completed ethical contributions to the project are also surveyed: deliverable contributions, peer reviewed publications and conferences and outreach, in order to give a sense of the scope of AI-PROFICIENT's ethical component.

The heart of the ethical validation is then presented through the quantitative implementation results of the ethical recommendations, as assessed by the ethics team and separately by the project partners. The ethics team was composed of Marc Anderson and Karën Fort, principally, with some additional input from Christophe Cerisara and other members of the UL partner. In the project partner side of the assessment, the Use Case leaders, main Industrial partner representatives, and the leader of WP6 participated.

The quantitative implementation results include results by category, by category and partner (anonymized), and overall results. A discussion of observations regarding implementation, which aims to be useful to those who might adopt the applied ethical method of AI-PROFICIENT, accompanies the results.

AI-PROFICIENT implementation results are further integrated into the WP6 validation through a conversion of ethical implementation results into general validation results, to be used in the other WP6 deliverables.

Finally, a comparison of AI-PROFICIENT approach with the approaches of other projects in the ICT-38 cluster is carried out, high-lighting complementarities and potentials for integration within the cluster. To this are added a discussion of limitations of the AI-PROFICIENT approach and suggestions for future related research.

Introduction

One of the biggest problems in AI ethics is the lack of paths to operationalization. As (Morley et al, 2021) have argued, there is "a significant gap … between theory and practice within the AI ethics field." (Prem, 2023) has summarized some of the approaches taken to close this gap. Nearly all of those approaches are very general in character however, or they content themselves with technical adjustment primarily, disconnecting users from software developers and the technology development companies within which they work.

In AI-PROFICIENT project we have pursued a different method for operationalizing AI-Ethics. The method was described in AI-PROFICIENT Deliverable 1.2, and subsequently published in (Anderson and Fort, 2022). To our knowledge there are no similar descriptions of methods developed to operationalize AI ethics at ground level in industry. The nearest approach we know of is (Berrah et al., 2021) who have concentrated on developing a system for ethics evaluation in Industry 4.0 as one dimension of performance evaluation, although they admit that attaining quantifiable performance indicators for ethics evaluation is difficult.

As (Nafus, 2018) argues, the development of AI systems is often about automation first, with the exploration of the human contextual element left as a minor aspect, if considered at all. In other words, the goal of automation decides what follows. If so, as we argue, it is liable to decide what follows with regard to ethics as well. This, in part at least, may be the reason that so many AI ethics frameworks have sprung up. Such frameworks are – consciously or unconsciously – various attempts to frame ethical engagement as an automated act, where one simply goes through the list of possible ethical problems or 'risks,' mechanically.

The AI-PROFICIENT ethics team's approach has been very context based, not merely at the level of heavy industry, but at the shop floor level. Our view is that the context should decide the kinds of questions asked or omitted. Our response to the automation of AI ethics has thus been to attempt a more un-regimented, or 'freewheeling' method, where the ethics team members explore the context first, discussing that context and related issues with both industrial partners and tech company individuals, reflect upon this exploratory data which has been uncovered, and reason out a tailored actionable solution. From there we have given recommendations and followed the progress of their implementation.

This process has given us results in the form of responses to the recommendations. From those results we can begin to understand what types of recommendations have worked, or have not worked, and how much the character, interests, and motives of the industrial partners, tech company partners, and individual software engineers impact implementation.

This is a flexible method, and an interesting method, but also a time consuming one. It differs from the method envisioned by the European Commission High Level Expert Group *Ethics Guidelines for Trustworthy AI*,[1] which implicitly advocate 'automating' the ethical engagement – e.g., through the repeated use of the Altai tool – and in fact leaving it in the hands of the AI developers[2] themselves. We do not discard the HLEG guidelines, however. We look to them as a rough very high-level guide map which we can look up to as needed, but from the ground of the factory floor conditions.

The current deliverable thus begins in Part 1: HLEG Guidelines for Trustworthy AI, with a review of the HLEG guidelines and a discussion about their strengths and weakness with regard to our project and heavy industry in general. We then describe how we have departed from the method of the guidelines and why.

In Part 2: Review of Methods Used, Recommendations, Recommendation Categories, and Research and Dissemination Results, used in AI-Proficient, we review our approach again in outline, which is

[1] Hereafter HLEG. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[2] The HLEG guidelines define developers in the context of AI as: "those who research, design and/or develop AI systems." (HLEG, 14)

given in early but much greater detail in AI-PROFICIENT Deliverable 1.2. We give examples of the recommendations, and describe how we categorized them, again from the ground up.

The core goal of Deliverable 6.4 was evaluation, in conjunction with the more technical evaluation of other Deliverables of AI-PROFICIENT WP6 and thus Part 3: Ethical Recommendations Review offers a systematic review of the Implementation results of the recommendations, in a number of graphs which show the data overall, by category, and broken down by category and by partner. Observations and discussion of the results follows. Part 3 ends with a description of how the ethics team has converted the implementation results according to the evaluation methodology agreed upon with other AI-PROFICIENT partners, to be included in the general evaluation of AI-PROFICIENT Deliverable 6.2

Part 4: Insights for Future projects gathers together Insights from the project for future researchers in the EU context for AI ethics in Industry 4.0 and heavy industry generally. We begin with a review and comparison of ethics approaches in other projects of the ICT-38 cluster in which we are included, trying to show how our method complements the other approaches while offering its own unique contribution. We then detail some insights gleaned from our particular applied ethics method, review the limitations of our method, and finally offer some suggestions for lines of future research.

Part 1: HLEG Guidelines for Trustworthy AI

1.1 Overview of HLEG Guidelines

The High-Level Expert Group Guidelines for Trustworthy AI (HLEG), were developed in 2018, with a final draft made available in 2019. To create the Guidelines, the European Commission brought together a group of 52 experts in the domain of artificial intelligence from among university researchers, NGOs, and industry leaders. A first draft was made public in 2018, and feedback was gathered from public consultation upon the document. It was then modified and released to the public again in 2019.

1.1.1 Outline of the HLEG Guidelines

The stated goal of the HLEG is to promote trustworthy AI. Trustworthy AI is defined as having three components, namely, it is *lawful*, *ethical*, and *robust*. Although lawfulness is not addressed directly, the component of lawfulness in its relation to Human Rights informs the development of the document. Subject to that condition the three components are developed sequentially.

Beginning with a general introduction which expands its three-component definition of trustworthiness, the HLEG clarifies how the three components should work together, how they are necessary and yet not sufficient for trustworthy AI and notes how potential tensions might arise between components. From there it describes the scope and audience for the guidelines.

It then goes on to describe some of the legal and regulatory apparatus already applying to AI in the EU, and adds that, while legal considerations are sometimes reflected in ethical and robust AI considerations, the latter often go beyond legal obligations. It also provides a disclaimer that the document is not meant to be taken as legal guidance. Ethical principles are then characterized as necessary because laws do not align at times with ethical norms or may be unsuited to address some issues. Finally, robust AI is characterized as necessary, technically and socially since AI systems may cause unintentional harm even though their purpose may be ethically sound.

The HLEG lays down a framework in three chapters, which build upon one another, from the more abstract to the more concrete, to achieve a beginning of trustworthiness in AI: **Foundations of Trustworthy AI**, **Realising Trustworthy AI**, and **Assessing Trustworthy AI**. This is a very top-down approach, contrary to our approach in AI-PROFICIENT project.

In **Foundations of Trustworthy AI**, the expert group aims to provide a part of the normative vision of the European AI future, to include democracy, human rights, and law. For the expert group, the best way to move from law to ethics is to move from the fundamental rights laid down in the Charter of Fundamental Rights of the EU, and the Articles 2 and 3 of the Treaty on European Union. (HLEG, 9) The expert group suggests that respecting human dignity – a 'human-centric approach' – is the commonality of these rights. Since the fundamental rights of individuals are viewed as stemming solely from their being human beings, with the moral status which that entails, and at the same time they are taken to be legally enforceable, then they can provide a link between lawful AI and ethical AI.

There are four families of fundamental rights in particular to be drawn from EU law, namely the above-mentioned Charter of Fundamental Rights of the EU,[3] and Articles 2 and 3 of the Treaty on European Union,[4] and related to prospective AI use. *Respect for human dignity*, i.e., the human worth of each individual, demands that humans not be treated as objects by AI, e.g., in being sorted or manipulated. *Freedom of the Individual*, the freedom of each to make life decisions, demands that AI not be used for unjustified surveillance, manipulation, or coercion. *Respect for democracy*, justice, and the rule of law, demands that AI foster rather than undermine democratic processes. Meanwhile, *Equality, non-*

[3] https://www.europarl.europa.eu/charter/pdf/text_en.pdf
[4] https://eur-lex.europa.eu/resource.html?uri=cellar:2bf140bf-a3f8-4ab2-b506-fd71826e6da6.0023.02/DOC_1&format=PDF

discrimination and solidarity require that AI be developed so as to avoid unbiased outputs. Finally, *Citizens' rights*, demand that AI be used, e.g., to improve government services when possible.

Upon these families of fundamental rights, the expert group advances four ethical principles, whose ethical instantiation can in many cases be carried beyond their pre-existing legal instantiation. These are the principles of: *respect for human autonomy*, *prevention of harm*, *fairness*, and *explicability*, each of which is intended to be a broader, more abstract, ethical formulation of the fundamental rights issues noted above.

The HLEG then goes on to highlight the possibility of tensions between the ethics principles advanced, exemplifying these tensions with the case of AI use for predictive policing, which they conclude may give rise to benefits in crime reduction (the principle of preventing harm), but on the other hand may curtail individual liberty (the principle of human autonomy). Theses tensions can only be resolved, if they can, through sustained reasoned evidence-based reflection.

In the second chapter, **Realising Trustworthy AI**, the HLEG takes the first step in its approach of moving from the abstract to the concrete. It defines developers, deployers, and stakeholders, and outlines their roles. It then proceeds to list seven requirements, which are an expansion upon the earlier four abstract principles and under each of which is gathered systemic, individual, or societal aspects which have particular relevance to AI both positively and negatively. The seven requirements are to be equally important and applied throughout the complete AI lifecycle.

Human agency and oversight should check that fundamental rights are being upheld in terms of data tracking and AI uses which support accessible education. It should facilitate human agency in order to allow informed decisions regarding AI use, and it should provide for human oversight mechanisms so that humans remain in control of AI systems.

The *Technical robustness and safety* requirement should mitigate potential harms of AI systems. It should build resilience to attack in the systems, considering malicious uses and dual uses. Safeguards, such as pausing for human intervention should be built in and the AI system should be accurate and this proportional to its effect on human lives. Reliability and reproducibility help test and improve the system to prevent unintentional harm.

Privacy and data governance require guarantees of privacy and data protection so that users can trust the systems. It includes checking quality and integrity of data to address dataset biases in advance and prevention of malicious data input to AI systems. Data should be accessible only according to well defined protocols.

Transparency is to be achieved through thorough documentation which allows for traceability of AI decisions and errors. Explainability, i.e., the ability to explain a system's technical processes and human decisions in a timely manner adapted to the individual user, makes up a further component. Finally, the fact of dealing with an AI, including its relevant limitations, should be communicated to users.

Diversity, non-discrimination and fairness are ensured by addressing unfair bias in data sets through oversight principles which examine the system's purposes and decisions transparently, diversity hiring, and directly removing biases in the data collection phase. The foregoing is complemented by designing user-centric, and universally accessible systems for equal access, and by incorporating mechanisms for long term participation of affected people.

Societal and environmental well-being should be sought through assessment of the AI system's environmental impact (resource and energy use), assessment of social relationships and mental and physical well-being, and effect upon electoral and democratic processes.

Accountability demands that AI systems be auditable through internal and external (and independent) auditor evaluations, that protection be given to those reporting legitimate concerns, and that tradeoffs be allowed for and evaluated in terms of ethical risks, with decisions documented and decision makers kept accountable. Most importantly perhaps, modes of redress should be available for adverse impacts.

The seven requirements need implementation and thus the guidelines go on to suggest some general technical and non-technical methods for realising them. Importantly – and we shall return to this later – the HLEG guidelines view the AI development as a continuous dynamic and evolving process.

Under *technical methods*, are included: translating the decided upon requirements of the AI system into its architecture so that it does or never does certain things (including self-learning AIs), adopting ethics by design approaches which develop AI systems to comply with norms from the beginning (e.g. secure, robust, and having fail-safe shutdown), employing various methods of Explainability, early and ongoing testing and validation of all components of AI systems under multiple metrics and with diverse testers, and instituting a variety of quality of service indicators for such metrics as functionality and performance but also for algorithm training.

Under *non-technical methods*, The HLEG guidelines encourage: specific regulation to support trustworthiness, adapting of corporate codes of conduct to the HLEG guidelines, drawing upon pre-existing standardization practices to enhance quality management of AI systems, potential certification by appropriate organisations, governance frameworks such as ethical officers or boards who can provide oversight and advice, educating involved parties – including the public – as to their potential roles in shaping AI technology, actively promoting social dialogue and open discussion on AI system use and impact, and fostering diversity in AI system design teams.

The HLEG guidelines third chapter, **Assessing Trustworthy AI,** provides a pilot version of an assessment list for operationalizing Trustworthy AI. The list contains general questions to ask relative to each of the requirements of chapter two. The group notes that the list is neither exhaustive, nor intended to guide legal compliance, and also that it needs to be tailored to specific use cases. Incorporating the assessment list into a governance structure is discussed, with an emphasis on high level management. To clarify the foregoing further, the ways in which different levels of management might act relative to the assessment list are outlined. It is suggested that use of the list be incorporated into existing practices of AI practitioners. The list was later modified and expanded after a period of public consultation and released as *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment* in 2020 and also adapted into a web-based tool: ALTAI.

1.1.2 General Discussion of the HLEG Guidelines

We can note a number of issues regarding the content and the directions of construction of the HLEG. This is not meant to be critical, but rather to outline the basis upon which the specifics of our approach, an approach of complementing the HLEG, can be better understood.

The Guidelines take a first ethical stance in assuming that AI systems are or can be generally beneficial. They do not countenance a situation where AI systems might be more harmful than otherwise. Examples of this stance include: “the aim of the Guidelines is to promote Trustworthy AI” (HLEG, 2), “AI is … a promising means to increase human flourishing” (HLEG, 3), “It is through Trustworthy AI that we, as European citizens, will seek to reap [the benefits of cutting-edge and ethical technology]” (HLEG, 3).

Thus, the opening position of the group is not neutral. Whether this is an ethically questionable stance generally we leave aside here. The point however is that this opening stance tends to render more difficult and controversial those types of guidance or recommendations, which, through a reflective assessment of the particular context of a Use Case (hereafter UC), find that on balance the use of AI in a particular situation is unwarranted or unwise ethically speaking. In other words, recommendations *not to use AI* in particular cases becomes more difficult after beginning from the positive stance toward AI use which the guidelines embody on the abstract level.

Following on this, the paradigm of trustworthiness itself could be questioned. The Guidelines suggest that trustworthiness is the lynchpin of ethical AI use and that their goal is to “ensure and scale Trustworthy AI.” Without trustworthy AI systems, unwanted consequences may occur, as the working group notes. The intent is laudable but yet may give rise to a false sense of security. Even if AI system development both *has access to* and *implements* the best possible ethical principles, the unwanted

consequences are likely to occur. To see this, one can consider the modern development of the airplane. Airplanes are trustworthy systems by all our usual standards. And yet the global average of airplane crashes of all types, while slowly declining, has averaged about 2000 per year for the past 40 years, with a number of large jetliner crashes being among these each year, and an average of more than 1000 fatalities per year. Thus, although airplanes are indeed trustworthy systems, the development of this trustworthiness, and even the perception of it, is a gradual process, with many setbacks along the way.

In other words, the aim of trustworthy AI in a strong sense – a sense very much open to being hyped and marketed – may not be operationalizable. Such an aim may have to give way to a more pragmatic aim: develop the most ethical AI systems that we can at any moment. The latter outlook leaves more room for specific operationalizations, i.e., 'we have done what we can to make this or that system ethically sound,' an approach which may fall short of the strong sense of trustworthiness noted but may nonetheless move forward practically in the development of AI ethics, a point which we will relate to our approach later.

The emphasis on robustness is also framed in a particular way in the guidelines, so that trustworthiness implies robustness. But robustness is implicitly defined around the premise that: "AI systems will not cause any unintentional harm." This leaves open a wide scope for developing AI systems which are robust but whose purpose is *intentional* harm, including but not limited to military applications.

In effect there is an uncertainty in the Guidelines, due to their being uneasily linked to law, as Human Rights, while at the same time being distinctly separated operationally from law and regulation.[5] The urge to link ethics to law plays out in the vision of ethics on offer. Ethical, for the Guidelines, involves "alignment with ethical norms," (HLEG, 7) essentially a deontological stance. But the pragmatic sources of this deontological stance are left unmentioned. Unless merely logically developed, norms are developed through a process of engagement with experience. In fact, even *if* logically developed, norms are sourced in the desire to bring individual and social human experience together consistently, and they need working out in that arena to have meaning and effect.

In a similar way, ethical AI development will arguably have to pass through a stage of trial and error, where we try some things, fail to be consistent in the individual and social arena, have setbacks and 'unwanted consequences', re-assess, and then try again. It may be that some AI developments cannot be made ethical, i.e., an ethical 'skin' cannot be laid over them. When, in the context of human dignity, the guidelines suggest – a very Kantian stance – that all people be: "treated with respect due to them as moral *subjects*, rather than merely as *objects* to be sifted, sorted, scored, herded, conditioned, or manipulated," (HLEG, 10) we should go on to ask whether perhaps some types of AI are, or can finally only be, *nothing but* systems which sift, sort, score, etc. If so, a clear ethical opening must be left for not using AI as well.

Finally, elements in the notion of the requirements of trustworthy AI indicate a lacuna in the approach of the guidelines. "While most requirements apply to all AI systems, special attention is given to those directly or indirectly affecting individuals. Therefore, for some applications (for instance in industrial settings), they may be of lesser relevance" (HLEG, 15). This is particularly relevant to AI-PROFICIENT. What is missing here, is the insight that industrial settings are deeply bound up with the human activity of work and of the creative products of that work. Once these insights are remembered, it becomes clear that ethics is embedded in both products and the machinery that humans use to create them, and further that AI uses in industry may very well have ethical effects which are not easily located in the direct relations of human to human or human to machine, but which show themselves at different scales through work and its products. A view biased toward mainly social uses of AI – in the unreflective sense of social uses – is ill suited to engage these effects.

[5] There are at least three disclaimers in the guidelines, which state that they do not provide guidance with regard to legal compliance with existing AI regulations in any way.

1.2 HLEG Guidelines in Relation to AI-PROFICIENT

The following is an overview of the relation of the HLEG guidelines in relation to AI-PROFICIENT project. It aims to show how the ethics team both followed the spirit of the HLEG guidelines and adapted them insofar as judging them to be unsuitable for AI-PROFICIENT.

1.2.1 Original Aims of the AI-PROFICIENT Project Proposal with regard to HLEG

The original positioning of the AI-PROFICIENT project was laid out in the project proposal. In general, the intention was to draw upon the HLEG guidelines to add an ethics by design approach to the project, solve potential ethical issues, and for this to complement a legal consideration of issues which might run across legal and ethical boundaries.

Particular tasks were envisioned as being developed in conjunction with an emphasis on a user centered approach sourced in some of the principles laid out in the guidelines. Among these were tasks 1.3, 1.5 and the tasks of WP6, directly, as well as task 5.3 (data privacy) and the tasks of WP4 (HMIs and explainable and transparent AI) indirectly.

As part of WP6 more specifically the aim was *to adapt the HLEG guidelines to the manufacturing domain while providing recommendations based on that adaptation*. (AI-PROFICIENT Project Proposal – Annex 1, 32). Ethical human-machine collaboration in conjunction with trustworthy AI, and with an eye to the Human-in-the-Loop paradigm, was a main component of this adaptation (AI-PROFICIENT Proposal, Part B – 5; Idem. Part B – 7). From there the results of this adaptation were to be carried into the final task of WP6 and integrated with the more technical result validations of the project, in order to guide future projects and European industry in relation to AI development for industry.

Instantiation of AI-PROFICIENT guidelines in terms of *recommendations* was implicitly envisioned as being mostly after the fact. Task 1.2 of WP1, which resulted in Deliverable 1.2, was scheduled to begin immediately and be completed by month 6. The latter was to be an analysis of ethical – and legal issues – regarding human machine interaction, data exploitation, and explainable and transparent AI. Meanwhile Task 1.5, which resulted in Deliverable 1.5, was to run between month 7 to month 12, and proposed following the HLEG guidelines through adopting a user centered approach. But WP6 Task 6.4 was scheduled to begin at month 23 and run to the end of the project in month 36, and only here were recommendations mentioned. In terms of dissemination of project results. It was also stated, with regard to WP7, that “finally, UI will provide final ethical recommendations and instantiate HLEG guidelines to manufacturing domain.” (AI-PROFICIENT Project Proposal, Part B – 66)

1.2.2 Aspects of the AI-PROFICIENT Project which the HLEG are ill suited to address

AI-PROFICIENT project was a research and development project in conjunction with real industrial conditions. The ethics team did not know enough of the shop floor context at the beginning to apply HLEG principles so that they could lead the scenarios of the project to an ethical outcome. Thus, it was determined that a promising approach was to let the ethical issues disclose themselves out of a familiarization with the shop floor context in conjunction with an embedded participation in proposed AI service design meetings.

So rather than imposing a framework blindly – that of the HLEG principles – upon a context that we did not understand and would not understand for some months we simply observed and participated as much as possible while letting the ethical issues present themselves. The idea was that if the issues disclosed could later be gathered under the HLEG guidelines, then they would be. But they should not be cropped to fit under the latter and if they need to be then the guidelines are insufficient and need supplementing.

From this beginning, the ethical recommendations grew naturally out of observing the work context directly or as described by the industrial partner engineers and participating in the technical meetings that worked to develop the AI services desired for them. Because covid regulations were in place in the beginning, the industrial partner engineers gave us and other project partners virtual opening tours of the plants, answering our initial questions. There then followed a period of meetings, in several stages, over several months specifically to understand the context of each Use Case, the problems which the industrial partners were interested in solving, and the interests and capacities of the developer partners in solving them. We took full advantage to ask further questions and fill out our picture of the Use Case contexts relative to potential ethical issues. At about the sixth month the ethics team began to make initial recommendations which were subsequently updated or added to as the AI service solutions evolved. It should be clear that this approach is the reverse of that envisioned in the HLEG guidelines and similar frameworks, whose tendency is to begin at the top and go through a systematic process of checking for ethical issues which might fall under the various principles.

Not only is such an approach tedious, but unless it is done over and over, it will tend to miss ethical issues which arise unforeseen and to lose sight of the process of development. Accordingly, the ethics team view is that developing the corrective recommendation to the arising ethical issue is best done by experiencing and following the issue as it arises in the process of development. Contrary to our view, the HLEG guidelines approach pushes the notion of the design being built upon the ideal, whereas the reality is that the work and tech development contexts from which one begins will inevitably dictate what can be achieved. Just as road maps, or now a GPS system, are most wisely consulted as a rough guide in driving but cannot substitute for keeping one's eyes on the actual road and surroundings, so the HLEG principles can be useful as a rough guide in various ways but are not suited to serve to as a beginning in engaging ethical issues in a project such as ours.

The outline of the above general considerations is essentially clear only after the fact, however. As with our recommendations, they arise out of the particularities encountered in the project. The three following points describe in more detail those particularities.

1) There was no stress laid upon particular ethical principles in the project proposal. The stated aim was that ethical principles guiding the project should *accord* with HLEG guidelines (AI-PROFICIENT Proposal, Part B – 4). The ethics team interpretation of this was that the guidelines should serve as a high-level check upon whatever ethical principles were applied, i.e., that the less general principles should at least accord with the spirit of the HLEG guidelines, rather than running contrary to them.

The nature of digital technology development is that it proceeds at a fast pace, driven by external interests, and in later phases it builds upon choices made earlier in the design cycle. Many of these initial choices have ethical implications. AI-PROFICIENT project largely followed this paradigm. On the other hand, the process of ethical analysis and application that would be able to instantiate the HLEG guidelines according to the stated goal, needs time and a good knowledge of the context. The instantiation of ethical principles to a new technology and its relations to humans is a matter of reflective morality for those developing and guiding the new technology in particular, and only secondarily an issue of custom – such as might be found in business or computer ethics codes which are essentially heuristics developed through the history of smoothing interactions between members of the profession. The HLEG guidelines are thus a cousin to older ethics codes in terms of their structure, but insofar they are also limited in their capacity or application to new and unconsidered technology and its applications.

In other words, the context of the AI-PROFICIENT project, as many such projects, is that you don't know what situations will arise, because the technology being developed is relatively new, it is being put to new uses, and it is pushing the worker to adopt new relations to the technology and to colleagues. Yet, as the philosopher John Dewey noted, the analysis of reflective morality, as opposed to a more custom driven morality, demands that we clarify the moral situation first (Dewey and Tufts 1932). Thus, unless the ethics team was to limit itself to analyzing what has gone wrong *after the fact* in the project, it would first need to clarify the moral situations in the wide variety of work contexts related to the project.

What makes the moral situations in AI-PROFICIENT different from the usual situations discussed in more theoretical ethics, is that the industrial context presents them very directly to us. The latter are much less abstract. The potential ethical issues are the degradation of the relations of the worker to

their fellow worker or to the process managers, or the degradation of their own work experience, through the introduction of the AI service or subsidiary changes related to it in the work context. The HLEG was developed to deal primarily with social AI uses, i.e. uses where the consequences of the AI developers design solutions are far removed from the developer's act, either because the AI use is affecting a large number of people – beyond the scale of normal human physical relations – or because the AI use is not grounded in a very spatial and temporal environment such as the industrial work context with its machines.[6] And to be fair, the creators of the guidelines acknowledge this with regard to the requirements: "While most requirements apply to all AI systems, special attention is given to those directly or indirectly affecting individuals. Therefore, for some applications (for instance industrial settings), [the guidelines] may be of lesser relevance" (HLEG 15). Nevertheless, our opinion is that *in industrial settings the affects upon individuals become very clear in ways they may not for more obviously social AI uses*, and thus changes to the suggested approach have to be made. The factory context is much closer and more physical for the developer than in social AI uses and the developers have to take it into account in designing the solution.

2) As noted earlier, in its third chapter, the HLEG guidelines themselves lean upon an assessment list of questions to be asked to AI developers. Including the fundamental rights impact assessment, the assessment list contains 152 main questions. Many of the questions – e.g., the question regarding discrimination – both have multiple aspects and are non-exhaustive. Moreover, even though typically very general, the questions apply to different levels of generality, thus requiring additional reflective judgment in application.

This engagement as a series of questions to the developers puts the weight upon the developer to select what seems important in a given context as well as expecting the developer to have a well worked out sense for judging the importance and generality of the different questions, including which questions might have *no relevance* to a given context. Moreover, what is important to the developer – through no fault of the latter – is likely to be what the developer is used to dealing with professionally. The notion that technology can and should always find a solution is much ingrained in tech developers and works against ethical engagement, as Avnoon et al. (2023) and Clark and Lischer-Katz (2023) have argued. Thus, for example, questions which can immediately be related to a technical solution are arguably more likely to be seriously considered by the developer in a project such as AI-PROFICIENT, than obviously human related questions which do not present themselves as often in the everyday technical work of the developer.

Beyond a recommendation that the assessment questions be integrated with existing principles of AI practitioners, there is no indication in the guidelines that the above difficulties relative to operationalization have been clearly considered in the HLEG; such considerations are left to the ethics and AI practitioners. In the best-case scenario, existing principles are likely to be few or structured relative to practices for internal professional collaboration. (Skaug Saetra and Danaher, 2022) note the dizzying profusion of terms for seemingly different domains of technology ethics, a situation which, without guidance, could easily lead to the AI developers of our project falling back upon minimal internal ethics codes – like professional codes of conduct – entirely insufficient to the specifics of adding AI to industrial shop floor situations. Also, AI practitioners may simply not have any sufficiently developed existing ethical practices to integrate with. If, on the other hand, the project partners do have such codes or principles, then the question of whose codes or principles to integrate with becomes a further issue, given that multiple partners are contributing to the project.

If the assessment list approach were carried out sincerely and directly by the developers the whole list might have to be gone through for each use case, multiple times, which is difficult. Arguably, the use of the checklist approach – the HLEG guidelines do not endorse this approach but the result comes very close practically – is tailored toward a certain way of systematic thinking – toward automatization – which appeals to tech development practitioners. But in promoting that approach, the human centeredness is lost for others who are not trained in such thinking. Also, practically, there is no time even for the checklist approach from the point of view of the tech developers.

[6] In fact, the hardware and processes supporting the AI are very much grounded in space and time as (Crawford, 2021) has shown, among others, but unfortunately this is often forgotten – sometimes deliberately.

Meanwhile, a selective use of the list presupposes a level of familiarity with potential ethical issues which belies the originality of the research being undertaken, but also renders the list itself dispensable: if you know what to look for already you don't need the list.

3) End user engagement through the project was a stated goal (AI-PROFICIENT Proposal, B – 4). But the practicalities of this goal were not indicated and could only be indicated with difficulty by proceeding 'downward' from the HLEG guidelines themselves. The industrial work context of a project such as AI-PROFICIENT, has particular difficulties in this regard. First, the workers are engaged under contract and thus may not be fully free – legally free – to accept or reject certain actions which might be deemed necessary by the employer, but which have ethical implications. In other words, law can easily be made to oppose ethics in the work context, because the moral situation involves voluntary activity (Dewey and Tufts, 1932). Second, implicit workplace customs and explicit corporate and union legalities, make a knowledge of actual end user engagement at the level of generality of the HLEG difficult. Without that knowledge it would be difficult to clarify the moral context in terms of validating the implementation of the very general principles of the HLEG regarding user engagement.

For example, the HLEG assessment list, under Human Agency and Oversight, asks the question: "Could the AI system affect human autonomy by interfering with the end-user's decision-making process in any other unintended and undesirable way?" (HLEG Assessment List, 7). The latter question cannot practically be answered in pursuit of ethical implementation (instantiation), until a number of far more specific questions are asked and answered to some reasonable degree: who is ultimately responsible within this operator team, how often does the relevant process manager officially give instructions, are there informal workplace traditions which are taken for granted with regard to responsibility, etc. The addition of the AI may now indeed add an un-ethical element of uncertainty as to the worker's responsibility in terms of some aspect of the manufacturing process, but we cannot get at it unless we get some knowledge based on these more specific questions. We may not be able to access it at all, but insofar as we can access it, generalities will not suffice.

1.2.3 Modifications Made based on the Above Difficulties

1) If the factory context is much closer and more physical for the developer than in social AI uses and the developers have to take it into account in designing the solution, then we can engage this situation best if we tailor our ethical approach to match it. As Widder and Nafus have shown (2022), the nearer tech developers are to encountering the end user and experiencing the responses of the latter to the technology, the more likely they are to consider the ethical dimension of their work.

The appropriate modification to the HLEG approach here then, is to begin by gathering detailed knowledge about the original work context in its physical and temporal aspects. Having gathered that knowledge, we then go on to clarify what is intended in the solution, again putting stress on its physical and temporal aspects. This means gaining at least a layperson's understanding of the design and development of the services at this very physical level. In each case the important questions which engage the moral context are: what is the operator doing already, where and how are they doing it, what is the issue – from the operator perspective – that project partners aim to solve with some use of AI and other technologies, what leeway is there in the design of the solution, what is the goal of the solution with respect to the operator (even if it that goal may be secondary).

At the same time, we modify the HLEG approach from something tending toward 'one and done,' to something more continuous. This means keeping the ethical dimension in the foreground for the developers and industrial partners and keeping it fluid and evolving, i.e., giving it the character of a process. Instead of coming in as 'experts' to declare 'we've checked this and this aspect so we needn't return to it,' it means following the development directly and continually and communicating to the developers that the ethics team is collaborating in that continuous mode.

2) Here the modification in our approach was not to leave it to the developers to explicitly ask ethical questions of their own practice. Indeed, it was not to ask questions *at all* in a systematic way with direct regard to potential ethical issues, e.g., 'have you checked that personal data is being secured?'

Instead, we used questions to gather the details of the context of each use case. We then continued asking questions during regular participation in the technical meetings in which the partners met to design the solutions. The ethical issues often presented themselves quite naturally through reflection or discussion on the base of contextual knowledge for each Use Case, sometimes in relation to the vague ideals of the highest requirements laid down by the HLEG guidelines, and sometimes straying outside of those ideals. When the ethics team did not get clear answers easily, we went on to make formal ethical recommendations that certain points within each Use Case[7] be formally clarified, e.g., to formally state who was ultimately responsible in some situation, or to give quantitative estimates of extra workload for AI feedback training or service reliability.

Another modification is that we also do not limit ourselves to directly AI related aspects of the contexts *at the lowest level*. There cannot be a good separation of the AI ethical aspect of the context and other potential ethical aspects, e.g., the physical state of the worker. But since these indirect aspects are not usually envisioned as part of the technical solution they would not be addressed unless we make a specific effort to bring them to the fore. Thus, the ethics team supplemented the approach of the HLEG by making recommendations such as: check the cohesion of the work team after service deployment and do a colour blindness test for the operator users (since some app colour choices proposed using a red and green schema).

3) A modification to the HLEG guidelines approach here was to survey the legal landscape around AI, note what actual legal scholars have said on the issues – since we do not claim competence in the legal area – and distill some general practical advice with regard to ways in which the project partners could participate earnestly in the legal approach to the issues. The ethics team also approached issues related to current laws or regulation, but which have an ethical aspect, by noting where and how ethics agrees with the spirit of current regulation, e.g., GDPR, in a particular context, and then giving ethical recommendations that would, if followed, implement the spirit of the laws.

A second modification was to keep recommendations very specific, to the point where their implementation, or rejection was readily apparent in developer meetings and later evaluation. When the recommendations are specific the responsibility for ethical development moves out of the realm of theory and into the practical level. The recommendations could still be rejected, but the approach tends toward clarifying the boundaries between legal obligation and voluntary ethical progress. It also tends to clarify the reason for the rejection explicitly or else it highlights the oppositions of ethical and non-ethical interests[8] which will eventually have to be overcome.

A third modification, overlapping with that of 2) above, was to make ethical recommendations specifically toward uncovering the context of a situation. If, as we suggest, developing ethics requires knowing the particulars of the situation in question, then logically, uncovering those particulars itself becomes a subsidiary ethical action included with the more general process. So here, as noted above, the ethics team made recommendations such as: state who is ultimately responsible within the operator team, state whether the process engineer or the team captain has responsibility over actions at the factory floor, etc. These are ethical questions because they feed into ethical outcomes at higher levels, but they are a mode of questioning that the HLEG guidelines and similar guidelines do not typically consider. The guidelines ask variations of: ‘have you been ethical in terms of x?’ Our approach is ‘tell us what is happening already or what you are planning here and here,’ and if it’s not ethical we’ll suggest why it isn’t and recommend an alternative.

Modifications With Regard to the Issue of Sustainability:

We must say something about sustainability here. As noted above, one of the seven requirements for realizing trustworthy AI in the HLEG guidelines, is *Societal and Environmental Well-being*, under which the resources and energy used by AI systems, and in general the environmental impact, should be considered. With respect to this requirement, early in the project the ethics team decided against pursuing the assessment of this requirement in a rigorous manner. Our reasoning was as follows. Given that the high-level expert group itself suggested that the guidelines were not ideal for

[7] Hereafter UC.
[8] E.g., the “corporate logics and incentives,” noted by (Green, 2021).

the industrial context, there is obviously no clear guidance in the HLEG guidelines on an environmental and sustainability assessment. A survey of research on the notion of environmental impact and sustainability in terms of AI, shows that these are very difficult terms to pin down indeed, particularly because AI systems depend upon very far-reaching processes – e.g., mining in developing countries for the resources with which the hardware for AI systems are built – which are almost never countenanced (Crawford 2021). To keep in line with our approach aiming at a quantitative assessment of ethics operationalization and ethics by design from the bottom up, we would need measurements and standards which are simply not there yet. The only hard data we know of is that of (Strubell et al. 2020), but this is for large language models, and does not fit our industrial context, which is using small data sets as noted in (Fernandez et al. 2023). In our context the sustainability and environmental impact of the AI solutions will be very much tied up with the environmental impact and sustainability of the existing plant machinery and the manufacturing processes with which those solutions will be integrated.

In other words, AI does not exist and is not developed, in a vacuum. It is already developed on the basis of practices which should have been questioned earlier – before the advent of AI – in terms of sustainability, i.e., manufacturing and machine practices generally, and all the more so in terms of AI-PROFICIENT. Accordingly, an assessment of environmental impact and sustainability from the bottom up, would need to first question existing practices in our manufacturing partners, before AI integration, and would, in doing so, go well beyond the scope of the project.

We felt that trying to carry this out – even if the manufacturing partners agreed to it – would detract from our main goals: providing a bottom-up human centered ethics by design, with clear results as to operationalization of ethics recommendations.

Nonetheless, environmental impact and sustainability remain extremely important, as witnessed in the **UN 2030 Agenda for Sustainable Development**. The latter lays out 17 Goals, of which two in particular, Goals 8 and 9, would be relevant to the project. Goal 8: *promote sustained, inclusive and sustainable economic growth, full and productive employment and decent work for all*, has been covered in part within the scope of the project through our particular attention to working conditions of the operators, and giving recommendations to ensure that the technology does not disrupt their work conditions. Goal 9: *build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation*, is entirely relevant, but again, to be addressed properly would have needed the far deeper commitment noted above. For example, in terms of target 9.4: upgrade and retrofit of industries toward sustainability, we would need to know the current sustainability metrics of the near and further resources use of the industrial partners along their supply chains, and in terms of 9.b the degree of their integration with and support for domestic technology development in developing countries. Uncovering these would be a project in itself.

What we have done within the context of the project is to take some steps toward better defining what sustainability is. Sustainability remains a very vague notion, not least in the Industry 4.0 context. So, we have undertaken a deeper consideration of the notion in terms of what an *ethical sustainability* would be, as opposed to mere sustainability. This was done, as noted under peer reviewed publications and conferences below, in the presentation by Anderson “Exploring the Idea of Ethical Sustainability for Digital Manufacturing” at the Service Oriented, Holonic and Multi-Agent Manufacturing Systems for Industry of the Future conference, 2023, with publication forthcoming in Proceedings of SOHOMA 2023, Springer Studies in Computational Intelligence. There, a set of practical aspects of ethical sustainability for industry were laid out, which could be used to assess an Industry 4.0 or digital manufacturing process in terms of sustainability.

Part 2: Review of Methods Used, Recommendations, Recommendation Categories, and Research and Dissemination Results, used in AI-PROFICIENT

2.1 Methods Used

To expand upon the outline of modifications and additions to the HLEG guidelines given in the previous section we will here describe the 'toolkit' of methods which we developed as part of our approach.

2.1.1 Design Meeting participation

Beginning in the first months of the project, one or both members of the ethics team attended all those design meetings which we considered to have a direct or indirect bearing on the ethical outcome of the project. In the last months of 2020 and the first several months of 2021 these were meetings at the level of the first descriptions of the plant contexts – within the limitations imposed by Covid at the time – and the initial selection of UCs by the partners. This evolved into meetings particularly for Deliverable 1.3 of Task 1.3, in the spring of 2021, where pilot demonstration scenarios were decided upon. From there the technical meetings which the ethics team attended included those for Deliverables 1.5 System architecture, 5.2 Semantic data model for integrate digital twins, Deliverable 4.1 Human-machine interaction and feedback mechanism, Deliverable 4.4 on Explainable AI approaches in the project, as well as meetings for Deliverable 6.1 Validation methodology.

The meetings were nearly always by video, except for the cases where technical meetings were combined with in person general assembly meetings. We attended between 50 and 70 meetings in each of 2021 and 2022, with one or several meetings a week on average, and were still attending meetings in 2023 particularly around Deliverable 3.5 Future scenario-based decision making, which has taken up some of the actions that represented continuation of work undergone within tasks 4.1 and 4.4.

In all meetings we observed tech and industrial partner interactions, on the premise that the culture of technical-industrial design collaboration could tell us things relative to a deeper and more embedded ethics by design. The ethics team also carried out ongoing ethical analysis of design solutions as they were presented or modified at meetings, and regularly asked for clarifications on various aspects of design solutions. To be thus constantly present in technical development meetings was the one of the central aspects of our approach. It was in these exchanges that ethical issues were disclosed and the germs of the best recommendations to address them were formed to be developed into fuller recommendations, so that the ethical reasoning which accompanied recommendations often referenced statements made or decisions taken by the partners in particular technical meetings.

2.1.2 Special Meetings

Along with regular participation in design meetings which followed the development of deliverables, we also attended a number of meetings specifically requested by project partners or which we ourselves requested. These meetings – particularly when requested by the partners – confirmed our sense that the approach was on the right track in terms of a ground up and evolving ethical engagement. When a partner asks: "can you go over what we are proposing with us and tell us whether you see any ethical issues?" we feel that regardless of the level of quantitative success which we are aiming for in evaluating implementation of recommendations we are succeeding in promoting a mindset toward ethical engagement which is very promising.

Université de Lorraine (UL) team meetings were also scheduled as necessary, particularly in the first year. Given that the UL partner has a large role in steering the project toward successful results and completion, and that the ethics team was part of the UL partner, we felt that periodic updates of our

progress were useful as one component which the larger UL team oversees. We also used these meeting as an opportunity to disclose our concerns or outlook about the overall progress of the ethical component and looked to the managing members of the UL team to help us get across the importance of participation in the ethical component to the other consortium partners.

2.1.3 Weekly Ethics Team Meetings

The ethics team itself met each week. In weekly meetings we discussed particular ethical issues as they arose in the project, design solutions being advanced by the partners, and the ethical recommendations to be given. We also discussed the evolution of the ethical aspect of the project, talking about new ways that we could engage the partners and better understand the shop floor conditions, worker manager relations, and opinions of the workers. Out of these discussions came the basics of our ethical evaluation method, questionnaires for the operators, particular stances or points to be made in presentations to the consortium, and so on.

It was necessary to go beyond the bounds of the project proper to get a wider view of the ethics of AI and technology and the differences between industrial and social uses of AI, and the particularities of the former, as noted earlier. Thus, we also discussed theories and applications of ethics generally in our meetings, various other contexts of AI use and their ethical issues, and approaches to ethics of AI which differ from our approach. From there, we regularly considered where we should direct our research collaborations toward publication, as directly related to the project and as secondary, and sometimes both, e.g., the uncertainty we encountered in the project regarding the Human-in-the-Loop concept.

2.1.4 External Expert Meetings

Our external expert meetings included making contact with permanent members of the INRS (Institut national de recherche et de sécurité pour la prévention des accidents du travail et des maladies professionnelles) and scheduling a number of meetings to discuss problems which they had observed in a multi-year study of a transition in the freight transportation industry – trucking – to an automated dispatch and routing system. These problems, which included, reliability of and resistance to the new technology, and increased workload and stress through retention of legacy systems with new systems, helped us to better understand the shop floor contexts of our own project. We then organized an expert presentation of our INRS contacts to make describe the above issues to the project partners.

Further communication with INRS contacts later was turned toward learning best practices toward developing our own survey for operators and helping several tech developer partners to create a questionnaire to be used in iterative discussion with operators and process managers within Task 4.4. In the latter task we recommended an iterative process for explainable artificial intelligence[9] option selections. The tech developers were to ask the operators and process engineers what they wanted in XAI and develop it based on those needs.

We also engaged in discussion meetings, in person or virtual, with researchers developing other approaches toward applied ethics in industry 4.0 contexts, including colleagues from LAMIH (Laboratoire d'Automatique, de Mécanique et d'Informatique Industrielles et Humaines) at Université Polytechnique Hauts-de-France, colleagues at the Université de Mons in Belgium researching ethics in the industry 5.0 context, and colleagues in southeast England researching AI ethics in the context of workers.

[9] Hereafter XAI.

2.1.5 Project General Assembly Meeting Participation

One or both of the ethics team members attended all of the project General Assembly meetings called since the beginning of the project. These meetings gave us a chance to receive a broader view of the evolution of the project, as well as an opportunity to meet with members of the partner consortium directly and discuss issues in a more comfortable and less formal manner. The latter is important to get a sense of the responses of the partner members to our ethics approach, e.g., to hear "we haven't carried out these recommendations yet," or "these recommendations are not practical," and to hear and discuss why, informally. It is necessary as well to better understand the different motives of the partner members and to see how individual character shapes responses to ethical engagement. Face to face discussion at a one-on-one level about project issues is particularly good to help gain this understanding.

The General Assembly meetings also allowed us to present – in formal presentations – some global overviews of the project ethics situation where we could highlight particular ethical issues, particular recommendations or recommendation categories partially or unimplemented, or give encouragement to the partners where we felt that our efforts had succeeded in bringing ethics in successfully. In several of these meetings we were able to answer questions to the EU project observers regarding our approach, which might otherwise remain unclear.

2.1.6 Plant Visits

The ethics team visited the Continental plant three times in total. The first and second time we were guided around the Combiline area and concentrated on getting a sense of the shop floor context of each Continental Use Case, the operator manager interactions, the HMIs already in place, and the pre-existing context relative to the technical aspects of the solutions proposed for the Use Cases.

In our third Continental visit in May 2022, we concentrated primarily on operator interactions. Over 5 hours we observed a full extruder shut down and restart with a change of production. Considering the cohesion and interaction of the operator teams has been central to our bottom-up ethics approach, and so we noted the movements of the operators, their different tasks, their interactions with one another and with existing HMIs, and the speed, timing, and duration of the various tasks during the shutdown, cleaning, and restart. Our original assessment of high ethical problematicity in Continental UC 1, which played a part in its being eliminated from consideration early on, was confirmed here as we saw the speed, physical range, and variety of stations over which the extruder operator had to carry out his job.

We visited the Ineos Geel plant in November 2021. Relative to Ineos 1 and Ineos 2 UCs, our main interest was to better understand the control room hierarchy and environment and the physical distances and conditions within the big bag hopper loading area and in relation to the control room – issues central to Ineos 2 UC. We had opportunity to observe the control room operators at work and noted the evident comfortable and friendly team cohesion and asked various questions about cohesion and hierarchy in the Geel plant context. Observations here fed into our evolved second set of UC specific recommendations regarding Ineos 1 and Ineos 2 UCs.

The ethics team visited the Ineos Cologne plant in February 2023, where we saw the general Ineos operations as well as the PE3 reactor area. The Polyethylene testing lab visit was of particular interest to us in the visit, because it was closely related to Ineos UC 3. The lab visit confirmed to us the relevance of our very physical and ground level approach because we observed that a very physical level of operations plays a major part, and perhaps *the* major part in Ineos operations. As the quality manager showed, most of the lab analysis is physical rather than chemical, e.g., analysis of the shape, colour, and density of the pellets. In the logistics area this was reconfirmed, as the physical manipulation of product was very evident – heavy product – at very clockwork rates, and again with physical issues of pellet dust and noise. While the physicality and temporality of the context where somewhat less relevant to the AI ethical issues in Ineos 3 UC specifically, they added to our sense that the heavy industry context must always take these aspects into account if ethical outcomes for AI integration are to be achieved.

2.1.7 Review of Deliverables as they were being written

As various project deliverables were being collaboratively written by members of the consortium partners, we reviewed those deliverables which we judged to have a direct or indirect relation to the ethical aspect of the project. The specific deliverables which we collaborated on are listed below in 2.4 Research and Dissemination Results. The most important deliverables for ethics were reviewed several times as partners added to them, and we generally reviewed the whole deliverable, i.e., not just sections directly related to ethics, in order to disclose latent ethical issues. This often occurred in parallel with deliverable specific partner meetings. The ethics team felt it important to use the same evolving approach to collaborative development of deliverables as for the larger project. Thus, we added direct off the record comments onto the deliverables, made or suggested wording changes as the deliverables advanced, and in some cases made ethical recommendations concerning the deliverable itself.

In these reviews and in discussions with project partners and task leaders we also evolved our recommendation format to embed recommendations and the reasoning behind them directly into the deliverables when appropriate, so that for the public deliverables, our ethics approach could be visible and usable by others wishing to reference our model of operationalizing ethics for AI and industry.

2.1.8 Continuous Monitoring and Recording of Ethical Implementation

Our approach involved continuously monitoring and recording changes or advances in implementation of our recommendations as they occurred. In most of the instances of full implementation the implementation was carried by parts, and in other cases we have still only achieved partial implementation. Thus, our running record of implementation gave us a sense of where the recommendations was at with respect to the UC and the progress of WPs and Deliverables.

At first, we recorded implementation progress or lack of progress on the subsequent versions of the CO (confidential) level ethical recommendations documents we sent to partners, with more informal notes being kept as well. Later we adopted the use of an online collaborative spreadsheet, in which we recorded our formal ethical recommendations in a more compact format. To this we added information on the ongoing implementation progress, our assessment of the implementation status of the recommendations, our evolving assessment of the partner(s) responsible for implementation, specifically what the ethics team considered to be a full implementation of the recommendation, and date the recommendation was made.

More recently we have created a version of the spreadsheet accessible by consortium partners, with a column for the partners to give their own assessment of the implementation result of recommendations for which they are responsible, as well a column to note their reasons if a recommendation was not or not fully implemented. These latter sections have been used to help us compare the ethics team's assessment of progress in operationalizing ethics to that of the partners and feeds into our insights to be carried forward into future EU level projects – see tables and charts below – about what kinds of recommendations tend to get implemented and what kinds do not, and what we can do in future to improve ethical recommendation implementation.

2.1.9 Research in AI or General Ethics or Technical Concepts

Ethics, as a branch of philosophy, is a discipline which traditionally includes a substantial research component in the form of reading and reflection, either toward theory or application. Thus, considerable reading research in primary and secondary literature was carried out for the project, in order to better understand the issues around operationalizing AI ethics in the heavy industry context. It should be *strongly* noted however that we have not proceeded directly from this research in making recommendations, i.e., *we base our recommendations primarily on observing and addressing contradictory tendencies between design solutions proposed and existing situations in the work (shop floor) context, while keeping in mind the dictum that at a minimum the work context for the operator (or engineer) should not be degraded by the solution*. Our approach is thus founded upon a pragmatic applied logic which proceeds from the bottom up in order to link to more abstract grades of ethical theory.

To a lesser extent we then go on to compare solutions proposed with existing regulations, and we also try to incorporate a deliberately positive aspect in transforming our researched knowledge in various fields beyond ethics into recommendations instantiating ground level best practices into the design processes of the project partners, e.g., applying psychological research insights into the explainable AI solution development process.

Our project related research into general ethics has centered around pragmatic and communitarian ethical theory primarily, as well as deontological theory. On the pragmatic side we aimed to gain insights respectively into how to better bridge applied ethical practice at the level of the shop floor and the technological design process, with a consistent higher-level account of human moral formation as a slow accreting growth, which has setbacks but advances through 'trying things out.' On the communitarian side we have tried to better understand the formal structure of ethical communities in order to engage more successfully the already formed communities of our industrial partners and tech developer partners – as a group of international members who come together as a community to collaborate specifically over an extended period of time for the AI-PROFICIENT project. From the deontological angle we attempted through research into deontological paradigms, to better understand the perspective of the current more high-level approaches – e.g., the HLEG guidelines – in order to integrate our alternative approach with what the majority of AI ethics researchers are doing. Thus, we draw from a number of philosophical ethical traditions less often drawn upon in AI ethics.

Our research into legal and regulatory texts covered full readings of the main regulatory texts themselves, e.g., the GDPR, as well as secondary texts which discuss the main texts, and general legal texts which intersect with some of the ethical traditions noted above.

We read a number of texts in sociology, psychology, and anthropology. These provide ideas toward best practices in operationalizing ethics. Here the knowledge of prior research in how workers and tech developers respond to ethics as a new but growing component of their practices interests us. Our readings of current empirical research show us that the working cultures and educational backgrounds of people who design technology – programmers and engineers – has a significant influence upon how far they are willing to go in considering and implementing ethical practices. This fact and the specifics of it, are invaluable in an applied approach such as we are attempting.

Technology related research generally and ethics of technology more specifically also made up a large portion of our background readings. The culture of tech design uses many terms easily which can be opaque to the ethicist unless they are researched more deeply. We kept a running glossary of tech terms encountered in design meetings. In this category go also various texts consulted concerning the Industry 4.0 paradigm and work automation generally.

Finally, the concept of work itself and the concept of the worker, which is more clearly defined in heavy industry than anywhere else perhaps, interests us. To that end we read a number of texts about the historical and evolving nature of work and the ethical aspects of work as such. Added to this were more technical readings related to empirical research on case studies and outcomes of automating work in various heavy industry sectors.

Below are listed a selection of the larger texts read fully for the project at the level of books or monographs. Readings of current AI Ethics, ethics of technology, and AI technical journal articles were numerous, but are not listed.

General Ethics:

Bentham, Jeremy. (1823) *An Introduction to the Principles of Morals and Legislation*. Oxford. Clarendon Press. [1789].

Bradley, F. H., (1927 [1876]). *Ethical Studies*, 2nd edition, Oxford: Clarendon.

Dewey, J., and Tufts, J.H. (1932) *Ethics*. 2nd Edition. New York. H. Holt and Company. [1908].

Green, T. H. (1884). *Prolegomena to ethics*. Oxford. Clarendon Press.

Kant, Immanuel, (1788). *Critique of Practical Reason*.

Lewis, C. I. (2017). *Essays on the Foundations of Ethics*. State University of New York Press.

Legal and Regulatory:

General Data Protection Regulation (2018)

European Data Protection Board Guidelines on the application of Article 65(1)a GDPR (2021)

EU AI Act Proposal (2021)

HLEG expert Guidelines for Trustworthy AI (2019)

OECD AI Principles (2019)

UNESCO Recommendation on the Ethics of Artificial Intelligence (2021)

UN 2030 Agenda for Sustainabile Development

Sociology, Psychology, and Anthropology:

Latour, Bruno. (2000) “La fin des moyens.” Reseaux, vol. 18, no. 100, pp. 39-58, 2000. https://www.persee.fr/doc/reso_0751-7971_2000_num_18_100_2211

--- *What is the Style in Matters of Concern?*. Van Gorcum. Assen, Belgium. 2008

Suzman, J. (2022). *Work: A deep history, from the stone age to the age of robots*. Penguin.

Technology and Ethics of AI and Technology:

Coeckelbergh, M. (2020). *AI ethics*. Mit Press.

Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Floridi, L. (2013). *The ethics of information*. Oxford University Press, USA.

Mazis, G. A. (2008). *Humans, animals, machines: Blurring boundaries*. State University of New York Press.

Russell, S., & Norvig, P. (2021). *Artificial Intelligence: a modern approach*, 4th US ed. University of California, Berkeley.

Van Den Eede, Y. (2019). *The beauty of detours: A Batesonian philosophy of technology*. State University of New York Press.

Work:

Applebaum, H. A. (1992). *The concept of work: Ancient, medieval, and modern*. Suny Press.

Smil, V. (2021). *Grand Transitions: How the modern world was made*. Oxford University Press.

2.1.10 Reasoning behind Recommendations

We have used our reading research to expand upon and back up the reasoning behind the ethical recommendations to help the project partners learn a culture of ethical engagement by practice as much as possible, rather than simply blindly carry out the recommendations. Interactions with project partners indicate that they have often taken time to read our ethical reasoning sections.

Ethical reasoning sections were included beginning with Ver. 1.0 of the UC specific CO level formal ethical recommendations documents, with citations of relevant ethical theory and supporting empirical research as necessary. They were given under the heading Ethical Issues and usually correlated numerically with the ethical recommendation of the corresponding number which followed in same the document. Sometimes they were more general and discussed the issue with regard to the cited

research. They usually referred to statements or decisions made in related design meetings as the design solutions evolved.

In several cases separate documents were written to clarify or expand upon particular ethical recommendations, e.g., the practical application of recommendation ETH ID 1.5-3 which recommended categorizing the humans (operators, engineers, etc.) in terms of their being processes with respect to the development of the platform architecture. The latter for example, was written after the lead partner of Deliverable 1.5 communicated to us their uncertainty about what we were looking for in the recommendation, and they helped that partner to better understand and fully implement the recommendation.

These ethical reasoning sections were often included in related deliverables in slightly modified form, making up one of the components (besides recommendations) of the ethical issues section which was soon adopted as standard for relevant public deliverables.

2.1.11 Ethical Recommendations

Along with the bottom-up consideration of the time and space work context experience which intersects considerably with them – since we use the recommendations to build up our picture of that context – the ethical recommendations are one of the central aspects of our approach. Based on our position that high level guidelines such as the HLEG guidelines are generally inadequate by themselves to actually 'get AI ethics done,' we needed another approach that was not abstract and general, even though it should be able to connect to the abstract and general eventually.

What the ethics team envisioned for the project therefore was a way of doing ethics where we could eventually show definite results at a very low level. Much of AI ethics and tech ethics literature seems to us to revolve in a continual circle of advancing abstractions, criticizing those abstractions, and then advancing further abstractions. This is a form of theory building of ethics indeed, but not a healthy form,[10] and completely divorced from actual practices. In short it becomes a round of talks about *the culture engaged in creating ethics in AI and tech* – almost a cousin of 'management speak' – rather than an embedding of ethics into AI and tech development practices. When it is not the former is too often merely critical rather than positive. The recommendations were therefore our way to 'get AI ethics done' at ground level, by beginning from the opposite position of most current approaches.

They allow the ethics team to keep track of a quantitative aspect for ethical engagement. The ethics team has kept track of how many of them were carried out and to what degree. They also allow that quantitative aspect to be merged with a more qualitative aspect. Having categorized them (qualitative), some idea can then be got (quantitative) of implementation results by category. Further one can begin to see how different consortium partners are able or willing to carry them out to different degrees, and even – though we did not undertake to record it here – how certain individuals among those partners are more or less apt or willing at implementing them.

Our recommendations were given at CO level in formal documents, but as noted earlier, most found their way into public deliverables of the project. The recommendations remained flexible. They were modified depending on new information regarding the work context or the design solution proposed by the tech developer. They were dropped if rendered not applicable by design changes. This flexible and evolving stance permitted a regular monitoring and recording of progress in implementation. It also allowed them to progress from UC level to task/deliverable level.

[10] In the case of AI and tech ethics it is very often also blind to the long slow tradition of ethics as a branch of philosophy, where there are many deeply thought out and consistent ethical theories, any one of which might serve to underpin an ethical development of AI and tech, but *only* if they are operationalized. It is relatively easy to find the names of Kant, Aristotle, Mill, etc. dropped into an AI ethics article preamble, but much more difficult to find anyone actually applying the respective theories at ground level.

2.2 Recommendations

The ethical recommendations are identified now according to a merger between the encodings decided upon in AI-PROFICIENT Deliverable 6.1 and our own original identification scheme. This identification is a progression from the numbering scheme originally adopted. Originally in the first version/round of recommendations given to the partners, the recommendations were numbered simply according to one of the eight UCs, thus e.g., UC PartnerX 2, 1).

The second version of recommendations still centered around the UC level, but now as engaged through Deliverable 1.3 of Task 1.3 Pilot Demonstration Scenarios. These recommendations were about issues arising through the progressive development of design solutions being outlined in Task 1.3. They also involved modifications or additions to recommendations already given, and in some instances, they focused on the Deliverable 1.3 itself. They were thus numbered according to one of the UCs but now in relation to the deliverable/task as well, e.g., UC PartnerX 2, 1.3-1).

From there on most deliverables and associated tasks began to diverge from a comprehensive focus on all UCs, to being about selected UCs depending on the interests of the different partners. We thus dropped the UC relation and retained the deliverable/task numbering format only, with new recommendations becoming Task x.x Specific, thus for Task 1.5 the numbering was 1.5-1).

Current numbering incorporates the Deliverable 6.1 encodings adopted with our original numberings. Thus, for example: ETH C UC 2 ID 1, corresponds to recommendation 1 of PartnerX_2 Ethical RecommendationsVer 1.0 document, ETH I-G UC 2 ID 1.3-1 refers to recommendation 1.3-1 of PartnerY_2 Ethical RecommendationsVer 2.0 document, and ETH ID 3.5-1 refers to recommendation 3.5-1) of T3.5: Future scenario based decision-making and lifelong self-learning_Ethical Recommendations Ver 1.0 document.

130 ethical recommendations have been given in total throughout the project. Some of these recommendations have been upgraded to N/A (not applicable). N/A status changes for some of the recommendations came about through design modifications, generally the dropping of some aspect of the design solution. The remaining portion of the N/A recommendations have come from one UC however, where the industrial partner decided to abandon the deployment stage of the UC altogether.

The general format of the recommendations followed the pattern: "*we recommend that you do x*." In the first versions we did not specify responsible partner precisely beyond the default of UC leader and relevant industrial partner, the latter depending upon the particular UC. Task specific recommendation documents were largely UC independent, so that we began to specify the relevant partners or partners for the recommendation. We distributed the CO level ethical recommendations documents to a group of ethical contacts designated early on by each of the consortium partners. These contacts were subscribed to our project ethics mailing list. In addition, we distributed by email the CO level documents to other partner members who were involved in each particular UC or task level deliverable. Finally, the documents were uploaded to our ethics Teams channel – open to the ethics contacts and others – to refer to as needed.

Below we provide a representative set of examples of recommendations (anonymized) beginning from one UC and proceeding to task level recommendations related to that UC, in order to give an example of the evolution of our recommendations.

Ethical Recommendations

UC code: XXXX_X_XXXX

Pilot: XXXX XXXXXX plant

UC full name: XXXX X XXXX XXXX

Version: 1.0

Dissemination Level: CO

Nb. Ethical recommendations are evolving. The first versions are directed more toward the work context on the industrial partner side. Later versions will shift toward the tech partner development side, implementation of AI and Human/AI interfaces and interactions, with recommendations for the tech partners.

Ethical Issues:

1) Human-in-command can be defined as *the possibility of deciding when and how an AI will be used (or not used) combined with the capacity to supervise the activity of the AI in the broadest sense*. According to this definition the operator has decision power over the AI use here insofar as the Partner X request was for this Use Case to incorporate human-in-command.

Given the above, there is some ambiguity in this Use Case regarding how the operator will use the AI proposals. Will the operator be expected to always use the proposal of the AI, or to judge the suggestions of the AI and decide when to use it based on his experience?

It was suggested in Q&A (Jan 19th and Feb 16th) that the restart is based significantly on the feeling of the operator and the operator's experience. There are 20 parameters which the operator looks at and which must be nominal for restart.

2) Currently the operators only increase rpms for cap compounds (main extruder). It was suggested in Q&A (Feb 16th) that in future the operator will adjust rpms for all extruders, which will add further tasks for the operator.

Recommendations:

1) The *default* position about whether the driver operator is always expected to follow the AI proposal should be specified either overall, or for various phases of AI integration if there is a trial period. A trial period with phasing in of the AI integration in stages should be implemented.

e.g., first 6 months – operator will consult AI proposals but use his own judgment whether to implement them.
next 6 months – operator must always implement AI proposals unless it is clear that AI proposal causes some major problem.

Formally clarify at what stages the operator has command over the AI to the point of ignoring its suggestions if he chooses (according to the human-in-command definition)

2) It should be clarified at the beginning whether some time is envisioned when the operator can stop looking at the restart parameters. If so, clearly separate this period from a trial period to come before in which he must continue to monitor the relevant parameters.

I.e., for building up *trust* in the *robustness* of the AI there should be a 'phase in period' in stages (combined with #1 above) where the operator's monitoring of the parameters is relaxed progressively (if it is going to be relaxed), rather than leaving it to be decided in an unplanned and ad hoc way.

3) There should be a *protocol* created to deal with the situation when the AI makes an error: Formally clarify who the operator is supposed to report it to.

Formally clarify under what_conditions the operator should report that the AI is in error (e.g., if the AI suggestion does not look right according to his previous experience)

4) The operator will be expected to adjust all extruders if AI is integrated. An estimate of how much more time this will take should be made. It should be determined whether the operator has enough time to do this added job and how much the added time to adjust will offset the reduced downtime in restarting the extrusion.

The added time for the operator to adjust should be factored into the setup time component of the KPI and used to decide about a minimum success rate threshold above which the AI is worth deploying eventually.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 957391.

Ethical Recommendations

UC code: XXXX_X_XXXX

Pilot: XXXX XXXXXX plant

UC full name: XXXX X XXXX XXXX

Version: 2.0

Dissemination Level: CO

Ethical Issues - Task 1.3 Pilot Specific Demonstration Scenarios:

Review of Task 1.3 response to Version 1.0 Recommendations (XXX Leading):

1) Regarding staged implementation:

Commitment to proceed in staged implementation of AI services, specifically to begin with a single extruder and evaluate both extra workload for operator and feasibility to extend to further extruders.

Clarification that AI suggestions are provided as guidelines rather than imperatives.

2) Regarding operator monitoring of parameters:

The proposal so far is for the operator to remain in command and monitor as before with addition of AI suggestions, thus status quo for monitoring.

3) *Recommendation to develop an AI error protocol remains unaddressed at level of Task 1.3, except in operator feedback to retraining system.*

4) Regarding estimates of extra time/work for operator and AI success rate:

Commitment to evaluate extra operator workload and success of system in first stage on a single extruder.

In general, a strong effort to address preliminary ethical recommendations.

Additional Recommendations, Task 1.3 Specific:

1.3-1) Regarding: What advantage do we provide to the final user? – "less experienced operators…"

More experienced operators should be able to give more valuable feedback for the retraining system. *Recommend that in consultation with Conti you select the most experienced operator(s) to interact with the first stage of the AI integration on the single extruder, and then have that/those operator(s) guide the less experienced operators if you proceed to multiple extruders*.

Oct 27th, 2021 – (Deliv 1.3 material moved to Gap Analysis) – recommendation unaddressed

1.3-2) Regarding: Addressing Ethical Consideration – "Operators will be provided an HMI…"

Unclear if this means a new HMI. *Recommend that you integrate the solution with existing HMI setup as much as possible to provide smooth transition to new AI services for operator*.

Oct 27th, 2021 – (Deliv 1.3 idem) – recommendation not addressed

Additional Recommendations, Task 1.3 Specific: (October 2021)

1.3-3) Re: UC2 Diagram – Deliverable 1.3

Recommend that you clarify which stakeholder is represented by the human symbol at UC2 – Extruder, (or which stakeholder is tentatively proposed), and designate Extruder separately, e.g. (at Extruder) or (component – Extruder)

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 957391.

Ethical Recommendations

T4.4: Explainable and Transparent AI Decision Making

Version: 1.0

Dissemination Level: CO

Ethical Issues - Task 4.4 (XXX leading)

General Discussion:

The task involves development of explainability and transparency with regard to AI services for Conti UC2, Conti UC5, Conti UC 10, and Ineos UC3. It was stressed in the last meeting that there is a tradeoff between accuracy and interpretability. To develop that task in the most human-centered manner, the tech partners should work to have maximum interpretability relative to acceptable accuracy. In other words, choose the explainability methods for each use case so as to use up the room for maneuver that you have relative to the minimum accuracy.

If the explainability is to be human centered, then it is necessary to consider the understanding of the user who is receiving the explanation. (Jacovi et al. 2022)[11] have shown that explainability in AI can be mapped to folk concepts of behaviour: *the person getting the explanation assumes that the AI has intentions like humans do*. If the AI process is not clarified by the explanation, the person receiving the explanation (e.g., operator) will fill in the blanks with imaginary assumptions which can easily be contradictory with further instances of AI explanation.

Further, (Ehsan et al 2021)[12] have shown that the background of the person receiving the explanation changes how they see the explanation and leads to problematic or mistaken ascriptions of intention to the AI. More specifically everyone prefers humanlike explainability – e.g., natural language explanations –, but fully natural-language like explanations are viewed by those with more technical exposure to AI as indicating more complex 'thought processes' in the AI.

The use of numbers primarily in the explanation affect the perception of the 'intelligence' of the AI as well. But including numbers in the explanation affects those with different backgrounds in different ways: those *with* a more technical exposure to AI tend to ascribe diagnostic value and methodological intention to the AI process. Those *without* a technical exposure to AI ascribe greater intelligence to numbers-based explanations even when the numbers-based explanations make the AI process *less* understandable. They follow "heuristic reasoning that associates mathematical representations with logic and intelligence." (ibid.) In other words: 'it looks complicated and mathematical so it must be intelligent.'

For everyone, the desire to *collaborate* with the AI *decreases* with the use of numeric explanations and increases with the use of natural language explanations.

These findings raise several issues in terms of AI-PROFICIENT project, particularly around the issue of *trust* in the AI. If we want to make the explainability human centered for the project we should consider the background of the users of the proposed explanations more closely before finalizing the choices of explanation. The different backgrounds will be those of the operators, the process engineers, and the data scientists of the project. We should clarify how much background knowledge of AI each group has – particularly the operators and process engineers – and if possible, for each individual who will be most affected by an eventual live integration of the AI. From there the methods of explainability should be developed. The background of the data scientists of the project is liable to engender a trust in the AI which is unwarranted in terms of validation of project goals which are related to

[11] Jacovi, A., Bastings, J., Gehrmann, S., Goldberg, Y., & Filippova, K. (2022). Diagnosing AI Explanation Methods with Folk Concepts of Behavior. ArXiv, abs/2201.11239. https://doi.org/10.48550/arXiv.2201.11239

[12] Ehsan, U., Passi, S., Liao, Q.V., Chan, L., Lee, I., Muller, M.J., & Riedl, M.O. (2021). The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations. ArXiv, abs/2107.13509. https://doi.org/10.48550/arXiv.2107.13509

explainability. The background of the process engineers and operators are likely to be susceptible to unwarranted trust in the AI process, particularly in numbers of heavy explanations.

(Jacovi et al. 2022) suggest that to be effective, explanations should be coherent, i.e., the information the explanation provides can be generalized to further instances of AI behaviour. But they also suggest that if contradictions occur between instances, this should not be viewed as failure, but rather an opportunity to develop *iterative explanation processes*, where the mental model of the (explanation user) is adjusted between instances of AI behaviour, until no more contradictions arise.

Jacovi et al. give suggestions to mitigate the potential effects of anthropomorphic bias. These include: attempting to understand the intent which the explanation user perceives in the AI and designing to account for it (modifying the explanation design); controlling the perceived intent (shaping the explanation user's mental model); clarifying to the explanation user that the AI process is not intelligent (through various possible methods).

The AI-PROFICIENT project is working with industrial processes which are more number oriented than concept oriented, and thus perhaps more difficult to engage with natural language explanations. The is a danger that the explanations are also liable to become 'number oriented' so as to create an unwarranted trust in the AI processes on the part of the operators and process engineers, based upon the perception of the AI services as 'logical and intelligent' due to the numbers. Along with the tendency for 'numbers oriented' explanations to result in decreased collaboration between human and AI, this might result in the operators and process engineers *'leaving the AI alone to do its thing'* so to speak, on the unwarranted assumption that *'the AI knows best*.'

The feature attribution methods LIME and SHAP are specifically proposed by XXX and XXX. Accordingly, something to watch for here, is whether the user (operator or process engineer) assumes a particular contextual interpretation of the input by the model, when in fact the model is using it otherwise. In terms of the numbers-oriented nature of the project, the potential for this failure should be uncovered by observing and questioning the users to get a sense of their assumptions with regard to the inputs.

If the timeline of the project allows for iterative development of explanation methods, then this should be used, after a user-based choice of the XAI methods. If the timeline does not allow for such an iterative development, then for each method of explanation adopted a special effort should be made *to show the operators and process engineers that the AI processes are not intelligent, through a clear mechanistic explanation of those processes*.

Recommendations, Task 4.4 Specific:

4.4-1) (All) *Recommend that you discuss directly and regularly (once a month) with the process engineers most closely involved with each of the UCs under consideration as you develop the transparency models. Recommend that the process engineers discuss similarly with the operators who will be most closely involved.*

4.4-2) (All) *Recommend that you formally clarify who will be the user(s) of the explanations for each UC (e.g., data scientists, process engineers, operators), on an individual level if possible.*

4.4-3) (All) *Recommend that in discussion with the process engineers you clarify what level of accuracy of the AI is acceptable for each UC and then you decide which options for explainability methods remain open based upon that.*

4.4-4) (xxxx; xxxx) *Recommend that you carry out a preliminary short survey, e.g., 10 questions of user background knowledge (process engineers and operators) regarding AI, to be used in adjusting for potential user assumptions during XAI development.*

4.4-5) (xxx, xxx, xxx) *Recommend that for all explainability methods destined for process engineers and operators, you review the models directly with the process engineers and/or operators as soon as possible after the prototype stage (or with a mock input and result), asking them directly: "do you understand this method generally?" and then adjust for their concerns.*

4.4-6) (All) *Recommend that, after you advance beyond the prototype stage you pursue an iterative development of explainability methods if project timeline allows.*

4.4-7) (xxx, xxx, xxx) *Recommend that if project timeline does <u>not</u> allow for the iterative development (recommendation 4.4-6), you have several sessions with the process engineers and operators where you present clear mechanistic explanations which characterize the AI processes as un-intelligent tools.*

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 957391.

2.3 Recommendation Categories

Below we give the definitions of the categories used in our evaluation of ethical results. Note that recommendation categories were developed *after* most of the specific recommendations had been given and were not disseminated to the project partners. This is in keeping with our bottom-up approach. The categories were created for two reasons. First, in order to gain insights into what types of ethical recommendations are most likely or unlikely to get implemented, so that we can then reflect upon the reasons for these differences and suggest where future research in ethical implementation and operationalization is needed for AI in Industrial settings. Second, in order to show how categorization can be done within a bottom up and context-based approach, i.e., have the categories reveal themselves organically from within the context themselves, in contrast to a top-down approach – as in the HLEG guidelines – which attempts to generalize every issue that might occur in advance and simply compartmentalizes odd ethical issues accordingly. In short have the categories, which are aimed at being useful to future researchers, tailored to the ethical issues, rather than cropping, or discarding, the issues to fit into pre-assumed generalizations.

2.3.1 Definitions of Categories

Recommendation Category	Definition
Protocol	Adopt a specific set of instructions regarding errors, new tasks, etc.
Human centering	Tailor aspects of development to individual users and develop services collaboratively with users (e.g., work with the operator to design something)
Design	Make changes or additions to technical or procedural elements of the solution
Insufficient Specs	Clarify aspects of the production or development process (e.g., in what format is operator feedback gathered, how many suggestions will AI give operator, what XAI methods will be used, etc.)
GDPR	Check whether a solution follows the spirit of GDPR regulations
Responsibility	Confirm or change who is responsible for tasks in some part of the process or what their new tasks will be
De-anthropomorphizing	Change anthropomorphic wording or thinking about AI
Simplification	Try simpler techniques first
Verify effects	Verify whether a proposed implementation would have some human effect (e.g., does AI service effect team cohesion?)
Timeliness	Implement certain other recommendations in a timely manner
Valorize experience	Make better use of human abilities/experience
Ethical rewording	Reword a text or redraw a diagram to better include the human contribution
Workload	Estimate how much, how long, how many, of some new task to be done
Evaluation	Assess whether some aspect of the workplace context is considered in the proposed quantitative outcome of the solution, e.g., acceptable error rate, or set a range for quantitative assessment of service, e.g., reliability

Training	Suggestion to provide specific training or implement services by stages

2.3.2 Process of Categorization and Agreement Results

Motivation:

The ethics team is interested in knowing not merely how many ethical recommendations have been achieved but what types of recommendations are more or less likely to be implemented. Thus, we required a bottom-up categorization of recommendations, i.e., one growing out of the specific issues encountered and one which could accommodate those issues, rather than discarding issues not fitting within a pre-developed categorization.

This helps understand how the industrial partners and tech developers see the ethical aspect, helps make suggestions or advance potential future lines of research to improve implementation of recommendations in poorly implemented categories, and makes it possible to apply insights from some categories toward better implementation of recommendations in other categories.

Methodology:

Manual annotation, in our case categorization, is not about measuring a physical reality (such as the height of Mont Blanc), but rather about quantifying a phenomenon (Desrosières 2008). This implies agreeing beforehand on conventions of equivalence: for example, in order to count unemployed people, we first have to agree on what defines unemployment.

These conventions should then be documented, for example in annotation guidelines, and the consensus should be measured, using inter-annotator agreement metrics (Artstein and Poesio 2008). This methodology was used to categorize the ethical recommendations of the AI-PROFICIENT project.

Process and Results:

One of the authors categorized the 120 recommendations *which had been made up to that point in the project* [13] into 15 categories (annotation #1), e.g., Human Centering, and Responsibility. They wrote annotation guidelines, precisely defining each category. The second author then read the guidelines and, without consulting the first round of annotations, did their own (annotation #2). They disagreed on the categories for 60 recommendations (50% of the cases), fully agreed in 38 cases (32%) and hesitated between the annotation #1 and another category in 22 cases (18%). The strict observed agreement is therefore 31.66%. If we consider the ambiguous cases as part of the agreement, we reach 50%.

The results show that the categories needed to be reviewed. We first improved some definitions in the guidelines such as the one for ‘Timeliness’ (removing the emphasis on order of implementation and redefining the category fully in terms of getting another recommendation implemented ‘earlier’), then we merged some categories (e.g., ‘Human centering’ with ‘feedback,’ and ‘training’ with ‘adaptation’), while adding new ones, such as ‘GDPR.’ Finally, we went over all the recommendations one last time, discussing each agreement case and making a consensus decision on them.

Four categories were assigned to at least 15 recommendations: Human centering, Design, Responsibility and Workload. On the other end of the spectrum, four categories were used less than

[13] The categorization process began at about the end of the second year of the project, at which point we had made 120 recommendations, and was thus a ‘snapshot’ of that point. Eventually 10 further recommendations were made, for a total of 130, as noted earlier, but those were not considered in the categorization process.

five times (four times for each): Evaluation, Training, GDPR and Verify effects. The results are satisfying since there is no prevalence of one category in particular and there are no useless categories either.

2.4 Research and Dissemination Results

The research of the ethics team on our methodology for ethical recommendations or on related issues was disseminated through the public level deliverables of the project, peer reviewed journal articles or peer reviewed conference proceedings, conferences, and digital fact sheets and info packs. Since the beginning of the project the ethics team has written, co-written, or contributed to, seven project deliverables (not including the present deliverable), five peer reviewed articles or proceedings, and ten conferences or public outreach events, as outlined below.

2.4.1 Deliverable Contributions

Besides the current Deliverable, the main project deliverable contributions by the ethics team were the following.

2021 Fort and Anderson "AI-PROFICIENT Deliverable 1.2: Legal and Ethical Requirements for Human-Machine Interaction," online at: https://ai-PROFICIENT.eu/wp-content/uploads/2021/09/D1.2-Legal-and-ethical-requirements-for-human-machine-interaction_v1.0.pdf

2021 Arnaiz et al. "AI-PROFICIENT Deliverable 1.3: Pilot Specific Demonstration Scenarios." online at: https://ai-PROFICIENT.eu/wp-content/uploads/2021/09/D1.3-Pilot-specific-demonstration-scenarios_v1.0.pdf

2022 Berbakov et al. "AI-PROFICIENT Deliverable 1.5: AI-PROFICIENT system architecture." online at:https://ai-PROFICIENT.eu/wp-content/uploads/2022/07/D1.5-AI-PROFICIENT-system-architecture_v1.1.pdf

2022 Lopez de Calle et al. AI-PROFICIENT Deliverable 2.5: Local automated control for quality assurance."

2022 Fernandez et al. "AI-PROFICIENT Deliverable 4.1: Human-machine interaction and feedback mechanisms (Design and specification)." Online at: https://ai-PROFICIENT.eu/wp-content/uploads/2022/07/D4.1-Human-machine-interaction-and-feedback-mechanisms-Design-and-specification_v1.0.pdf

2023 Pujic et al. "AI-PROFICIENT Deliverable 4.4: AI-PROFICIENT approach for XAI." online at: https://ai-PROFICIENT.eu/wp-content/uploads/2023/02/D4.4-AI-PROFICIENT-approach-for-XAI.pdf

2023 Van Loock et al. "AI-PROFICIENT Deliverable 6.5: Best Practices and Lessons Learnt."

2023 Fernandez et al. "AI-PROFICIENT Deliverable 6.6: AI-PROFICIENT Validation methodology." online at: https://ai-PROFICIENT.eu/wp-content/uploads/2023/02/D6.6-AI-PROFICIENT-validation-methodology-final-version.pdf

2.4.2 Peer Reviewed Research Publications

The peer review article or conference proceeding publications of the ethics team directly or indirectly addressing ethical issues disclosed in the project, or our methodological approach for the project, included the following.

Marc M. Anderson and Karën Fort, “From the Ground Up: Developing a Practical Ethical Methodology for Integrating AI into Industry” in AI for People Special Issue in *AI & Society - Journal of Culture, Knowledge and Communication* (2022). This article surveys current approaches in AI ethics for industry and AI ethics generally and the poverty of practical application in such high level and abstract approaches. We then describe the embedded and bottom-up ethics approach developed for the project by the ethics team and provide examples of our direct ethical recommendations in the project and the ethical reasoning which accompanied them.

Marc M. Anderson and Karën Fort, “Human Where? A New Scale for Defining Human Involvement in Technology from an Ethical Standpoint” in *International Review of Information Ethics* (2022). This article discusses the history and evolution of the Human-in-the-Loop term and related terms since their adoption in the early 1950s and argues that it has been given multiple and conflicting meanings over the past seventy years. We then consider the ethical import of the terms and argue that they are unsuitable for AI and tech ethics. Finally, we go on to suggest a new scale for human interaction with AI and automated systems based on the notion of participation in the community developing the technology.

Marc M. Anderson, “Some Ethical Reflections on the EU AI Act”, IAIL 2022, Best Short Paper Award - Imagining the AI Landscape after the AI Act 1st International Workshop on Imagining the AI Landscape After the AI Act. *CEUR workshop Proceedings. Vol 3221, Sept. 2022*. ISSN 1613-0073. http://ceur-ws.org/Vol-3221/. Here Anderson considers the European Commission AI Act proposal development process and the general aims of the act in terms of ethics as a practice. He argues that the conflation of law and ethics needs to be re-examined with regard to the AI Act proposal. Specifically, it is not evident that the AI Act regulation is ethically grounded, as witnessed in the Act’s objectives, its consultation process, and the ‘speed paradigm’ evoked in the regulatory process and definition of AI.

Marc M. Anderson and Karën Fort, “Ethical Internal Logistics 4.0: Observations and Suggestions from a Working Internal Logistics Case” in Service Oriented, Holonic and Multi-Agent Manufacturing Systems for Industry of the Future. *Proceedings of SOHOMA 2022, Springer Studies in Computational Intelligence* (2022). In this publication, in keeping with our aim of operationalizing ethics at ground level, we take a closer, but anonymized, look at one of the UCs of the project which falls under the category of Internal logistics, namely that of AI aided OCR text recognition of ‘big bag’ labels at one of the factories. We discuss, with examples, some of the real-world difficulties encountered as well as some successes in achieving implementation of our ethical recommendations in the UC. We then discuss potential sources of the difficulties, particularly the top-down framework vision of the Industry 4.0 concept and conclude with some suggestions regarding a consistent path for an ethical internal logistics.

Marc M. Anderson, “Exploring the Idea of Ethical Sustainability for Digital Manufacturing” Service Oriented, Holonic and Multi-Agent Manufacturing Systems for Industry of the Future. *Proceedings of SOHOMA 2023, Springer Studies in Computational Intelligence* [forthcoming]. This article discusses the notion of ethical sustainability in Digital manufacturing, in comparison to mere sustainability. It defines ethical sustainability on the basis of ethics as a practice of removing contradictions in manufacturing processes. It is argued that an ethical sustainability would be one which moves beyond a ‘parsimonious with resources’ paradigm, by adjusting action, embedding creativity, and bringing the human individual and human society back into the manufacturing process.

2.4.3 Conferences or Public Outreach

We disseminated our research results and discussed our method and the context of the project at a number of venues, either indirectly or indirectly. Conferences or presentations attended with direct or secondary relevance to the project included the following.

2021 Marc Anderson, “Commentaires sur la Charte OLKi” Intelligence Artificielle et Vie Privée (dans le cadre des projets DigiTrust et Open Language and Knowledge for Citizens (OLKi) France (June 10)

2021 Marc Anderson and Karën Fort, “Human Where? A New Scale Defining Human Involvement in Technology Communities from an Ethical Standpoint” AI-MAN (ICT-38) Projects Cluster Workshops Series Online – “Ethical and Legal Issues of Artificial Intelligence in Manufacturing” (25 November)

2022 Marc Anderson, “Plenary Presentation of Key Findings - Breakout Session 12: Psychological Approach for Data Labelling”, AI for Future Manufacturing Theme Development Workshop, CLAIRE Confederation of Laboratories for Artificial Intelligence Research in Europe (May 10)

2022 Marc Anderson, “Some Ethical Reflections on the EU AI Act”, IAIL 2022 - Imagining the AI Landscape after the AI Act, 1st International Workshop on Imagining the AI Landscape After the AI Act, Vrije Universiteit Amsterdam, Amsterdam (June 13)

2022 Marc Anderson and Karën Fort, “Ethical Internal Logistics 4.0: Observations and Suggestions from a Working Internal Logistics Case”, 12th International Workshop on Service Oriented, Holonic and Multi-Agent Manufacturing Systems for Industry of the Future – SOHOMA’22 (September 22-23)

2022 Marc Anderson, “Is the Future of AI Ethics Interdisciplinary?”, Where AI Ethics Should Go, von Weizsäcker Zentrum (University of Tübingen) and the Archives Henri Poincaré, Germany (June 30-July 1)

2022 Marc Anderson, “Ethique dès la conception dans un projet d'informatique industrielle : l'exemple du projet AI-PROFICIENT,” ISET - CRAN, (Ingénierie des Systèmes Éco-Techniques - Research Center for Automatic Control, Nancy), France. Invited Seminar (December 16)

2023 Marc Anderson, “AI as Philosophical Ideology: A Critical look back at McCarthy’s Program,” Philosophy in Technology Workshop 2nd Edition. Wrocław University of Science and Technology; the Pontifical University of John Paul II, and the Polish Academy of Arts and Sciences (April 28-29)

2023 Marc Anderson, “Operationalizing AI Ethics in Industry 4.0,” The Future is WOW 2023: Bringing AI Technology to the Production Line. Mechelen, Belgium (June 8, 2023)

2023 Marc Anderson, “Exploring the Idea of Ethical Sustainability for Digital Manufacturing” in “Sustainability for the digital manufacturing era,” 13th International Workshop on Service Oriented, Holonic and Multi-Agent Manufacturing Systems for Industry of the Future – SOHOMA’22, Annecy, France (September 28-29)

2023 Izaskun Fernandez, Kerman Lopez de la Calle, Eider Garate, Regis Benzmuller, Melodie Kessler, Marc Anderson, “Human-Feedback for AI in Industry,” in the Sixteenth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services, CENTRIC 2023, Valencia, Spain (November 13-17) (Izaskun Fernandez as presenter) [forthcoming]

2023 Marc Anderson, “AI Ethics and the Lessons of History,” in the 2nd International Conference on the Ethics of Artificial Intelligence (2ICEAI), Porto, Portugal, (November 28-30) [forthcoming]

Part 3: Ethical Recommendations Review

3.1 Implementation Results of Ethical Recommendations – Ethics Team Assessment

The following graphs and charts display our results for implementation of ethical recommendations. Results were assessed by the Ethics team in an ongoing manner. As noted earlier, we kept a running record of efforts toward implementation. There is a qualitative aspect to the assessment here since we had to plan as to what constituted a full implementation. Some recommendations had several aspects.

Our position was to take anything short of full implementation as partial implementation unless nothing at all was done. Thus, partial implementation covers a range, from 'mostly implemented' to a bare effort at some aspect of implementation. The latter stance is in part necessitated by having to adopt an attitude which presumes good faith on the part of the consortium partners, i.e., wherever a recommendation could not be explicitly verified by the ethics team we must depend upon second hand reports or confirmations by consortium partners members that they are carrying out the recommendation.

But the qualitative aspect is moderated by the fact that commitments were made in deliverables which relate to some recommendations, some recommendations addressed additions to deliverables directly which we could verify in the deliverables themselves, and we can see the changes in design solution related to ethical recommendations as they occurred and were discussed in technical meetings which we attended. In other words, despite the qualitative aspect, there was much that we could verify at first hand.

3.1.1 Overall Implementation Results

In Figure 1 below we indicate the overall results as a proportion of the total of 121 recommendations which were kept for the ethics team assessment, for each of our three outcomes: fully implemented, partially implemented, not implemented.

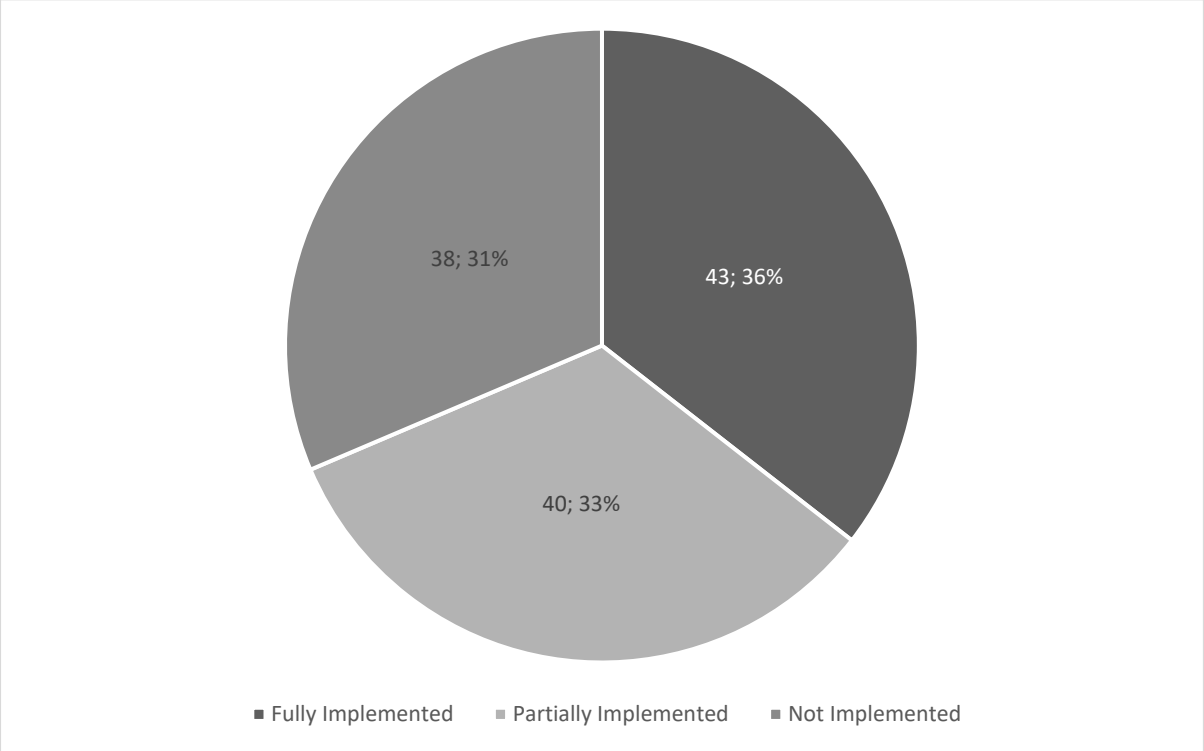

Figure 1: Overall Results of Ethical Recommendations as Assessed by Ethics Team

3.1.2 Implementation Results by Category

In Table 1 below we give the results according to category, as assessed by the ethics team, for each of our three outcomes as a proportion of the total recommendations under that category.

Recommendation Category	Fully implemented	Partially implemented	Not implemented	Total in Category
Protocol	0	5	6	11 (+2 NA)
Human centering	11	10	2	23 (+0 NA)
Design	6	6	2	14 (+1 NA)
Insufficient specs	2	4	4	10 (+0 NA)
GDPR	1	1	0	2 (+2 NA)
Responsibility	7	7	7	21 (+1NA)
De-anthropomorphizing	3	0	0	3 (+1 NA)
Simplification	4	2	1	7
Verify effects	1	0	2	3 (+1 NA)
Timeliness	0	1	3	4 (+0 NA)
Valorize experience	1	1	1	3 (+1 NA)
Ethical rewording	6	0	0	6
Workload	1	2	6	9 (+0 NA)
Evaluation	0	0	3	3
Training	0	1	1	2
Total	**43**	**40**	**38**	**121 (+9 NA)**

Table 1: Overall Results by Category

In Figure 2 we give the overall results by category, as assessed by the ethics team, in chart form.

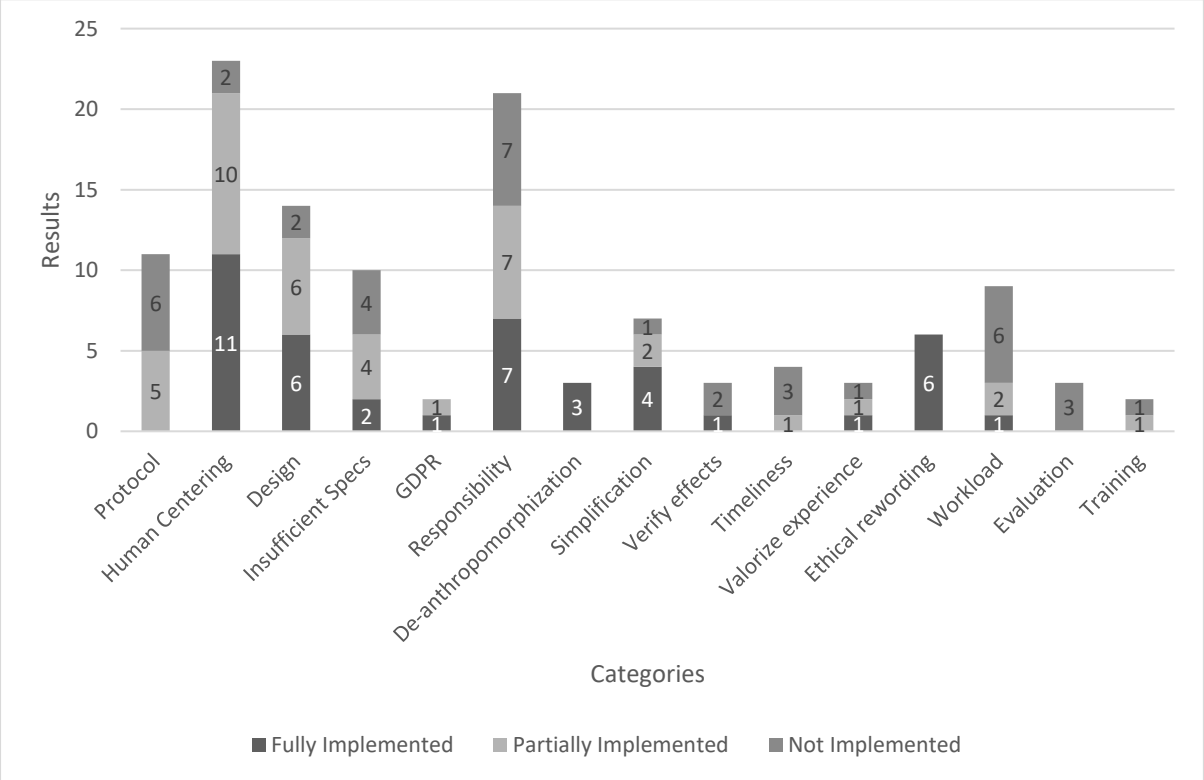

Figure 2: Results of Ethical Recommendations by Category as Assessed by Ethics Team

In Figure 3 we give the same information in chart form, with the categories most fully implemented as a percentage of the total recommendations in the category in descending order from left to right.

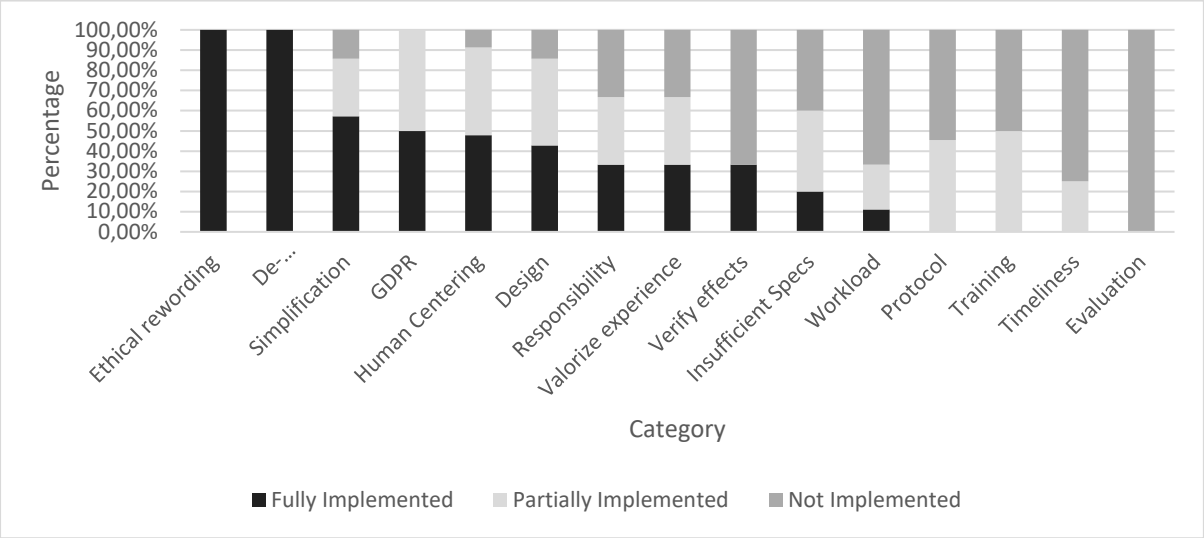

Figure 3: Results of Ethical Recommendations by Category as Assessed by Ethics Team - Percentage of Total

3.1.3 Implementation Results by Category and Partner

In Figure 4 below we give as a heat map the recommendation implementation results, as assessed by the ethics team, both categorized and sorted according to consortium partner(s) we judged to be responsible for carrying them out. For each recommendation category in intersection with a consortium partner, the saturation level in blue or red indicates the proportion of fully, partially, or non-implemented recommendations carried out by that partner relative to the total number of recommendations given to the partner in that category. Note that since some recommendations were the responsibility of several partners the total number of recommendations for a category does not correspond exactly to those given in the table and figures above because responsibility for some recommendations was formally assigned to more than one partner, and sometimes the same recommendation was carried out (or partially) by one responsible partner but not by the other(s).

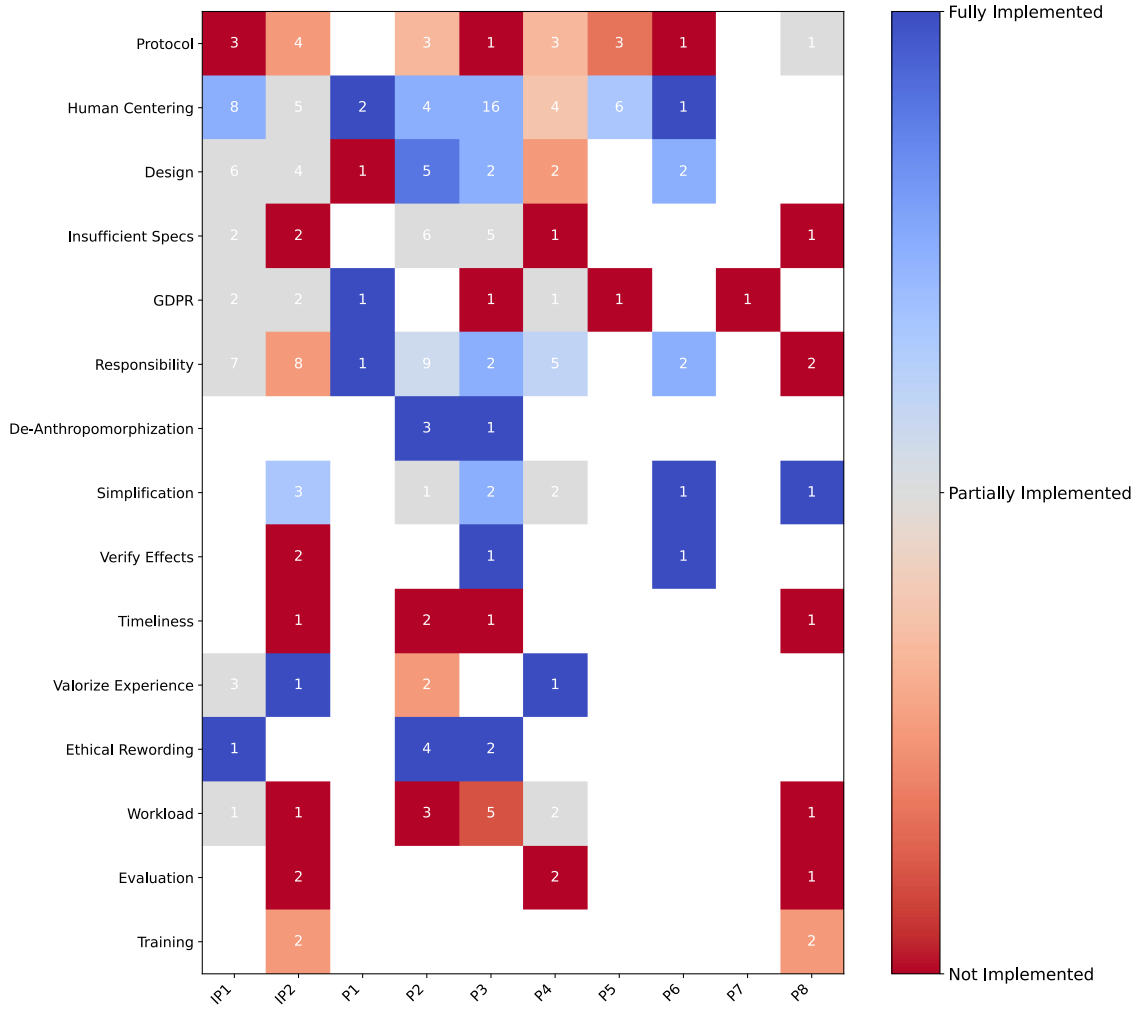

Figure 4: Result of Ethical Recommendations by Category and Project Partner(s) as Assessed by Ethics Team

3.2 Implementation Results of Ethical Recommendations – Partner Assessment

Below we give the Implementation Results as assessed by the AI-PROFICIENT consortium partners. These results, as well as the above implementation results as assessed by the ethics team, will be discussed further below in Section 3.3.

Our procedure in arriving at the partner assessment was as follows. Once the ethics team had completed its assessment, the assessment was made available on an online worksheet and the consortium partners were given five weeks to review and agree or disagree with the ethics team assessment of implementation results. If they agreed they need do nothing. If they disagreed, they could give their own assessment of the implementation status, either upgrading or downgrading for each recommendation. They were also given the opportunity to give reasons for non or partial implementation, or updates in case they foresaw implementation forthcoming.

3.2.1 Overall Implementation Results

In Figure 5 below, the overall results as a proportion of the total of 108 recommendations which were kept by the project partners for assessment, are indicated, for each of the three outcomes: fully implemented, partially implemented, not implemented. Note that by the time the partner results assessment was carried out, more recommendations were judged NA, in particular all the recommendations for one UC which formally abandoned by the project partners.

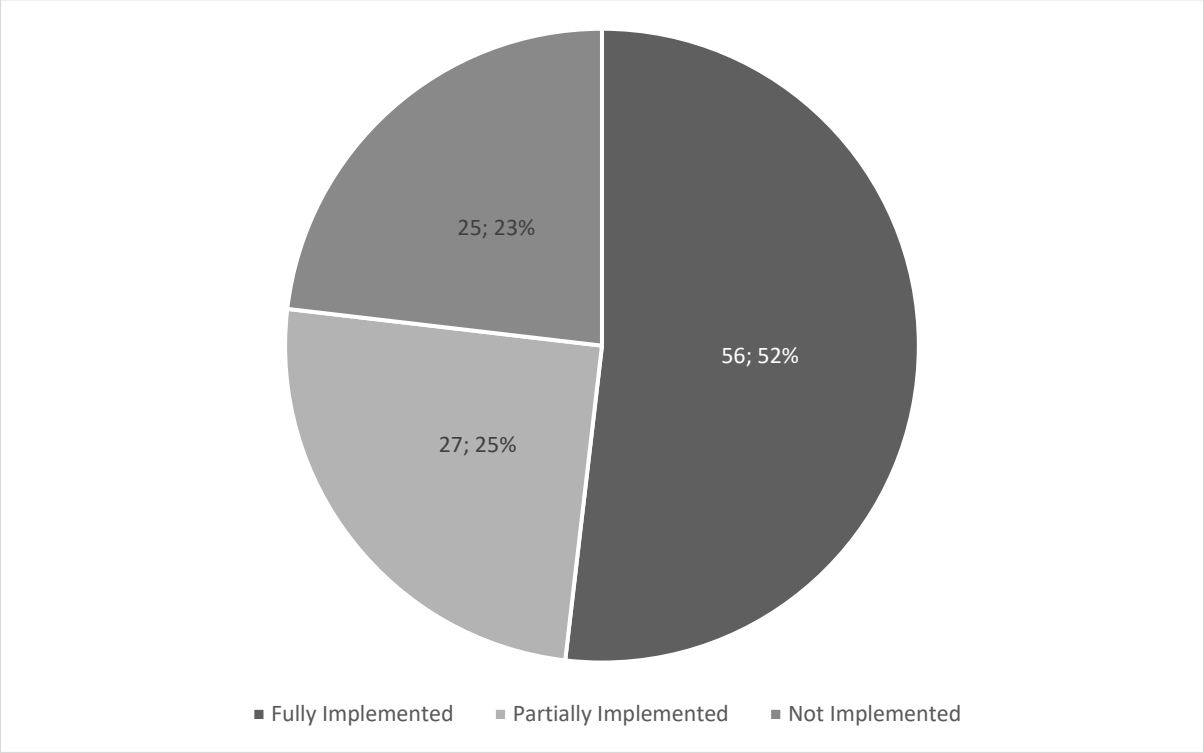

Figure 5: Overall Results of Ethical Recommendations as Assessed by Project Partners

3.2.2 Implementation Results by Category

In Table 2 below we give the results according to category for each of our three outcomes as a proportion of the total recommendations under that category.

Recommendation Category	Fully implemented	Partially implemented	Not implemented	Total in Category
Protocol	2	4	5	11 (+2 NA)
Human centering	17	4	0	21 (+2 NA)
Design	5	5	1	11 (+4 NA)
Insufficient specs	2	1	3	6 (+4 NA)
GDPR	1	1	0	2 (+2 NA)
Responsibility	7	7	6	20 (+2 NA)
De-anthropomorphizing	3	0	0	3 (+1 NA)
Simplification	6	1	0	7
Verify effects	1	2	0	3 (+1 NA)
Timeliness	1	0	2	3 (+1 NA)
Valorize experience	2	0	1	3 (+1 NA)

Ethical rewording	6	0	0	6
Workload	3	1	3	7 (+2 NA)
Evaluation	0	0	3	3
Training	0	1	1	2
Total	**56**	**27**	**25**	**108 (+22 NA)**

Table 2: Overall Results by Category as Assessed by Project Partners

In Figure 6 below, the results as assessed by project partners, according to category are given, for each of the three outcomes as a proportion of the total recommendations under that category.

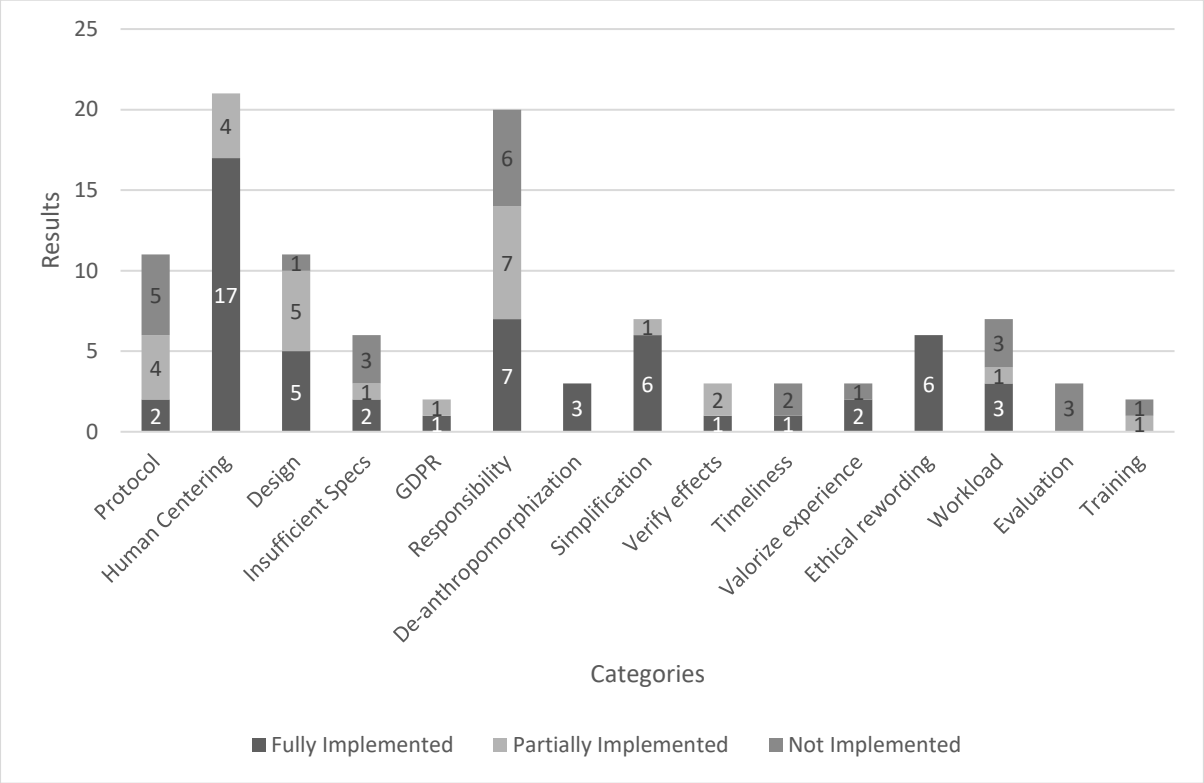

Figure 6: Results of Ethical Recommendations by Category as Assessed by Project Partners

In Figure 7 below, the same information is given in chart form, with the categories most fully implemented as a percentage of the total recommendations in the category in descending order from left to right.

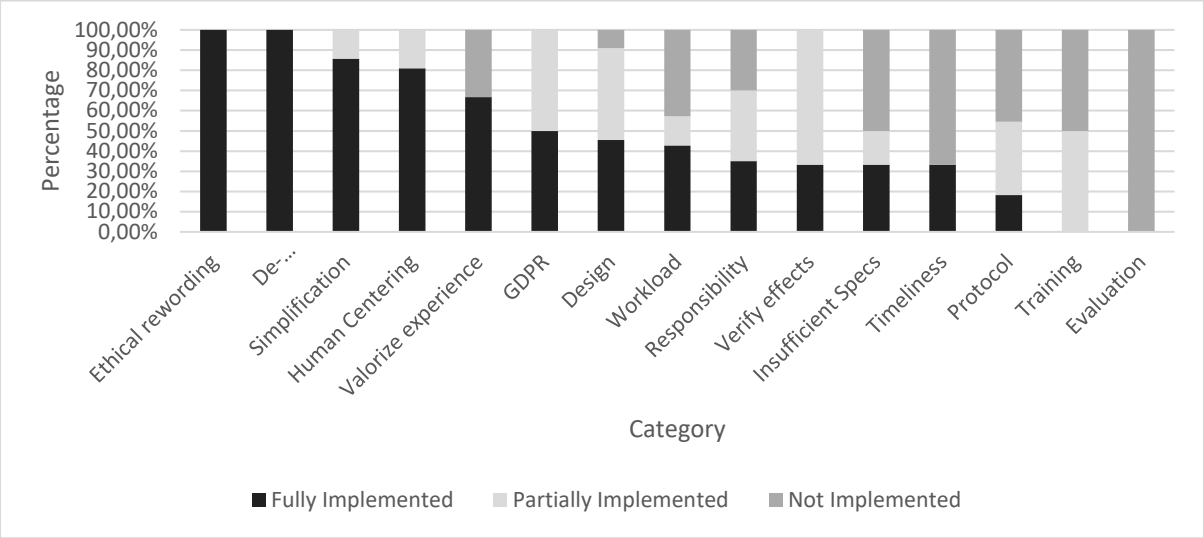

Figure 7: Results of Ethical Recommendations by Category as Assessed by Project Partners - Percentage of Total

3.2.3 Implementation Results by Category and Partner

In Figure 8 below, we give, in heat map format, the results according to category, as assessed by the project partners, for each of our three outcomes but now sorted additionally by responsible partner or partners (anonymized). The colour intensity of each square indicates the combined proportions of fully, partially, or not implemented, relative to the total recommendations under that category – number given in the square – assigned to that partner. Note that here the total number of recommendations for a category does not correspond exactly to those given in the table and figures above because responsibility for some recommendations was formally assigned to more than one partner, and sometime the same recommendation was carried out (or partially) by one responsible partner but not by the other(s).

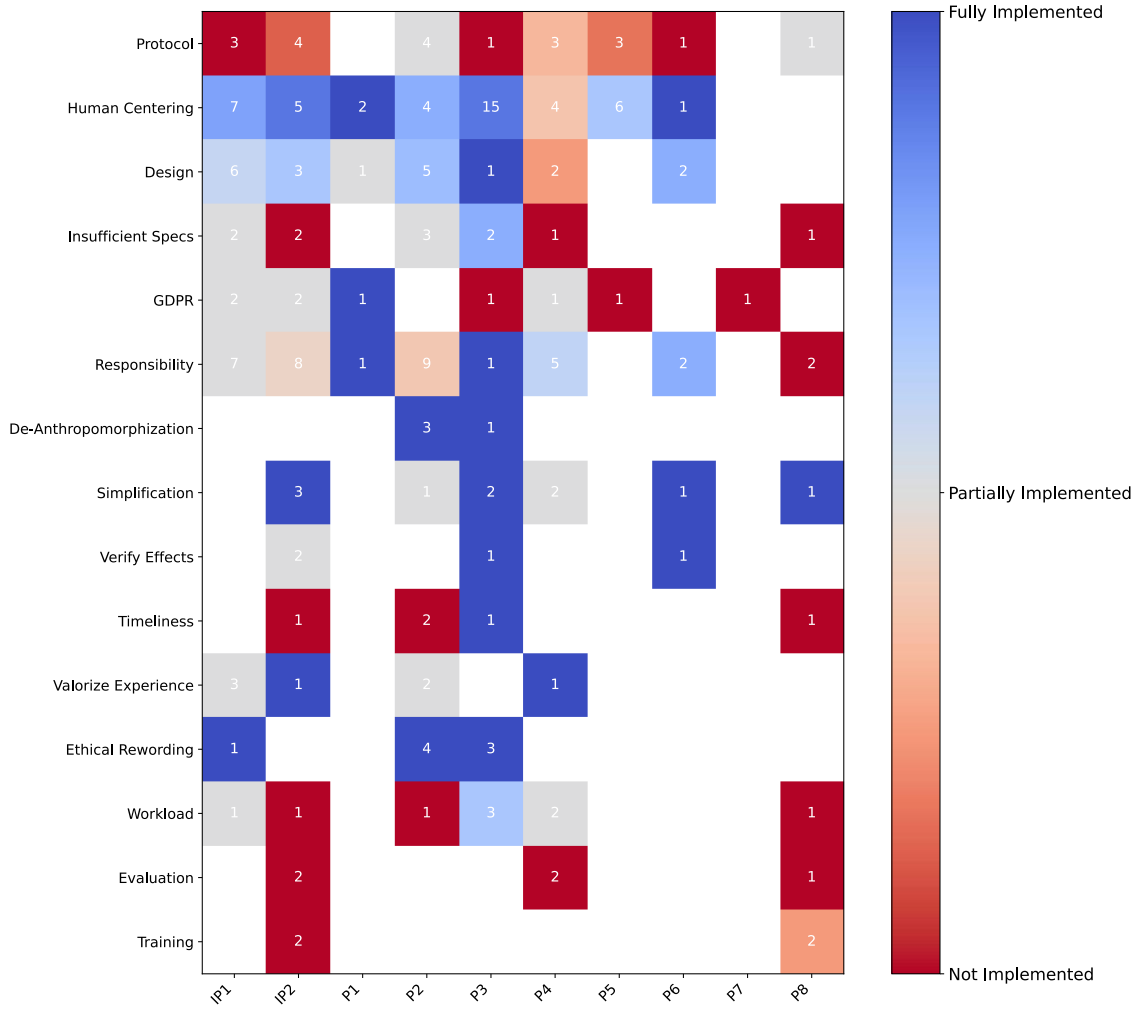

Figure 8: Results of Ethical Recommendations by Category and Project Partner(s) as Assessed by Project Partners

3.3 Discussion of Implementation Results of Recommendations

Below we discuss the implementation results of the ethical recommendations given, first in terms of what the ethics team observed and then regarding the significance of those observations.

3.3.1 Comments on Methodology

Given that the ethics team's approach is one which evolves, the implementation results recorded above are an indication of a point near to but before the formal end of the project. We cannot wait until after the end of the project to make a final assessment, but although the results may change somewhat before the end of the project, we have good reason to think that our ethical implementation results are a reasonably accurate picture of the success of our ethics by design approach. We thus have a basis

from which to evaluate the strengths and weakness of our approach and to provide some insights into what remains to be done in future projects.

We should stress that the implementation results are entirely due to the partner's efforts. The ethics team did not push actively for implementation in any sense of regular 'moralizing.' Rather we clarified the UC contexts, gave advice to the partners in the form of the recommendations, inquired and updated from time to time regarding ethical progress, participated continually in design discussions, and remained constantly available to the partners for consultation on ethical issues.

3.3.2 Observations Regarding Overall Results

As presented in Figure 1, with regard to the total number of recommendations given, after subtracting those judged NA, 36% were partially implemented, and 33% were fully implemented. 31% of the recommendations remained un-implemented. There is a minor chance that both percentages might increase slightly by the end of the project, considering that the completion of the evaluation Work Package of the project includes Deliverable 6.4. Since the deliverable must undergo review and the review will normally only convene while the project is ongoing, we cannot get final post project ethical assessment results, however.

The partner overall assessment, Figure 5, stands at 52% of recommendations fully implemented, 25% partially implemented, and a further 23% not implemented.

3.3.3 Observations Regarding Results by Category

Observing implementation results by category, Figure 2, we see that they contain two clear extremes, although this is offset by the fact that these categories contain relatively small numbers of recommendations. Under the categories Evaluation and GDPR, no recommendations were carried out. Conversely, under the category of De-anthropomorphizing of descriptions of AI in deliverables and under that of Ethical rewording, all recommendations given were carried out. As is clear in Figure 6, the partner assessment did not change this.

Further, as assessed by the ethics team, Figure 2, well over half of the recommendations within the three categories with the largest total number of recommendations were implemented either partially or fully. Taken in descending order of number of recommendations, these were Human Centering, Responsibility, and finally Design. By comparison, in Figure 6, the Human Centering category was the only category which differed considerably with regard to the project partner assessment. In that category a number of recommendations which the ethics team had judged to be partially implemented were upgraded to being fully implemented.

Looking at the heat map, Figure 4, of the ethics team assessed results of implementation sorted by partner responsible for the recommendation, as well as by category, we note that a high degree of full or partial implementation follows upon the conjunction of certain partners with certain categories. As examples of this, referring to the data used to generate the heat map colour intensity, partner 2 (P2) partially or fully implemented 6 of 9 Responsibility recommendations and 5 of 5 Design recommendations, while Industrial partner 1 (IP1) fully or partially implemented 8 of 8 of the Human Centering recommendations which they were responsible for, and Partner 3 (P3) fully or partially implemented 16 of 16 in the same Human Centering category of recommendations. The opposite applies as well, with some partners implementing relatively few of some categories, e.g., Industrial partner 2 (IP2) implemented only 3 of 8 Responsibility recommendations.

The project partner assessment of results by partner and category, Figure 8, shows some shift in the heat map toward the fully implemented or blue colour intensity, and shows this for some categories and for some partners. But some category rows and partner columns remain unchanged, e.g., the results of P8 and the Evaluation category row. Also, notably, similar general trends can be seen in the partner assessed heat map, i.e., very strong implementation within the Human Centering and Design rows, accompanied by relatively strong implementations overall by Partner 2 (P2) and Partner 3 (P3).

3.4 Discussion of Observations

Below the observations regarding ethical recommendation implementation results are discussed, first as to overall results, then results by category, and finally as to results by category and partner. Insights are also provided regarding the differences observed between the ethics team and partner assessments.

3.4.1 Overall Results

The main takeaway for the ethics team is that *ethics can successfully by operationalized at ground level in an Industrial AI context* by taking an approach which embeds the ethicist in the process of technological development. The success rate as assessed by the ethics team and as assessed by the project partners is also quite consistent. The partners did not simply upgrade all or most implementation results to fully implemented, as they were perfectly free to do. Even though the partner assessment is more optimistic than that of the ethics team, one might also have expected project partners to always assess as good or better than the ethics team, but this was not the case. Several partners downgraded some recommendation results from full to partial or even to not implemented, despite the opinion of the ethics team.

In the partner assessment of overall results, the upgrade from 36% to 52% for fully implemented came equally from partially and not implemented portions. The upgrade of certain recommendations which led to a higher percentage seems to have come mainly – as indicated by the partners – from the assumption that the full implementation of those recommendations would be achieved by the final AI services deployment phase. Whether that is so, we will not be able to objectively verify. Nonetheless the ethics team is confident that our assessment is an essentially accurate picture of the ethics by design achievement of the project. By including the partner assessment for comparison, we leave room for arguing that a more objective figure for the results lies somewhere between the two assessments. We also note that the formal abandonment of one UC shortly after the ethics team assessment, made up for a good deal of the overall change of percentage, because a number of ethical recommendations connected to that UC, which had not been fulfilled, were rendered NA, thus raising the overall success rate in the partner assessment.

Given the overall results, one could still ask whether this is a practical rate of success for ethics. Two responses at least could be offered to this. First, *practicality for ethics is arguably complex and unique by comparison with other domains*. Thus, regardless of the exact numbers, the approach used in the project demonstrates a practicality appropriate to ethics as historically understood, in terms of prompting the partners to consider ethics at the level of the shop floor and technical developer design level, and also in inspiring ethical action relative to the recommendations. The approach has habituated the technical developers and industrial partners to think ethically, and to look beyond mere technical aspects of solutions in the project. The ethics team hopes this habit and attitude will remain with project partners beyond the project. Insights into the relative willingness or unwillingness of project partners to implement different categories of recommendations – e.g., Human Centering and Evaluation – are another gain, and one we can base future studies upon and refine the approach to address.

Second, *if being practical is measured similarly to quantitative evaluations of technical success, then the expectations for improvement for an operationalized ethics should remain comparable to those for technical success*. Under its technical categories, the AI-PROFICIENT project proposal aims for efficiency improvements ranging between 1-9%. Taking this as a guide, if the ethics approach is to be evaluated by similar standards, then overall result outcomes of 33% partially implemented and 36% fully implemented are at least as good as the project technical outcomes. Modest gains of 10-25% - e.g., in efficiency (ZVEI, 2012) are typically envisioned as practical in industrial process and manufacturing automation, rather than aiming at complete transformations of manufacturing processes. It would not be fair to demand more of practical ethical results insofar as held to similar quantitative evaluation standards and applied in similar contexts.

Viewing ethics pragmatically, the above paradigms should not be exclusive. A quantitative tending aspect for ethics can co-exist alongside a more ideal and qualitative aspect, with the latter acting as a simple map to orient the development of the former.

3.4.2 Results by Category

The results by category show that the most difficult recommendations to get implemented are Evaluation, where the goal is to set user centered quantitative benchmarks for error and reliability, and Workload where the goal is quantitatively estimating aspects of new AI service-related tasks which the worker will have to take on.

One possibility which accounts for the difficulty with Evaluation recommendations is that aiming to formalize a benchmark may imply that *the system is not and may never be reliable enough to use in the end, so that we should not use it*. The prevailing technology paradigm, however, is that a technical solution is always possible. The latter notion, which dampens ethical engagement, is heavily embedded in the tech developer worldview, as argued by both Clark and Lischer-Katz (2023) and Avnoon et al. (2023). Hagendorff (2022), notes that this issue of deciding not to use technology is rarely put on the table in AI ethics, thus the developer milieu has an even lower chance of being exposed to or considering it.

Estimating quantitative figures in advance is clearly not impossible however, because our project partners carried it out for the most implemented categories of Human Centering, Design, and Responsibility. When one considers that avoiding quantitative evaluation is also counter-intuitive in light of tech developer partiality toward technical solutions, then it seems that the problem with regard to Evaluation type recommendations does indeed stem from the ‘technology always finds a way’ outlook, in which technology can never be inappropriate to a context.

Difficulty implementing Workload recommendations, while it has a crossover with the above issue, appears to be related to the historical and institutional frameworks around work, more than in the attitudes of the technology development partners. Requesting workload estimates tends to create further work, and such estimates may be difficult to arrive at. Yet both of these issues appear to be soluble.

The larger problem may be that workload recommendations bring attention to the fact that the result of developing and integrating more technology quite often simply *adds more work elsewhere* (Crawford, 2021). Often this is work which is not factored in, despite the often-stated goals of developing the technology to lessen the workload. Evidently, a mere shifting of work elsewhere, without accounting for it, *practically* contradicts such stated goals.

It does not *actually* contradict it however if it is never formally acknowledged so that the shop floor worker is made aware of it. The ethics team’s Workload recommendations often attempted to get such an acknowledgement. If the shop floor worker becomes aware of it, it then tends to require some justification, and possibly legal justification with regard to work contracts.

This may explain why there was a tendency in some of the instances where Workload recommendations were implemented, for some project partners to simply take on the extra workload themselves, or to shift the extra work from shop floor level to management. Such a result is good in one sense, insofar as clarifying the workload through recommendations has both a neutral result with regard to the shop floor worker and indicates a way in which applied ethics can engage the issue. But it is a stopgap measure rather than a satisfying result. The latter would need a context founded public rethink, in good faith, regarding the reasons we are deploying technologies in many cases.

The above also illustrates a significant problem faced in operationalizing ethics in industry and beyond: that law and ethics have a difficult relationship and what is legal may not be ethical. To impose upon a worker in a contract, may not be ethical in terms of the contradictions involved with regard to a company’s publicly stated aims for technology adoption. It may be perfectly legal however, or legal because unchallenged. In that case, the legal aspect will tend to win locally, unless and until the ethical aspect gains a wide enough hearing to begin to question and then change the law.

3.4.3 Results by Category and Partner

With regard to the heatmaps, where results are broken down according to partner as well as category, we can see that, allowing for the categories just discussed, some of the partners are more willing than others to implement recommendations. This held in terms of industrial partners, viewed separately, as well, where one of the partners has a proportionally higher rate of full or partial implementation that the other partner. One developer partner had little interest in ethical recommendations and suggested that they 'didn't really understand' the ethical aspect. Still other partners showed mixed engagement, depending upon the issue in question or on their interest in particular UCs. The responses to ethical recommendations were sometimes ambiguous for these partners: 'we will carry out *x* if it is efficient.' This indicates that efficiency took precedence over ethical concerns, and also that if implementation proved difficult there remained a 'way out.'

After adjusting for the number of recommendations, the assessment by the ethics team and by the partners, both showed Human Centering to have the best results in the heat maps. Considering the users (operators and process engineers) and developing the services in collaboration with them were things that were readily implemented. Anecdotally – in discussion with the industrial partner engineers – the worker tendency to not use a new service or technology if it does not work properly, probably helped this result for Human Centering.

But the primary reason for a good result with Human Centering may be that such recommendations tend to be positive. They require additional work, but they do not obstruct lines of technical developments which the partner is pursuing. One can operationalize them by giving options to the works, clarifying who does what, discussing with process engineer and work team heads, and developing short surveys to uncover the worker's background knowledge and expectation so as to adjust accordingly. The ethics team strove to make all recommendations positive, but Human Centering recommendations are particularly well suited to this positive approach.

Partner 3 (P3), had the best record of full or partial implementation, as noted above, proportional to the number of recommendations they were responsible for. They were closely followed by Partner 2 (P2). A number of individuals within P3 engaged with us in implementing the recommendations over various deliverables and different issues. Also, interestingly, as can be seen in Figure 2 and Figure 6, and perhaps because of the ethical engagement in question, a good portion of the upgrade from partially to fully implemented in the partner assessment, was made by P3. In other words, *the partner, which was most interested in ethical engagement, was also the partner most inclined to view their engagement as more successful*.

A strong commitment in the partners who seriously engaged with the ethics aspect of the project is indicated by the higher proportions of successful implementation by those partners who were responsible for the largest number of recommendations. The ethics team's subjective assessments of different partner reactions to recommendations generally, back this up: those partners who had the highest proportions of full or partial implementation were also those who tended to contact the ethics team on a regular basis for clarifications or extra meetings to discuss potential ethical issues in UCs which they led. Those partners even seem to have welcomed an ethical engagement, rather than viewing it as a bother.

3.5 Methodology for Conversion of Deliverable 6.4 Ethical Recommendation Implementation Results into General Evaluation Results of Deliverable 6.2

One stated goal of Deliverable 6.4 at proposal stage, was to assess AI-PROFICIENT in relation to ethics. This included providing recommendations and evaluating positive and negative impacts at shop floor level (AI-PROFICIENT, Annex 1, 33). The latter suggested post deployment impact analysis. The ethics team approach has been to embed ethics by design from the beginning however, rather than to wait and 'correct' potential negative impacts after the fact. The post impact approach would be too late. It would not contribute to correcting ethical issues – except for future projects – and it would not make good use of time that could be used to generate evolving solutions to those ethical issues.

The fact that AI-PROFICIENT deployment has been somewhat delayed confirms our approach. For most of the use cases – except those which were formally abandoned – deployment had been completed by the end of project, but not early enough to allow a final ethics evaluation after deployment. We suggest that the quantitative analysis of recommendation implementation results, as given above, is a practical substitute for post deployment analysis. In other words, the evolving recommendations given engaged the impacts before deployment and the ongoing partner responses to them allowed the ethics team to generate insights from the project which are equivalent, practically, to those of the original aim.

The differences and integrations of the ethics team approach with the more technical evaluation of the project are described below.

3.5.1 Description of Deliverable 6.6 Evaluation Methodology

Deliverable 6.6 methodology combines evaluations with regard to production level key performance indicators (KPIs) to arrive at a percentage of improvement. Along with this, end user requirements, functional requirements, and user experience are evaluated. User related evaluations are measured in terms of KPIs of usability, usefulness, etc., disclosed through user surveys begun after initial deployment. For the latter, there are a number of surveys with questions directed at the end users. These components are to be combined with the ethics evaluation component in order to arrive at a validation for each use case.

In Deliverable 6.6 the ethics component of evaluation was envisioned similarly, as a number of questions to be asked under five categories. But these categories grew out of the actual issues raised and recommendations given during the project by the ethics team. They are thus grounded in the ethics team's bottom-up approach, rather than imposed from above as they would be in the HLEG guidelines paradigm.

The ethics team has therefore decided to convert the implementation results data from our approach into the categorized formula for ethics evaluation described in Deliverable 6.6., rather than asking these questions over again at a more general level. In other words, *the questions have already been asked and answered in a more specific and practical way through the analysis of ethical issues, generation of recommendations, and responses of the partners in implementing them*. In this way we keep the paradigm of the messier ground level ethics by design by the embedded ethicist at the forefront, while still integrating with the WP6 evaluation paradigm.

3.5.2 Method of Conversion of Results

The Ethics Team has converted the *recommendation implementation results as assessed by the project partners*, rather than the results as assessed by the ethics team. The reasoning here is first, that the actual implementation results lie somewhere between the ethics team results assessment and the results as assessed by the project partners, and since we cannot do better objectively, we have chosen to ere on the side of the partner assessment, since the results are to be combined into the general

project evaluation. Second, and related, this choice brings the ethics results for the overall evaluation more in line with the final choices of the project partners regarding UC development, in particular the formal abandonment of one UC late in the project, which shifted the 13 related recommendations to N/A status.

In consultation and agreement with the WP6 leader, a change has also been made with regard to scoring the results. Thus, whereas Deliverable 6.6 envisioned a simple yes or no, or fully implemented or not implemented result for overall evaluation, scoring 1 or 0, respectively, we have thought it better to give a score of .5 for partial implementation, which is finer grained and a better reflection of actual results of specific recommendations.

On the basis of the above, the ethics team has gone over all recommendations and decided which of the five categories decided upon in Deliverable 6.6 a recommendation should go into, a choice which was self-evident in most cases as the five recommendation categories were formulated by the Deliverable 6.6 contributors on the basis of the specific ethics team recommendations. Finally, the formula agreed upon in Deliverable 6.6 was carried out.

3.5.3 Ethical Recommendation Evaluation Results for Deliverable 6.2

Below in Table 3 we give the weighted calculation for each UC of ethical recommendation results as gathered into the five general groups decided upon in Deliverable 6.6. The totals for each UC then give the percentage compliance for the UCs to be carried forward into the general project validation analysis of Deliverable 6.2.

Use Case	**[(ETH_IDX Result) + … + (ETH_IDX Result)] * 100/NQ * (Total WEIGHT of Group 1)** **(1 = full implementation, .5 = partial implementation, 0 = not implemented)**	**Totals** **(rounded up to hundredths)**	**Final Result = % Compliance** **(rounded up)**
ETH C UC 2			
Group 1 GAI	[1 + 0] * 100/2 * 2/7	14.29	
Group 2 ERRH	[.5] * 100/1 * 1/7	7.15	
Group 3 WkL	[.5] * 100/1 * 1/7	7.15	
Group 4 IN	[1] * 100/1 * 1/7	7.15	
Group 5 EtbD	[1 + 1] * 100/2 * 2/7	28.58	
Use Case Total:		64.32	65
ETH C UC 3			
Group 1 GAI	[1] * 100/1 * 1/2	50	
Group 2 ERRH	NA		
Group 3 WkL	NA		
Group 4 IN	[.5] * 100/1 * 1/2	25	
Group 5 EtbD	NA		
Use Case Total:		75	75
ETH C UC 5			
Group 1 GAI	[.5 + 0 + 1] * 100/3 * 3/13	11.54	
Group 2 ERRH	[0] * 100/1 * 1/13	0	
Group 3 WkL	[0 + 0 + .5] * 100/3 * 3/13	5.13	
Group 4 IN	[0 + .5 + 0 + .5] * 100/4 * 4/13	7.70	
Group 5 EtbD	[1 + 1] * 100/2 * 2/13	15.39	

Use Case Total:		39.76	40
ETH C UC 7			
Group 1 GAI	[.5] * 100/1 * 1/7	7.15	
Group 2 ERRH	[1] * 100/1 * 1/7	14.29	
Group 3 WkL	[1] * 100/1 * 1/7	14.29	
Group 4 IN	[1 + .5 + 1] * 100/3 * 3/7	35.72	
Group 5 EtbD	[.5] * 100/1 * 1/7	7.15	
Use Case Total:		78.6	79
ETH C UC 10			
Group 1 GAI	[.5 + .5 + .5] * 100/3 * 3/14	10.72	
Group 2 ERRH	NA		
Group 3 WkL	[1 + 1 + 1 + .5] * 100/4 * 4/14	25	
Group 4 IN	[1 + 1 + 1 + 1] * 100/4 * 4/14	28.58	
Group 5 EtbD	[1 + 1 + 1] * 100/3 * 3/14	21.43	
Use Case Total:		85.73	86
ETH I-G UC 1			
Group 1 GAI	[0 + .5 + 0 + 1] * 100/4 * 4/13	11.54	
Group 2 ERRH	[0 + .5 + 0 + 0] * 100/4 * 4/13	3.85	
Group 3 WkL	[0] * 100/1 * 1/13	0	
Group 4 IN	[1 + .5 + 0 + 0] * 100/4 * 4/13	11.54	
Group 5 EtbD	NA		
Use Case Total:		26.93	27
ETH I-G UC 2			
Group 1 GAI	[1 + 1] * 100/2 * 2/18	11.12	
Group 2 ERRH	[.5 + .5 + 0 + 0 + 0] * 100/5 * 5/18	5.56	
Group 3 WkL	[1 + 1 + 1 + 1 + 0] * 100/5 * 5/18	22.23	
Group 4 IN	NA		
Group 5 EtbD	[1 + 1 + .5 + 1 + 1 + 1] * 100/6 * 6/18	30.56	
Use Case Total:		69.47	70

Table 3: Use Case Level Ethical Recommendation Compliance Calculation and Result

3.5.4 Discussion of Results

The most successful ethics compliance was achieved in ETH C UC 10. That UC developed a Quality Analysis tool which included explainable AI. A relatively high proportion of Human Centering recommendations were made in the UC (4/13), and all were implemented fully except one which was partially implemented, and this aligns with the earlier observations regarding that category.

There may be some use to future observers of the AI-PROFICIENT ethics by design method to understand the recommendation results under the more general categories of the general evaluation. The full names of these categories can be found in Deliverable 6.6. But it should be remembered, as

has been argued throughout, that the ground up approach of disclosing issues is the best practical way to engage the general categories and advance and operationalize related recommendations. If this is kept in mind, then the more general categories could perhaps serve as useful guide maps to help the process.

Part 4: Insights for Future Projects

In Part 4: Insight for Future Projects we review and compare parallel ethics approaches in other projects, as well as distilling insights from the AI-PROFICIENT ethics approach, discussing limitations of the approach, and advancing suggestions for complementary future research.

4.1 Review and Comparison of AI-PROFICIENT Ethics Approach with approaches of other projects in the ICT-38 Cluster

The ICT-38 cluster of projects had the goal of integrating cutting edge AI technologies into the manufacturing domain. Projects were generally expected to have an ethical component as stated in one of the specific impact contributions included in the project call, namely: “structurally enhanced research and innovation capacities in this area, through structured transdisciplinary expertise, research and practice networks of the highest ethical and methodological standards across Europe,” (European Commission Horizon 2020 Work Programme 2018-2020, Part 20. Cross Cutting Activities, pg 149, 2020). More specifically “Ethical principles, as expressed by the high-level expert group on Artificial Intelligence should be followed and recommendations for instantiation in the manufacturing domain should be developed,” (EU Commission, Funding & Tender Opportunities, AI for Manufacturing, ICT-38-2020, 2019). The ICT-38 projects were also envisioned as building upon the earlier but overlapping projects of the ICT-26 call to realize a European AI-on-demand platform, which issued in the AI4EU AI-on-Demand Platform. Consequently, the ethics team considers it appropriate to review the ongoing ethics approaches of the other ICT-38 projects, and then compare their approaches with our own. By doing so we hope to better integrate the AI-PROFICIENT approach with the spirit of the higher-level research goal of the cluster, outline the specific outcomes of our approach, and provide material for reflection upon what types of future research might be needed along these lines.

4.1.1 Assistant

The Assistant project adopts an ex-ante and ex-post approach, with the goal of discussing and formulating responsibility in particular. They review theoretical streams engaging responsible AI, noting that ethical AI which falls under a group of general discourses are most often separated from ‘concrete development’ of AI systems within projects and fail to consider that concrete development (Buchholz et al. 2022). They then place ethics by design as a sub aspect of Responsible Research and Innovation, detailing its implementation in ethics boards, and frameworks such as the HLEG guidelines. The authors argue that this is not enough however, and they suggest two methodological approaches: design for values and human-centric design. Design for values depends upon embedding values in the design process and therefor also in AI systems. Human centric design on the other hand uses social science methods such as interviews and group discussions to capture the needs of the humans for the design process, and then transform these needs into values (Ibid.). From this, Assistant uses a high-level architecture document to outline responsibility issues beforehand, and then revisits issues in the low-level architecture context. The high-level document is iteratively developed at three points in the project, calling for concrete steps such as a responsibility map for distributing moral and legal responsibilities, and workshop formats for achieving interoperability of components in ways which include all stakeholders. The low-level engagement concentrates on risk assessment and especially definitions which contextualize the latter so as to facilitate a risk management process for partners.

We note that participation is facultative and the attempt is not made to define values (Ibid.). The process is iterative, which agrees with our own approach, but it is not continuous, which would seem more advantageous, although a continuous approach also adds more work. The focus on stakeholders developing the technology and benefiting financially from it, can remove a focus on the end users – workers – who do not have a proportional stake in those terms, but who have a much higher stake in their own terms, i.e., the value of their work methods and experiences. Since the risk management process is facultative, nothing definite is formalized which would allow a quantitative measurement with regard to non-participation in the process by the technology and industrial partners. This is a deficiency which our own ethics team, as well as teams in other projects – e.g., TEAMING.AI – have tried to address.

4.1.2 COALA

Coala project has a fairly comprehensive approach to ethics in ICT 38. Besides explanations of bias in machine learning, they discuss methods for avoiding or mitigating bias, privacy, trust, and autonomy. They also integrate certain parts of the ethical discussion with their historic arguments from philosophy, namely that of autonomy, in Kantian and Rousseauian terms. Their approach asked the partners themselves to identify ethical problems, through the use of a focused ethical questionnaire. Bias did not come out as a current issue, and privacy was viewed as a legal concern predominantly. Coala project used an ethics questionnaire, as well as the Altai form developed as a counterpart to the HLEG guidelines. They generated quantitative results with regard to the most relevant ethical issues impacting their project. They also created an ethics manager and ethics board to discuss ethical problems democratically, with advice from an external advisor. The admission that the full board was not convened at the time of publishing the deliverable and that no issues had yet been brought forward, but that the board would convene regularly as deployment phase begins (Coala Public Deliverable 7.3, pg. 26) illustrates some of the difficulties in instituting ethics by design through a more democratic and framework centered approach, such as that adopted by Coala, i.e., the ‘machinery’ of larger overseeing groups and framework heavy methods can be somewhat slow to come into action.

We agree on many of the general problems raised by Coala, and their respective recommendations although our identification of problems and our recommendations are perhaps more specifically contextual. We agree particularly on the position that “the worker must control the device not the other way around” (Coala, Deliverable 7.3, pg 19). We also think the use of the Altai assessment to get actual assessment results is appropriate, and, as an unambiguous demonstration of the tool in the projects of the cluster, it is complementary to our own approach.

AI PROFICIENT ethics approach differs however, in attempting not to focus the potential ethical issues under certain pre-selected categories, but instead taking things as we find them. We think this unregimented attitude helps us uncover some ethical problems that would otherwise remain hidden.

The AI-PROFICIENT ethics team did not carry out a sustained integration of philosophical ethical theory looking to historic traditions, judging it more important to move to active operationalization, the aspect which is most often left undone in AI ethics. We did adopt a pragmatic – to some extent Deweyan – outlook in some aspects, however, particularly in our emphasis on ground level experience and active and evolving engagement of the ethical problems arising from project technical solutions that were continually changing. Thus, we are pleased to be able to complement Coala’s philosophical engagement in that sense, from other philosophical traditions.

4.1.3 EU-Japan.AI

We have not found specific ethics focused deliverables for the EU-Japan.AI project, although (Adams 2022) discusses the larger AI related social issues around human robot interaction, workplace surveillance, and job loss through AI driven automation.

These issues are all good to note. Complementary to them, our own approach has concentrated first on operationalizing rather than discussion of broader social issues around the development of AI. Nonetheless our approach has tended to uncover the social issues in ground level contexts also, e.g., the tension between AI systems surreptitiously adding more work in a move which is relatively unethical and yet quite legal, depending on the context.

4.1.4 knowlEdge

The knowlEdge project applies the HLEG guidelines to the conceptual architecture of the project platform, across a number of layers of data integration, data analysis, AI and data analytic, knowledge management, and finally smart decision support functionalities (Wajid et al., 2022). It applies HLEG at the very high level of its seven requirements for trustworthy AI, without breaking the requirements down into their respective sub-questions. The way in which the project satisfies each requirement is then outlined, occasionally with one project component taken to fulfil several HLEG requirements.

The knowlEdge approach Illustrates that, when a checklist paradigm is adopted, the questions can remain very general, but they can be minimally satisfied with a very general response accordingly. Thus, for example, that explainable AI is implemented *as such*, giving human users the chance to intervene in every decision cycle of the system, according to the HLEG definition of Human-on-the-Loop, is taken to satisfy the HLEG requirement for Human Agency and Oversight, while the Digital Twin component *as such* – the ability to model outcomes before real world deployment – is taken as one of two components which satisfy the Technical Robustness and safety requirement. The relative poverty of the HLEG as a self-sufficient approach – and similar top-down framework approaches – is thus made visible. The human has the opportunity to intervene indeed, but the details of the opportunity, the training of how to intervene, the obligations and responsibilities with regard to intervening, the question of whether the intervention formats and user interfaces suit the user, etc., all remain undeveloped.

4.1.5 STAR

Star project approached ethical and legal analysis through a series of questions tailored to each pilot within the project with a stated focus on privacy issues and human-centric design (Soldatos and Kyriazis 2021). The approach was implemented in online co-creation workshops.

While the aim and spirit of human centric design is laudable, the ground level operationalization of the term is not expanded enough however, beyond the introduction of the somewhat new term operator 4.0 and the concept of the human digital twin (HDT), which, as (Montini et al. 2021) note, is addressed in very few works. The expansion of the digital twin concept is a good step. It retains several deficiencies, however. One is that it is, inevitably, a generalization of the human, and perhaps a very thin generalization. A second, and more pressing problem, is that the positive impacts which (Montini et al. 2021) describe under ‘Worker Well-being Monitoring,’ are arguably, given the factors outlined – e.g., decrease of absenteeism, productivity increase – positive only for the manufacturing management, while being negative for the worker. Thus human-centric design, if it does not take care to focus on the actual human worker – as opposed to the generalization which is the HDT – is only very questionably linked to and made a prominent part of a full ethical approach. Before it could be made fully ethical one would have to clarify among other things, the motives behind the ‘positive impacts’ and the point of view of the actual human workers.

4.1.6 MAS4AI

Bias and its elimination through better management practices for data, is the focus of the MAS4AI approach to ethics. A number of general ethical issues are highlighted, including privacy, power imbalances that AI might cause, transparency of AI systems, environmental impacts, job loss, autonomy and control issues, and accountability. Four guidelines are also mentioned; the Asilomar principles,

Ethically Aligned Design, HLEG guidelines, and the OECD recommendations, and reviewed in high level terms. Very general recommendations are then made, which culminate in the suggestion that pilot implementation partner leaders use the HLEG Altai tool to self-assess their systems. One recommendation, although general, which is unique to the MAS4AI project – although COALA has an ethics manager – is the involvement of an ethics mentor 'with appropriate expertise in ethics of new and emerging technologies,' for those pilots which have significant ethics risks.

The reference to potential job loss is good, as this is a rarely raised issue. The ethics mentor recommendation is also very welcome. The expansion of the concept of ethics mentor would have been useful in terms of the methodology to follow, the level of generality – shop floor or high level – the mentor would engage, and the possibilities for making the best use of such a mentor, and also as to the background of the mentor: should the latter have a formal training in ethics and philosophy?

4.1.7 TEAMING.AI

The TEAMING.AI method combines legal and ethical requirements but is oriented more toward the former. It proceeds downward from the HLEG guidelines to formulate 6 Ethics Requirements, which are very general in application: consultation of everyone affected by the AI technology, continuous availability of a human contact person, creation of a safety risk management plan, availability and explanation of accessible written information regarding the use and impact of the technology, ongoing assessment regarding discriminatory impacts of the technology, and finally partner self-assessment through the Altai tool. From there it moves to provide a set of 12 legal requirements grounded in the GDPR, and a further set of 7 legal requirements grounded around the notion of high-risk AI systems. Evaluation of compliance to requirements is proposed in terms of naming requirements, checking claims of compliance (yes/no) and checking evidence. Compliance results can then be rendered in a knowledge graph. The goal is standardized auditability and automated compliance verification.

We appreciate the focus on actually checking for compliance, although we wonder if, as a combination of Altai and a highly structured format, it might miss some important ethical problems which are difficult to uncover through a generalized and automated approach. The TEAMING.AI approach is much more legal and regulatory focused than the AI-PROFICIENT approach. The AI-PROFICIENT team attempts to draw ethical practices from the spirit of regulations. In this sense the TEAMING suggestion, that their project be able to deal with high-risk applications, even though they may have none, is an attempt to look ahead to cover all eventualities. Our ethics team differs in that we try to 'recover' the ethical sense of the regulations, or potential regulations, by recommending related best practices, but only for issues that are arising out of the specific context. We do not try to cover all eventualities, neither ethically nor legally. The TEAMING ethics team differs from us also, in advocating a mechanical use of requirements or recommendations, where the issue has not become evident in context, e.g., the TEAMING requirement #3 to institute a safety risk management plan, despite the fact that, as they note, 'no significant risks to health or wellbeing should exist.' (TEAMING.AI Deliverable 1.3 Policies, pg. 11)

4.1.8 XMANAI

Complementary to the focus of their project, XMANAI highlights the fact that lack of explainability in AI is unethical as such (Lampathaki et al. 2021). The ethics approach then seems to rest upon the fact of developing explainability services as a general good. They also stress conformity to various European regulations and directives, particularly with regard to data, and that the information regarding data use will be provided to all partners. Conformity thus becomes the responsibility of each partner but partners are to be aided by open discussions. Participation in demonstrators is to be voluntary and formally consented to.

Beyond the generalization regarding explainable AI, a more detailed and specific exploration of the potential ethical aspects for explainable AI, particularly as to how it might be tailored to the human user for whom it is developed, would be helpful. Accordingly, without the latter, it remains very much a technical solution, in comparison with the human-centric focus adopted by, e.g., Star project. XMANAI

has more focus on data than other ICT 38 Cluster projects with a more detailed plan for ethical data management, which the informed consent requirement completes practically.

4.1.9 AI-PROFICIENT in the ICT 38 Cluster

Based on the above outlines of and comparisons with other projects, what AI-PROFICIENT ethics team has brought to the cluster in terms of unique methods or methods which parallel others in the cluster, includes the following.

We have demonstrated how having a dedicated ethicist with a philosophical background in ethics can help contribute to non-technical viewpoints on ethical issues of AI in heavy industry. Along with this we have integrated and tested some aspects of the Pragmatist ethical tradition particularly, which complements the Coala project integration of Kantian notions of autonomy.

We have fleshed out the ground up aspects relative to the actual physical human user, which complements the other human centered approaches in particular, such as the Human digital twin approach outlined in Star project.

An attempt has been made in this deliverable, as well as the earlier AI-PROFICIENT Deliverable 1.2 and a supporting published article, to clarify the differences between ethics and law, and to show the overlap and the difficulties in treating the two together. In this we complement the more legal and compliance-oriented focus of Teaming.AI

The ethics team has demonstrated the use of very specific recommendations at the ground level, along with a more extended treatment of ground level issues, e.g., in uncovering them through specific recommendations. Along with that we have made a link to a more fine-grained application of the HLEG by moving from the bottom up, and moving beyond theory to actively *doing* ethics, in other words a more flexible method than one which relies on getting consensus among various partners (or ethics boards) before acting, even though the latter – as in COALA's approach – is also important.

Above all, the most important contribution to the cluster has been the stressing of the quantitative aspect for operationalization, aimed at specifically measuring ethical implementation successes at higher resolution. This has been done with a methodology aimed at uncovering what does and does not get done ethically in various categories, so as to try to understand why it does or does not get done.

4.2 Insights from our ethics approach

One insight gained from the ethics methods applied in this project has been that an ethical culture of tech development could be built with sustained effort. Over the three years of the project, the ethics team has seen the partners who are interested in incorporating ethics by design in the project grow more comfortable in discussing and considering ethically related adjustments to what would otherwise have been purely technical solutions.

Partners were visibly influenced by other partners' successes in carrying out ethical recommendations. Research such as that of (Ellemers et al. 2019) has shown, as one might expect, that moral social 'surroundings' have a psychological effect on individuals which plays a large part in their morally associated actions. Perhaps because of the latter effect, the project partners did not simply upgrade everything to yes, which might have been expected if it was just a matter of getting a higher score, in a context in which they were cut off from other partners. In some cases, they downgraded our assessment, or they upgraded from no to a partial result. Working together on ethical recommendations with other partners thus leads to a more objective result, i.e., objective in the sense outlined by such philosophers as Josiah Royce (add) and John Dewey: the moral social context is gradually created and evolving but nonetheless objective, because all of us will to make it objective as a moral community which is a counterpart to the communities of nature and science.

Failure to implement recommendations is a major issue, but not one specific to the project, nor one which taints the ethical results of the project. Some partners simply do not implement recommendations or implement selectively. There are no quick control-based solutions to this issue. Also, applying ethics takes time, and ideally would not be limited by deadlines, but be ongoing. In many cases partners insisted that the recommendations would eventually be completed by the end of the project – they may be but though the ethics team could not wait for this, still our results are promising even as they stand.

The issue of having things hidden from view is also a major issue for applied ethics. When there is a community of partners and partners can see ethical lapses within the community, it encourages the sources of those lapses to do better. The building of the tech community as a more visible community in terms of its efforts and results should thus be a priority for research. The big names of Big Data actively work against this visibility in many cases, and, for example (Fort, 2023) has noted with regard to ChatGPT, we don't know anything about it, or even what it is for. Dewey and Tufts (1932) have argued that one benefit of large corporations in the moral realm, is that the size of modern corporations creates a visibility for ethical issues related to the actions of those corporations which can go missing in individual moral behaviour, i.e., *for large corporations, unethical actions and motives are potentially always on display*. Such corporations often actively work against this of course, as do smaller corporations, which means the challenge of making things more visible, quantitatively – as we have attempted here – and in opposition to euphemisms and 'management speak,' is one which we must embrace.

Thus, the ethical way forward seems to be one of slowly bringing more people onboard and developing a culture of applied ethics so that custom takes hold, practically and flexibly. It is a matter of solving conflicts of value at various levels of generality, as well as expanding and re-interpreting more specific problems at higher levels of generality, while bringing the more general ethical issues down to specific levels in order to get at them. The financial motives of manufacturing, and corporations will have to be addressed in this process as well.

4.3 Limitations

The partners probably had not experienced a hands on and ground up approach before. This was thus a limitation in getting the ethics by design process moving. The ethics team considers the adopted approach to be a relatively new approach to applied ethics of AI in industry. The partners probably did not know what to expect when the project began, and the ethics team evolved and created the approach as the project went along. Some of the earliest recommendations – those made in the first four months – may have remained un-implemented wholly or in part because the partners were still unsure of the ethics approach. We observed that some of the more willing partners became more comfortable with the general approach as the project progressed.

The main limitation, in parallel with the above observation, is that, across categories, implementation of recommendations depends a lot on the attitude which the partner, as an organization, takes toward ethics. This makes it difficult to quantify ethical results such that they could be taken to be completely objective and unbiased. Yet it also points out several facts which are a way forward for ethics at the organizational level: if an organization can have an ethical attitude, then we can cultivate that ethical attitude.

This leads to a related limitation: the roles played by particular individuals in the ethical implementation results were clear. The specific influences of those roles are missing from the results, however. A range of interest was observed among individual participants in each partner organization. Some individuals were quite interested in operationalizing ethics, while others were mostly un-interested. With regard to the written deliverables for which an ethical issues section containing recommendations was explicitly included, the task leader in charge of the deliverable often pushed for the section to be completed as part of the deliverable completion. Often, this was not just a matter of mere routine however, because we note that some deliverable ethical issues sections were left relatively undone, while in other instances the task leader suggested to include an ethical issues section in a deliverable for which none was planned.

The bias limiting the objectivity of the study in quantifying ethics also shows that an approach of operationalizing ethics at ground level is the right way to go precisely because it is developing the applied ethics techniques to deal with such biases and locating where those techniques need to be applied. In other words, knowing that some organizations do not have a culture conducive to applying even direct ethical recommendations, leads to asking why they don't, and to how an ethical culture can be instilled in the organization. It also leads to asking how certain individuals working in the organization could serve as entry points for instilling an ethical culture.

4.4 Future Research to follow up on

In a pragmatist ethics in the Deweyan tradition, developing the relation between the individual and society – e.g., by improving the social context so that the individual can flourish – is a main focus of a practical application (Dewey and Tufts, 1932). In this project that relation is located at the level of the work organization. The 'individual' is the industrial worker or the individual software engineer working within the tech company, whereas the 'society' is, respectively the group of industrial and tech company project partners. This is good ground upon which to operationalize ethics. It can also frame the questions going forward.

Research thus needs to be done on how the internal culture of the partner organization (industrial or tech company) and the relation between the organization and the individual working within it, influences their response to a ground up ethical approach such as this one. Further research should focus on the paradigm of relative and incremental ethical results in industrial and other contexts, i.e., progress should be made on an ethical approach which aims at bettering a context or proposed solution, rather than an either-or approach simply transposed by fiat – usually to no effect – from high level norms, or from laws and regulations.

For AI ethics in the heavy industry context and work context more generally, there is also the opportunity to study how an applied ethics method such as the one presented here can locate and bring out into the open instances of the uneasy contradiction mentioned earlier, between additional work, caused by the adoption of AI and related technologies, which is legally allowed but unethical, as opposed to extra work which is legally allowed but also ethically consistent. Studies bringing this contradiction out in the open of a public discussion would go an important step further.

Further, the fact that there was not a large difference between the partner assessment of implementation results and the assessment of the ethics team also present an opportunity, particularly considering that many upgrades were in a few categories and by a specific partner. A good line for further studies here would be: how much parallel ethical assessments and the possibility of peer review of ethical assessments at operational level, influence how developers or industrial partners rate their own performance. Industrial contexts of AI ethics could serve practically to integrate such studies with psychological and related research to uncover positive practices to be applied.

Finally, it would be interesting to know whether a longer or shorter time frame works better for operationalizing AI ethics at ground level in the industrial context. Would less than a year give enough time to activate the ethical culture we have spoken of, or would a period longer than three years be better?

Conclusion

Deliverable 6.4 is the completion of work begun in AI-PROFICIENT Deliverable 1.2. It shifts from a description of the methodology used by the AI-PROFICIENT ethics team in its ethics by design approach to AI in heavy industry, to a description of the results and insights gathered from the use of that methodology.

The HLEG guidelines for trustworthy AI have been considered in detail, as to their aims, usefulness, and shortcomings with regard to AI-PROFICIENT. The modifications, relative to the particular context

of AI-PROFICIENT, of the framework and checklist approach implied in the HLEG guidelines have been stated and the reasoning behind those modifications given.

An outline review of the methodology and activities, over three years, which gave rise to the recommendation implementation results was made. The implementation results were then given in various formats and from two angles: that of the ethics team assessment and that of the project partner assessment. From this comparison we see that the AI-PROFICIENT ethics approach is workable and it can provide some much-needed substance to an ethics component included in an overall project evaluation carried out in quantitative terms.

Going forward, the opportunity which the heavy industry context provides for interactions of AI system developers with individual humans in physical contexts, should be stressed. Rather than abstract virtual 'datafied' representations of humans, often in large numbers beyond a 'humanly' understandable scale, here one can interact with real people, at least to some extent. Interacting with real people is arguably the key to operationalizing ethics. If the related experiences of such interactions can be kept in sight be those developing technologies, the clues they give toward operationalization should be transferable to other contexts where people have tended to become mere numbers.

Thus, it would also be interesting to study which aspects of the ground up methodology presented could be useful to AI ethics in other contexts. The above insights regarding the participation of the project partners developing the AI systems, should apply with some modification to non-industrial contexts. AI system developers (the individuals) in many contexts are part of teams hired by corporations, and, as in AI-PROFICIENT, their corporations interact with other corporations. What has worked in getting these individuals in AI-PROFICIENT to operationalize ethics, is likely, with some patience and experimentation, to work in other contexts; at least we should try.

In the end, the aim, for ethics applied to AI systems in the EU context, should be uncovering more humane, more comprehensive, and more refined, ways to operationalize ethics. In many cases, judging by the large and growing body of frameworks and literature on AI and technology ethics, we already know what should be done ethically, we have the ethical ideals. Now we have to get down to brass tacks and do it.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957391.

References Cited

Adams, Andrew (2022). The Social, Legal And Ethical Implications Of AI For Manufacturing. https://www.eu-japan.ai/the-social-legal-and-ethical-implications-of-ai-for-manufacturing/

AI-PROFICIENT Project Proposal (2020)

AI-PROFICIENT Project Proposal – Annex 1 (2020)

Anderson, M. M., & Fort, K. (2022). From the ground up: developing a practical ethical methodology for integrating AI into industry. AI & SOCIETY, 38(2), 631-645.

Avnoon, N., Kotliar, D. M., & Rivnai-Bahir, S. (2023). Contextualizing the ethics of algorithms: A socio-professional approach. new media & society, 14614448221145728.

Artstein, R. & Poesio, M. Inter-Coder Agreement for Computational Linguistics Computational Linguistics, MIT Press, 2008 34, 555-596

Asher, N., Castets-Renard, C., Loubes, J.M., Risser, A. (2022). Coala Project. Public Deliverable 7.3 Report on the application of ethical principles for AI in manufacturing – interim.

Berrah, L., Cliville, V., Trentesaux, D., & Chapel, C. (2021). Industrial performance: an evolution incorporating ethics in the context of industry 4.0. Sustainability, 13(16), 9209.

Buchholz, J., Lang, B., & Vyhmeister, E. (2022). The development process of Responsible AI: The case of ASSISTANT. IFAC-PapersOnLine, 55(10), 7-12.

Clark, J. L., & Lischer-Katz, Z. (2023). (In)accessibility and the technocratic library: Addressing institutional failures in library adoption of emerging technologies. First Monday, 28(1). https://doi.org/10.5210/fm.v28i1.12928 (Original work published January 16, 2023)

Crawford, K. (2021). The atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.

Desrosières, A. (2008). Pour une sociologie historique de la quantification : L'Argument statistique. Presses de l'école des Mines de Paris.

Dewey John, Tufts, James Hayden. Ethics. H. Holt and Company, New York, (1932).

Ellemers, N., Van Der Toorn, J., Paunov, Y., & Van Leeuwen, T. (2019). The psychology of morality: A review and analysis of empirical studies published from 1940 through 2017. Personality and Social Psychology Review, 23(4), 332-366.

European Commission. Horizon 2020 Work Programme 2018-2020

EU Commission, Funding & Tender Opportunities, AI for Manufacturing, ICT-38-2020 https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/ict-38-2020

European Commission High Level Expert Group, Ethics Guidelines for Trustworthy AI, 2019. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Fernandez et al. (2023) "Human-Feedback for AI in Industry," in the Sixteenth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services, CENTRIC 2023, Valencia, Spain (November 13-17) [forthcoming]

Fort, Karën. (2023) Les enjeux éthiques de l'IA vus par le prisme du traitement automatique des langues. Conference Presentation. JSALT 2023. https://umotion.univ-lemans.fr/informatique/jsalt-2023/video/9500-karen-fort-les-enjeux-ethiques-de-lia-vus-par-le-prisme-du-traitement-automatique-des-langues/ (accessed July 15, 2023)

Graux, Hans. (2021) TEAMING.AI Deliverable 1.3 Policies.

Green, B. (2021). The Contestation of Tech Ethics: A Sociotechnical Approach to Ethics and Technology in Action. ArXiv, abs/2106.01784.

Lampathaki, F., Agostinho, C., Glikman, Y., & Sesana, M. (2021, June). Moving from 'black box'to 'glass box'Artificial Intelligence in Manufacturing with XMANAI. In 2021 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC) (pp. 1-6). IEEE.

Montini, E., Bonomi, N., Daniele, F., Bettoni, A., Pedrazzoli, P., Carpanzano, E., & Rocco, P. (2021). The human-digital twin in the manufacturing industry: Current perspectives and a glimpse of future. Trusted Artificial Intelligence in Manufacturing: A Review of the Emerging Wave of Ethical and Human Centric AI Technologies for Smart Production, 132-147.

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., & Floridi, L. (2021). Ethics as a service: a pragmatic operationalisation of AI ethics. Minds and Machines, 31(2), 239-256.

Nafus, D. (2018). Exploration or algorithm? The undone science before the algorithms. Cultural Anthropology, 33(3), 368-374.

Prem, E. (2023). From ethical AI frameworks to tools: a review of approaches. AI and Ethics, 1-18. https://doi.org/10.1007/s43681-023-00258-9

Sætra, H.S., Danaher, J. To Each Technology Its Own Ethics: The Problem of Ethical Proliferation. Philos. Technol. 35, 93 (2022). https://doi.org/10.1007/s13347-022-00591-7

Soldatos, J., & Kyriazis, D. (2021). Trusted Artificial Intelligence in Manufacturing; Trusted Artificial Intelligence in Manufacturing: A Review of the Emerging Wave of Ethical and Human Centric AI Technologies for Smart Production; A Review of the Emerging Wave of Ethical and Human Centric AI Technologies for Smart Production.

Strubell, E., Ganesh, A., & McCallum, A. (2020, April). Energy and policy considerations for modern deep learning research. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 09, pp. 13693-13696).

Wajid, U., Nizamis, A., & Anaya, V. (2022). Towards Industry 5.0–A Trustworthy AI Framework for Digital Manufacturing with Humans in Control. Proceedings http://ceur-ws. org ISSN, 1613, 0073.

Widder, D.G., & Nafus, D. (2022). Dislocated Accountabilities in the AI Supply Chain: Modularity and Developers' Notions of Responsibility. ArXiv, abs/2209.09780.

ZVEI - Zentralverband Elektrotechnik-und Elektronikindustrie e.V. (2012). More energy efficiency through process automation.
https://www.zvei.org/fileadmin/user_upload/Presse_und_Medien/Publikationen/2013/januar/More_nergy_efficiency_through_process_automation/ZVEI_Energienutzung-englisch.pdf