Evaluating the acceptability of ethical recommendations in industry 4.0: an ethics by design approach

Marc M Anderson, Karën Fort

Evaluating the Acceptability of Ethical Recommendations in Industry 4.0: An Ethics by Design Approach

Marc M Anderson[1] Karën Fort[2]

Abstract

In this paper we present the methodology we used in a European Horizon 2020 project in order to evaluate the implementation of the ethical component of the project. The project is a three-year collaboration between a university partner and industrial and tech partners, which aims to research the integration of AI services in heavy industry work settings. An AI ethics approach developed for the project has involved embedded ethical analysis of work contexts and design solutions and the generation of specific and evolving ethical recommendations for partners. We have performed an ongoing evaluation and monitoring of the implementation of recommendations. We describe the quantitative results of these implementations: overall, broken down by category, and broken down by category and responsible project partner (anonymized). In parallel, we discuss the results in light of our approach and offer insights for future research into the ground level application of ethical recommendations for AI in heavy industry.

Keywords: artificial intelligence; ethics; heavy industry; Industry 4.0; operationalizing

1 Introduction

AI Ethics in general suffers from a lack of operationalization. (Morley et al., 2021) have highlighted a number of reasons which might lie behind what they call "a significant gap … between theory and practice within the AI ethics field." Meanwhile, various approaches have been taken to address this gap, for example those summarized by (Prem, 2023), but nearly all remain either at a very general level, or if they engage an operational level they remain very much within the range of technical adjustments to the algorithm or data, i.e. they disconnect an engagement of ethical concerns from the humans involved, for example the users or the software engineers working for the technology companies.[3]

There are three motivations behind this research. First, operationalizing AI ethics depends upon finding methods to implement ethics at a ground level, but a large part of implementation involves the willingness of software engineers to actually implement the methods. There are few studies which explore, as we do here, the question of what software engineers, developing AI systems in technology companies, are ready to put in place with regard to AI ethics. (Widder and Nafus, 2022) provide strong general insights into this question, but do not go as far as a quantifiable level for operationalization. Secondly, the issue of the potential of implementing AI ethics at operational levels as it overlaps with general ethical implementation concerning the humans and human contexts affected by the AI, has not been explored, either in terms of categorization or quantifiability, a deficiency which

- [1] marc.anderson@inria.fr; dr.marcanderson@gmail.com
  LORIA, UMR 7503, Université de Lorraine, Inria and CNRS, Campus Scientifique, 615 Rue du Jardin-Botanique, 54506 Vandœuvre-lès-Nancy, France

- [2] karen.fort@loria.fr
  Sorbonne Université/LORIA, UMR 7503, Université de Lorraine, Inria and CNRS, Campus Scientifique, 615 Rue du Jardin-Botanique, 54506 Vandœuvre-lès-Nancy, France

[3] Hereafter *tech companies*, or *tech company partners* in case we refer specifically to the technology company partners collaborating within our project. *Software engineers*, in contrast, refers to the individuals working within these tech companies.

we try to redress. Thirdly, and closely related, the question of whether there is an actionable typology of AI Ethics categories for operational level implementation which is more fine-grained and includes the overlap described above, deserves study.

Beginning from an approach of giving operational level ethical recommendations, our interest has been focused on knowing not merely how many ethical recommendations have been achieved but what types of recommendations are more or less likely to be implemented. Thus, a categorization of recommendations was required, but a bottom up categorization, i.e. one which grew out of the specific issues we encountered – in their human context – and which could accommodate those issues, rather than one which discarded issues which didn't fit with a pre-developed categorization. This would help understand how the industrial partners and tech company partners see the ethical angle, help make suggestions or advance potential future lines of research to improve implementation of recommendations in poorly implemented categories, and apply insights from some categories toward better implementation of recommendations in others.

In this article we thus present a quantitative review of the general and categorized results of direct ethical recommendations given to partners of an EU Horizons 2020 project. In Part 2, we begin with an overview of project context and timeline, the methodology adopted, and our approach for assessing the implementation result status of recommendations. In Part 3, we describe our category definitions along with examples, and our categorization process for recommendations. Part 4 presents, in various figures, implementation results relative to recommendations, in terms of overall results for the project as a whole, then by category, and finally broken down both by category and implementing partner, followed up, in each case, by observations and discussion. Both the ethics team assessment and the project partners assessment are given and discussed. Part 5 discusses limitations with regard to industrial technical contexts, while Part 6 offers some suggestions for future research regarding less implemented recommendation categories. Finally, we conclude with some reflections on interesting aspects of our findings.

To the best of our knowledge our approach to quantifying the implementation results of specific ethical recommendations is entirely new. In other words, there is no 'state of the art' with regard to the approach we are describing. Even on the subject of embedded and practical approaches to applying ethics of AI in industrial settings the existing research literature is sparse verging on non-existent. Much of the literature, such as (Tahaei et al., 2023) consists of extremely broad overviews, which assume a generic notion of industry practically unsuited to our aims, and which engage or summarize ethics under the usual few abstract headings such as explainability and fairness. Research which does address heavy industry e.g. (Singh et al., 2023) also tends to remain at an abstract level, or as with (Ciobanu and Meșniță, 2022) it remains at both an abstract and a technical level.

The publication of (Anderson and Fort, 2022), who develop a bottom up approach for ethics of AI in industry and that of (Berrah et al., 2021) who develop a system for integrating ethics as a dimension of performance evaluation, while admitting that quantifiable performance indicators for ethical integration are difficult to attain, are the nearest to our approach.

2 Project Overview & Methodology

The project began in 2020, with the aim of researching various aspects of AI integration in heavy industry settings in three manufacturing sites located in France, Belgium, and Germany.

Ethics by design had to be scaled up gradually since at the beginning there was only the initial project proposal to go by. The first opportunity for consideration of the project's ethical issues was thus offered in the preliminary stage, approximately a three-month period, in which the contexts of the possible Use Cases (hereafter UC) were presented by the manufacturing partners, question and answer sessions were held, and certain UCs were selected to go forward in the project.

At this stage of the project the aim was to develop a baseline of the current situation of the humans involved in each UC – primarily the shop floor level operators or maintenance technicians – within their work environment, in relation to the other workers, machines, and automated systems already in place, tasks being done, time to carry out those tasks, and physical spaces.

A set of baseline questions were asked to flesh out these contexts. Then an initial summary of ethical interest in each UC was made, in keeping with a call to the other partners to summarize their interest in each UC in terms of

what problems they had interest in and competence in engaging. On its part, the ethics team summarized potential ethical difficulties in the UCs and rated each of them in terms of these difficulties. The first ethics team recommendations – or pre-recommendations – were thus appraisals of each UC at three levels of difficulty: high, medium, or low in terms of ethical issues, with an outline of why each was rated as it was. The rating was arrived at by an appraisal of what was proposed for the AI, in comparison with the fleshed out initial context of the UC.

The ethics team did not specifically recommend not choosing certain UCs, but made clear which were the most problematic in terms of ethical issues. Subsequently, on that basis, the partners – the ethics team was not part of the decision committee – decided to eliminate the ethically most problematic UC from the project. The ethics team's reasons for rating this UC as high in terms of ethical issues, was, among others: the operator had very little time to act or react so that AI suggestions might add stress to an already small time frame, and the human agency aspect was ambiguous. The UC was abandoned even though its solution, which had been envisioned as a way to increase efficiency and eliminate downtime in the initial stage of the production process in question, was the stated top priority of one of the industrial partners. Other UCs were also eliminated from consideration for other reasons, e.g. the tech company partners did not find the problems within them to fit with their technical competencies or interests.

The ethics team considers the elimination of the ethically problematic UC, after discussion, to be a major success, and a full implementation, but could not include it in the quantified results because it was different in kind and unique in terms of quantification. But it is important to mention it here because it belongs to that category of actions centered around the often-forgotten or dismissed question, highlighted by Hagendorff (2022), of *whether in certain contexts AI systems should be used at all*. This question is rarely discussed in AI ethics, but here an ethical appraisal which resulted in not using the technology was followed by the partners in a significant way.

2.1 Recommendations

After the initial UC selection, the embedded ethicist participated in the ongoing design and technical meetings of the project, which is estimated at about 50 to 60 meetings per year for each of the first two years. At the same time, the ethics team participated in the development of the work package tasks of the project as well as in reviewing and contributing to the related deliverables. Thus, the initial sets of recommendations were specified according to a particular UC – approximately the first 7 months of the project – and were concerned with the initial solution plans for each UC relative to its work and worker context. Later recommendations tended to be specified according to a particular UC context within a task. Since the latter were dealing with written deliverables as well as design plans discussed in meetings, the recommendations regularly concerned the planned solution described in the deliverable and sometimes they also concerned the wording or content of the deliverable itself. The recommendations thus tended to become more precise as the AI service solution became more developed. In this way recommendations evolved, were integrated with one another, and gave way to new recommendations as solutions advanced and changed. This approach is only possible if there is really an ethics by design in practice, rather than merely in theory.

A total of 130 recommendations have been issued to date. The recommendations were given formally, i.e. by means of a written document which described the UC context, reviewed the ethical issues in question relative to the context, and gave precise numbered recommendations to address the ethical issues. The following are some anonymized examples of actual recommendations, with other examples provided below in Table 2.1 according to category.

- **x)** *Recommend that you formally clarify to the operator if the operator's role changes with regard to checking tread alignment, e.g. no longer has to check or checks less frequently. If the operators will no longer check, clarify this to the TBM operators*.

- **x.x-x)** (Partner X) Regarding labeling of good and bad images for alarm: *Recommend that you clarify who will label the images and estimate how many images need to be labelled and how long it will take*.

- **x.x-x)** (All) *Recommend that for Task X.X Use Cases where explicit and implicit feedback will be combined (xxxx_2; xxxx_5), you develop a best practice of tagging the data resulting from that feedback to indicate that active operator choices (explicit feedback) make up part of it*.

The ethical recommendations documents were dated and disseminated to the designated ethical contacts for each project partner as well as other involved partner members, and they were deposited on the project Microsoft Teams ethics channel files, in conformity with the practice adopted for other work within the project. The recommendations were also discussed in technical meetings, both as the ethics team became aware of them and as the technical solutions progressed. Sometimes a partner asked for special meetings to discuss potential ethical issues. Additional recommendations or additional aspect of recommendations might then be added.

2.2 Assessment

The Ethics team kept track of the state of implementation of each recommendation as tasks, deliverables, and UCs progressed. They also outlined what a full implementation would consist of. Based on this ongoing assessment, and in the interest of keeping the complexity of the approach within reason, it was decided to put the results of implementation under three categories. Fully implemented indicates that a recommendation was fully implemented according to what we expected, or has been committed to with evidence that it will be by project end. Partially implemented indicates that some effort was made toward implementation, or in the case of recommendations with several aspects, one or more but not all aspects were implemented. Not implemented indicates that no evident or proven effort was made for that recommendation.

The default assumption is thus negative tending. A recommendation is considered as not implemented unless at least some evident effort was or is being made toward implementation, and is considered partially implemented unless *all* aspects are evidently implemented or in process of implementation.

In addition, 9 recommendations were classed as NA (not applicable), by the ethics team, because changes in a proposed solution – often independent of ethical recommendations – rendered them irrelevant.[4] These 9 have not been included in the total, thus reducing it to 121.

Clearly the approach is imperfect in terms of being exact, because unless the ethics team waits until the project is complete there is still some small chance that all recommendations could be implemented. On the other hand, in most cases the time for implementation is practically passed in the project. Thus, a certain pragmatic imperfection has to be accepted. But this imperfection aligns with the project ethics team's embedded ethics by design approach, whose aims – in contrast with an all or none approach – are that 1) better some ethics, as much as we can get, be actually implemented than none at all, and 2) a demonstration of ethics by design starting from a ground up approach and relying on very specific recommendations to change the design path of a technology development project be shown to produce some results which can be improved in later efforts, and 3) insights from the approach be gathered and suggestions be made toward bettering future efforts in ground up applied ethics of AI and technology.

3 Categorization of Ethical Recommendations[5]

3.1 Definitions of Categories

Recommendation Category	Definition
Protocol	Adopt a specific set of instructions regarding errors, new tasks, etc. Example: ETH X UC X ID X - *Recommend that you* d*evelop a protocol to address the allowed for 5 to 10% of cases when the AI misrecognizes text, i.e. state the steps the operator will take in case of AI error.*
Human centering	Tailor aspects of development to individual users and develop services collaboratively with users (e.g. work with the operator to design something) Example: ETH ID X.X-X – (Partner X; Partner Y) *Recommend that you carry out a preliminary short survey, e.g. 10 questions, of user*

[4] NA recommendations have not been included in the calculation of percentages for implementation.
[5] Note that recommendation categories were developed after the specific recommendations had been given and were *not* disseminated to the project partners.

	background knowledge (process engineers and operators) regarding AI, to be used in adjusting for potential user assumptions during XAI development.
Design	Make changes or additions to technical or procedural elements of the solution Example: ETH ID X.X-X – *Recommend that you separate the x prediction from the x prediction service if possible, so that the operator can devote clear attention to x prediction faults only*.
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.) Example: ETH X UC XX ID X.X-Xa – *"Required interaction with the operator - What do we expect from him?": Recommend that you estimate how many potential suggestions as soon as possible*.
GDPR	Check whether a solution follows the spirit of GDPR regulations Example: ETH ID X.X-XX – *Recommend that you ask operator consent for natural voice feedback related to Task X.X (and tasks of WPX generally as relevant) accompanied by clear information as regards the purpose and storage time of the voice data*.
Responsibility	Confirm or change who is responsible for tasks in some part of the process or what their new tasks will be Example: ETH X UC X ID X.X-X – Figure 6: *Recommend that you clarify whether the operator will be expected to review the root cause output*.
De-anthropomorphization	Change anthropomorphic wording or thinking about AI Example: ETH ID X.X-X – *Recommend that if X.X-X not implemented then have several sessions with process engineers and operators, to present clear mechanistic explanations of AI processes as un-intelligent tools*.
Simplification	Try simpler techniques first Example: ETH X UC X ID X – *Recommend that you consider using non-AI statistical process control method if feasible, to avoid greater complexity and error*.
Verify effects	Verify whether a proposed implementation would have some human effect Example: ETH X UC X ID XX – *Recommend that you monitor cohesion of operator team regularly watching for changes in social interaction of the team members; ask the operator team members*.
Timeliness	Implement certain other recommendations in a timely manner Example: ETH X UC X ID X (update) – *Recommend that you begin developing protocol at Deliverable X.X stage*.
Valorize experience	Make better use of human abilities/experience Example: ETH X UC X ID X.X-X – *"AI training task:" Recommend that you dedicate an experienced operator who has a demonstrated interest in this to do it and reduce that operator's work proportionally*.
Ethical rewording	Reword a text or redraw a diagram to better include the human contribution

	Example: ETH ID X.X-X – *Recommend that you use alternatives to 'exploit/exploiting' human activity and knowledge, i.e. alternatives which emphasize human participation*.
Workload	Estimate how much, how long, how many, of some new task to be done Example: ETH X UC XX ID X.X-X – *"Required datasets for solution development (Quality of Service – Feedback System):" Recommend that you estimate in advance how much time this will add to operator's or quality manager's workload, either as a whole, or per 'meaningful units of product characteristics.'*
Evaluation	Assess whether some aspect of the workplace context is taken into account 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 Example: ETH X UC X ID X.X-X – *Recommend that you clarify whether the blurry image situation for OCR recognition is also included in the X% error KPI*.
Training	Recommendations to provide specific training or implement services by stages Example: ETH X UC X ID Xc – *Recommend that you institute a trial period where AI suggestions are first cleared by process engineer, shifting responsibility to the latter*.

Table 1 Definitions of Ethical Recommendation Categories with Examples

3.2 Process of Categorization and Results

3.2.1 Methodology

Manual annotation, in this case categorization, is not about measuring a physical reality (such as the height of Mont Blanc), but about quantifying a phenomenon (Desrosières, 2008), which implies agreeing beforehand on conventions of equivalence: for example, in order to count unemployed people, there first has to be agreement on what defines unemployment.

These conventions should be then 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 project.

3.2.2 Process and Results

One of the authors categorized the recommendations already made at that point in the project - 120 total - into 15 categories (annotation #1). They wrote annotation guidelines, defining precisely 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%.

These results show that the categories needed to be reviewed. First some definitions in the guidelines such as the one for 'Timeliness', were improved (the emphasis on order of implementation was removed and the category was defined fully in terms of getting another recommendation implemented 'earlier'), then some categories were merged (e.g. 'Human centering' with 'Feedback,' and 'Training' with 'Adaptation')[6] and new ones added, such as 'GDPR.' Finally, all the recommendations were reviewed one last time, with each agreement case discussed and a consensus decision made on them.

[6] The categories that were merged are not presented in the table above.

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

It should be noted that new recommendations were added in the meantime and that 130 recommendations have now been made in total at this point of the project, with 9 of these being re-categorized as NA, as explained above.

4 Implementation Results of Recommendations

4.1 Comments on Methodology

The ethics team assessment of the results was an ongoing process, in which partner moves toward implementation were recorded as they arose, from about the sixth month of the project onwards. A final review of the status of all recommendations was then carried out prior to writing. We then presented all partners with an online worksheet of all recommendations and our assessment of their status, with instructions to give their own assessment of the implementation status of recommendations. If they agreed with the ethics team assessment they need make no changes. If they disagreed with our assessment, they could upgrade or downgrade, and give reasons for the change, as well as reasons why the recommendation was not implemented. After a five week limit, the data was gathered from the partner assessment for comparison with our own assessment. That data is presented in the Figures displaying partner assessment result. It can be considered as the independent partner assessment of the implementation status of recommendations, although in many respects it agrees with the ethics team assessment.

Note also that the ethics team did not directly push for implementation of the recommendations. In other words, the ethics team did not engage in 'applied moralizing'. The team's role was viewed as that of giving the recommendations and the reasoning behind them, discussing them in meetings, being available for any partners who needed ethical advice, and occasionally inquiring whether particular recommendations had been addressed for purposes of keeping track of results.

4.2 Overall results

In Fig. 1 and Fig. 2 below, the overall results as a proportion of the recommendations which were kept, are indicated, for each of the three outcomes: fully implemented, partially implemented, not implemented.[7]

[7] Note that the overall results as assessed by the Project Partners are based upon 108 recommendations kept and 22 rendered NA (as opposed to 121 kept and 9 NA for the ethics team). This difference is due to project partners deciding to formally abandon one UC fairly late in the project, thus rendering all of its recommendations NA for the project partner assessment. The ethics team, however, decided to retain their assessment of the 13 recommendations of the UC, since that assessment had been completed before the formal abandonment occurred.

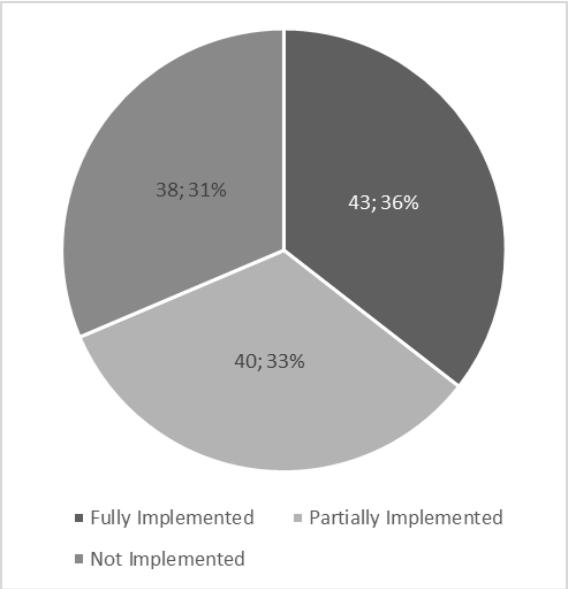

Fig. 1 Overall Results of Ethical Recommendations as Assessed by Ethics Team

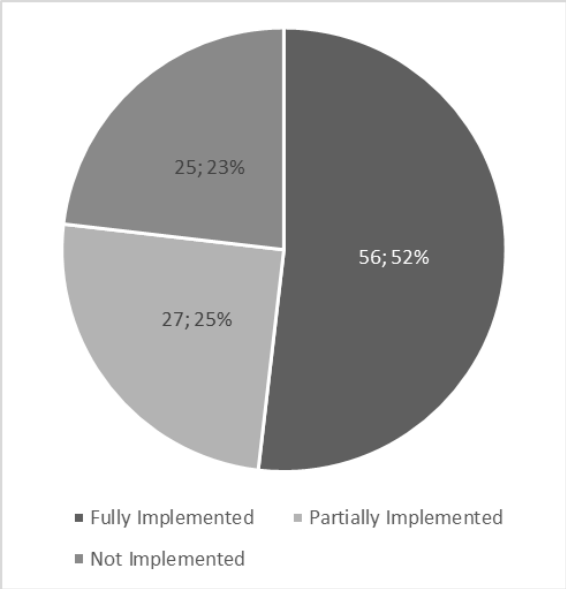

Fig. 2 Overall Results of Ethical Recommendations as Assessed by Project Partners

4.2.1 Observations Regarding Overall Results

As observed in Fig. 1, overall, 36% of the recommendations were implemented fully according to the ethics team assessment, with a further 33% partially implemented. 31% of the recommendations remained un-implemented.

As the project is not completely finished as of writing, there is a marginal possibility that some of the implementation results may change. This possibility is taken into account by the project partners in particular, as can be seen in the somewhat more optimistic overall assessment of implementation results presented in Fig 2. For the partners, 52% of the recommendations were fully implemented, 25% were partially implemented, and 23% remained un-implemented.

4.2.2 Discussion of Overall Results

The overall results show that *ethics can be operationalized* in the industrial AI context at ground level through the approach of embedding the ethicist in the technological development process. The difference between the rate of success as assessed by the ethics team and that assessed by the partners is also quite consistent. We did not for example, find the partners simply re-assessing the results so as to upgrade everything to fully implemented, which they were free to do. In fact, some partners even downgraded some results from full to partial, or from partial to not implemented. For example, deliverable commitments in one UC to a training and observation phase for process engineers with regard to digital twins, which the ethics team assessed as partial fulfillment, were downgraded to no by the respective industrial partner, and clarifications regarding AI model training input by industrial partners which the ethics team had accepted as full implementation, were downgraded to partial by one of the tech company partners.

Moreover, the upgrade toward fully implemented in the partner assessment of Overall Results from 36% to 52% drew equally from the partially and not implemented portions. In many cases, as indicated by the partners to us, the upgrade of certain recommendations to a full or partial implementation result is based upon the partner(s) insistence that full implementation will be and can only be achieved at the final deployment phase of AI services. The ethics team is confident that the results presented in our own assessment give an accurate picture of what has been achieved, but by including the partner assessment, there is room for arguing that the more objective figure lies somewhere between the two assessments.

One could still question whether this rate of success for ethics is practical however. There are at least two responses to that question. First, that *in a purely ethical sense practicality is arguably complex and unique in comparison with other domains*, and second, that *if being practical is to be measured in the same sense as quantitative evaluations of technical success, then the expectations for improvement should remain comparable to those for technical success*.

So, in the first place, regardless of numbers, the approach demonstrates a practicality appropriate to ethics as historically understood, in terms of success in getting the partners of the project to consider ethics at a specific level of generality – here the shop floor and tech design level – and in provoking ethical responses to the recommendations. Tech company partners and industrial partners are prompted and habituated to some degree, to think ethically, i.e. to expand their thinking beyond the narrowness of the purely technological solution, an outlook which one hopes will remain with them beyond the project. One can also begin to get some idea of what kinds of categories of recommendations are welcomed (e.g. Human Centering), or avoided (e.g. Evaluation), and thus of where to direct future reflection and then develop a practical approach to address those which are avoided.

In the second place, if the evaluation of technical achievements and the initial goals of the project are taken as a guide, the project aims for efficiency improvements under various technical categories which range between 1-9%. Thus, if the approach is to be measured by similar standards, it is fair to say that the total outcomes of 36% fully implemented and 33% partially implemented is at least as good as the project technical outcomes. Industrial process and manufacturing automation typically envision modest gains of 10-25% as practical, e.g. in efficiency (ZVEI, 2012). They do not aim at wholesale transformations of manufacturing processes. It would not be fair to ask more of a practical ethical approach insofar as it is held to the same quantitative standards and applied in similar contexts.

From a pragmatic ethical viewpoint, this suggests that the above two paradigms need not be exclusive. Ethics can and should have its quantitative tending aspect coexisting usefully alongside its more ideal and qualitative aspect. The former gradually builds into the latter, and looks to the latter as a simple map to keep oriented.

4.3 Results by Category

In Fig. 3 and Fig 4. below, the results according to category are given, for each of the three outcomes, as a proportion of the total recommendations under that category.

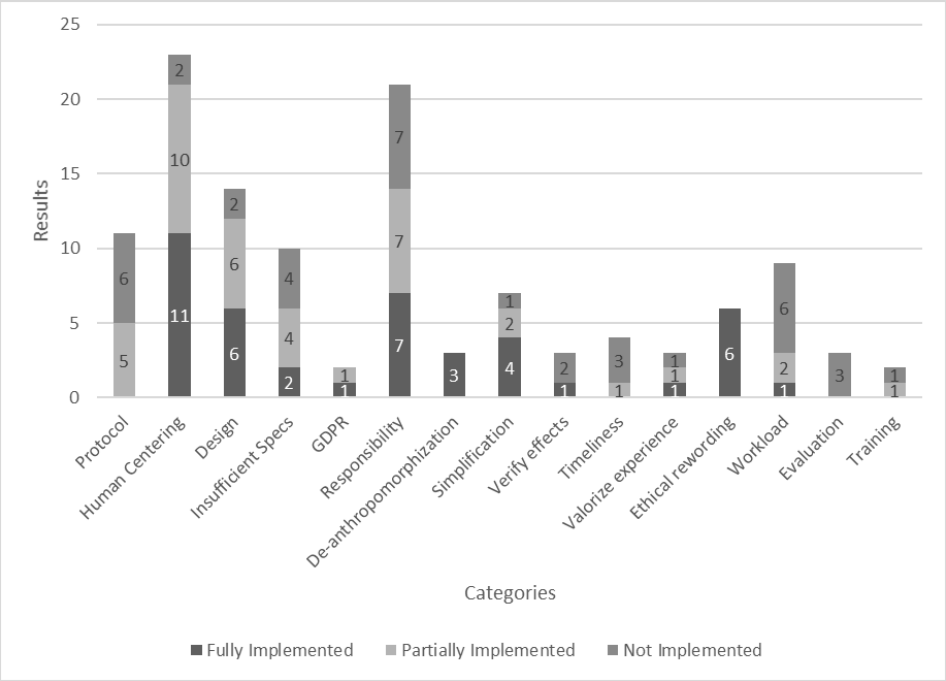

Fig. 3 Results of Ethical Recommendations by Category as Assessed by Ethics Team

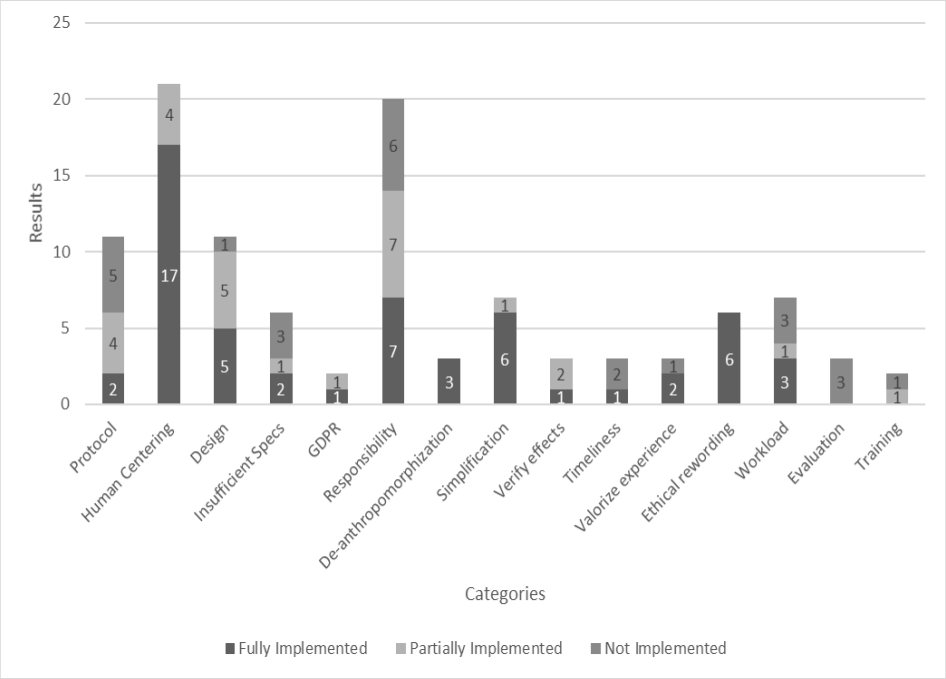

Fig.4 Results of Ethical Recommendations by Category as Assessed by Project Partners

In Fig. 5 and Fig 6. 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, with higher total recommendations for a category deciding ties in percentage, e.g. in Fig. 5, Responsibility has 21 recommendations compared to only 3 for Valorize Experience, even though both are at 33%.

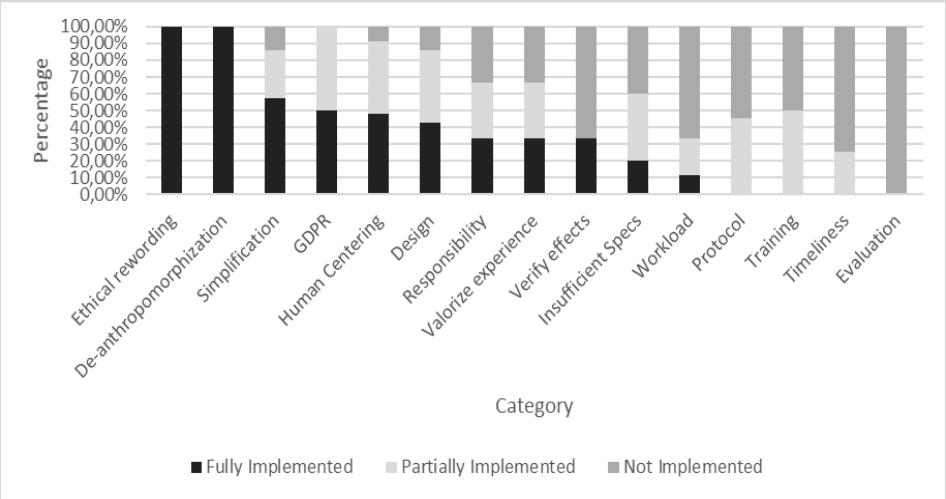

Fig. 5 Results of Ethical Recommendations by Category as Assessed by Ethics Team - Percentage of Total

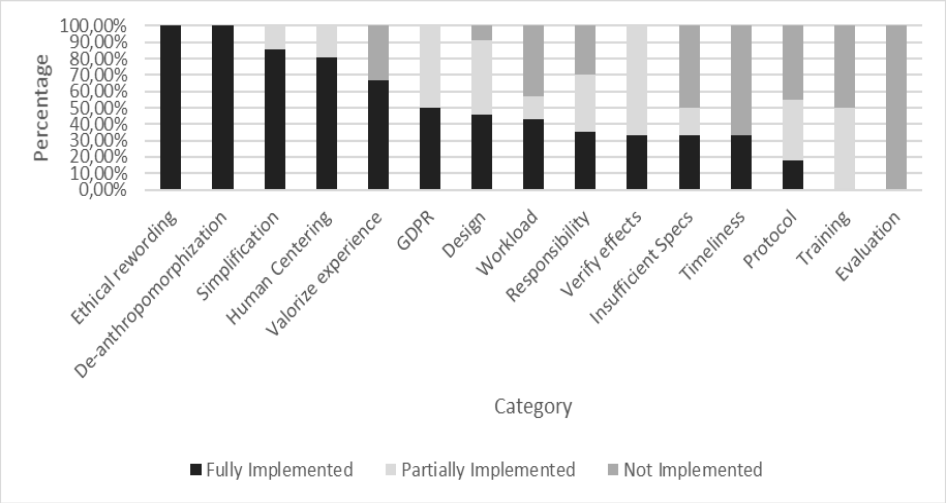

Fig. 6 Results of Ethical Recommendations by Category as Assessed by Project Partners - Percentage of Total

4.3.1 Observations Regarding Results by Category

Although somewhat offset by the fact that the number of recommendations for the categories in question are relatively small, the results by category, Fig. 3, contain two clear extremes. All recommendations for Ethical rewording and for De-anthropomorphization of AI descriptions within deliverables were carried out. No recommendations under the category of Evaluation were carried out. This did not change in the partner assessment, as seen in Fig. 4.

In the ethics team assessment, in each of the three categories with the largest total number of recommendations, well over half the recommendations in each category were either fully or partially implemented. In descending order of number of recommendations, these were Human Centering, followed by Responsibility, and then Design. The partner assessment did not differ substantially, Fig. 4, except for the category of Human Centering, where many implementations judged to be partial by the ethics team, were upgraded to fully implemented.

4.3.2 Discussion of Results by Category

In the results separated by category, what seems clear is that the recommendations which are most difficult to get implemented are those which aim to set user focused quantitative benchmarks for error and reliability – Evaluation – and recommendations to quantitatively estimate various aspects of new tasks to be given to the worker – Workload.

The reasons behind the poor implementation of Evaluation type recommendations may be that setting a benchmark can be taken as implying the possibility that *the system is not, and may never be, reliable enough to use ultimately, and so we should not use it*. Such an implication tends against the notion that tech can and should always find a solution, a notion deeply ingrained in software engineers and one which works against ethical engagement, as Avnoon et al. (2023) and Clark and Lischer-Katz (2023) argue.

If estimating quantitative figures in advance seems to be difficult and avoided, nonetheless it is clearly not impossible, since the project partners did it for recommendations under the best implemented larger categories of Human Centering, Design, and Responsibility. That quantitative evaluation is avoided is also counter-intuitive given the already noted software engineer urge to find quantitative and technical solutions. It seems then, that the root of the problem for recommendations of the Evaluation type does indeed lie in the ingrained notion that the technology can always succeed eventually, and conversely that it cannot fail or be inappropriate to a context.

The relative failure of the Workload recommendations in the project context seems more related to work and its historical and institutional frameworks as such, than in the attitudes of technology companies, even though it has a definite crossover with the above issue of publicizing problems with technology in quantitative terms.

Asking for workload estimates tends, in itself, to create more work to some degree, and it may not be easy to carry out such quantitative estimates. Both of these are issues that appear to be soluble however.

More problematic is that workload recommendations bring attention to the fact that developing and integrating a technology is often just *adding more work somewhere else*, (Crawford, 2021) and sometimes work that is never factored in, even though the overall stated goal is generally to adopt the technology to lessen the workload. If it turns out that in some cases there is merely a shifting of workload to others areas where it is not accounted for, then of course this *practically* contradicts the stated aim.

It does not *actually* contradict it however, unless formally acknowledged – an acknowledgement that many of the ethics team's Workload recommendations attempted to get – so that it is known in a manner such that the shop floor worker also becomes aware of it. Then it also tends toward requiring some justification and legal adjustment in relation to work contracts.

In support of this explanation, in the several instances in the project where the Workload recommendations were followed there was a tendency for the project partners to take on the extra workload themselves, or shift it from shop floor level to management. This is a relatively good result in view of operationalizing ethics, in the sense that bringing the extra workload out in the open with recommendations toward quantitative estimations, is something applied ethics can help to tackle, and it is at least a neutral result in terms of the worker. But it is not a satisfying result. A satisfying result would require a public and honest rethink on the reasons we are deploying technologies, and one which proceeds from the given context.

It 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 the 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.

4.4 Results by Category and Partner

In Fig. 7 and Fig. 8 below, are given, in heat map format, the results according to category, for each of the 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 sometimes the same recommendation was carried out (or partially) by one responsible partner but not by the other(s).

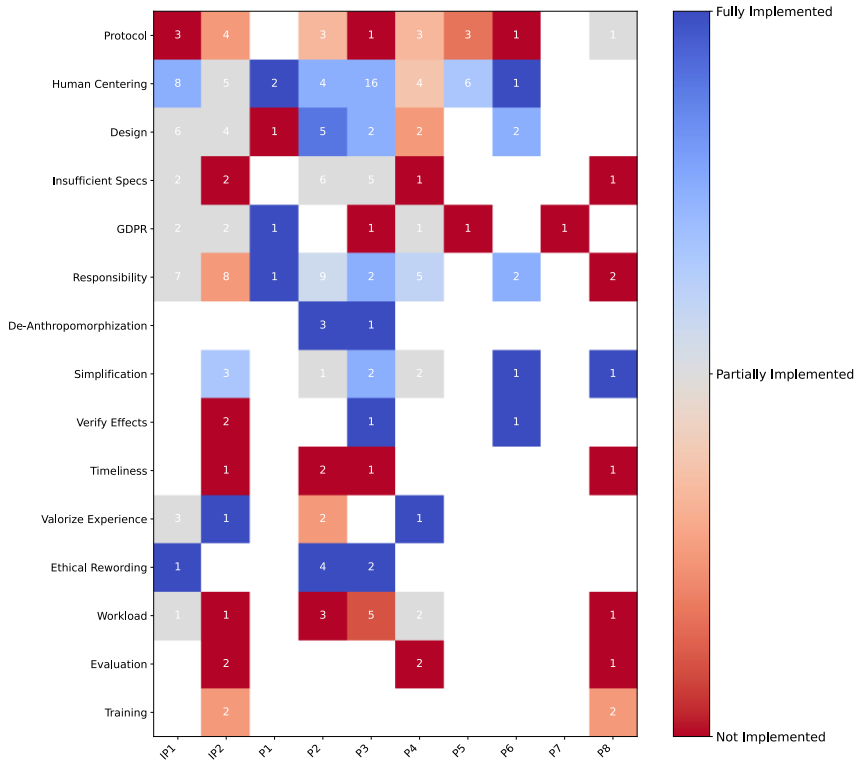

Fig 7. Results of Ethical Recommendations by Category and Project Partner(s) as Assessed by Ethics Team

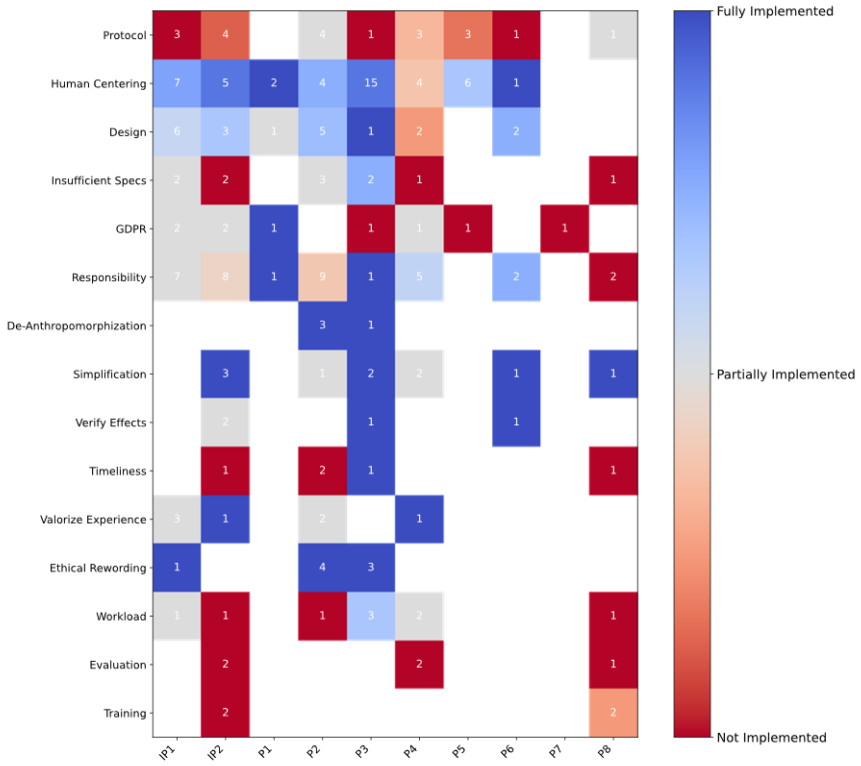

Fig 8. Results of Ethical Recommendations by Category and Project Partner(s) as Assessed by Project Partners

4.4.1 Observation Regarding Results by Category and Partner

In the heat map of results by category, Fig. 7, as assessed by the ethics team, further broken down according to partner responsible for a recommendation – sometimes multiple partners –, we observe that the conjunction of some partners with some categories produced a high degree of partial or full implementation relative to the total recommendations given for that category. Thus, to give as an example the specific results which arrive at the colour intensity indicated: tech company Partner 3 (P3) fully or partially implemented 16 of 16 of the Human Centering recommendations for which they were responsible and Industrial Partner 1 (IP1) implemented 8 of 8 fully or partially in the same category, while tech company Partner 2 (P2) partially or fully implemented 6 of 9 Responsibility recommendations and 5 of 5 Design recommendations. On the other hand, Industrial Partner 2 (IP2) only implemented 3 of 8 Responsibility recommendations.

In the heat map of implementation results by category and partner as assessed by the project partners themselves, Fig. 8, some movement toward the blue or fully implemented colour intensity can be observed both for some categories and some partners. But also, some category rows and some partner columns remain unchanged, e.g. those of the Evaluation category and the results of tech company Partner 8 (P8). Moreover, the same general trends in the project partner assessed heatmap can be seen, e.g. strong implementation in the Human Centering and Design rows, and relatively strong implementations overall by Partner 2 (P2) and Partner 3 (P3).

4.4.2 Discussion of Results by Category and Partner

In the results broken down according to partner in the heatmaps, it is clear that, allowing for the categories just discussed, some partners are much more willing to implement recommendations. This held between industrial partners viewed separately as well, since one of these partners had a proportionally higher rate of full or partial implementation than the other. It should be noted also that one tech company partner more or less ignored ethical

recommendations altogether, suggesting that they 'didn't really understand' the ethical aspect. Other partners were mixed in their engagement depending on the issue at stake and their interest in particular UCs. For these partners their responses to ethical recommendations were sometimes ambiguous, e.g. 'if it is efficient, we will do *x*', which can indicate that efficiency takes precedence over ethical concerns, but also that there is a 'way out' if implementation proves bothersome.

The Human Centering category had the best results in the heatmaps both as assessed by the ethics team and by the partners themselves, after adjustment for the number of recommendations. Recommendations to consider the workers (users) and collaborate with them in developing the services seem to be readily taken up most of the time. This may be helped by the fact – as noted anecdotally by the industrial partners – that the workers will tend not to use a new service if it doesn't work properly, so it is important to get the service right.

Primarily, however, it is also encouraged by the fact that Human Centering recommendations tend to be *positive*, in the sense that they require additional work, but they do not obstruct lines of technical development pursued by the partner. They can be operationalized by giving the workers options in the service, by considering how to build on technology already used by the workers, by clarifying who will do what, by discussion with process engineers and heads of work teams, and by giving short surveys which uncover the worker's background knowledge and expectations and adjusting accordingly. In other words, though the ethics team tried to make all the recommendations positive rather than prohibitory, Human Centering lends itself particularly well to this approach.

As noted earlier, tech company Partner 3 (P3) had the best record of full or partial implementation, proportional to the number of recommendations they were responsible for, closely followed by Partner 2 (P2). Multiple individuals within P3 engaged with us in implementing ethical recommendations over various deliverables and issues. Also, interestingly, and perhaps because of the ethical engagement in question, as can be seen in Fig. 7 and Fig 8., 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*.

The higher proportions of successful implementations for partners who were responsible for the largest numbers of recommendations seems to indicate a strong commitment in those partners to attempting to engage with the ethical aspect of the project. This impression was backed up by the ethics team's subjective assessments of different partner reactions to recommendations generally: the partners who had the highest proportions of full or partial implementation were also those who had a tendency to contact the ethics team regularly either for clarifications or to ask – without prompting – to set up meetings to discuss potential ethical issues in UCs which they led. These partners did not mind engaging the ethical aspect and even seem to have welcomed the engagement.

5 Limitations of our Study

Allowance should be made for the partners not having experienced a hands on and ground up approach. The ethics team considers that approach to be a relatively new approach to applied ethics of AI in industry. The partners did not know what to expect when the project began, and the ethics team developed the approach as the project went along. Accordingly, it may be that some of the earliest recommendations – some were made in the first four months – remained un-implemented wholly or in part because the partners were still unsure of the ethics approach. It was 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 very much on the attitude which the partner, as an organization, takes toward ethical concerns. This makes it difficult to quantify ethical results such that they could be taken to be completely objective and unbiased. But it also points out several facts which promise a way forward for ethics at the organizational level: *an organization can have an ethical attitude*, and if it can, then *that ethical attitude can be cultivated*.

It also leads logically to a related limitation: the roles of particular individuals in the ethical implementation results was very obvious, but the specific influences of those roles are missing from the results. A range of interest was observed among individual participants in each partner organization, with some individuals showing sustained interest in ethical operationalization. In terms of the written deliverables for which an ethical issues

section containing recommendations was explicitly included, the task leader in charge of the deliverable often pushed to have the section completed as part of the deliverable completion. This was not just a matter of mere routine however. 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 had been planned.

The bias which limits 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, to know that some organizations do not have a culture conducive to applying even direct ethical recommendations, leads obviously to asking why, and then to asking how such a culture can be instilled in the organization, and finally to asking how certain individuals working in the organization could serve as entry points for instilling an ethical culture.

6 Future Research

In a Deweyan friendly ethics the relation between the individual and society is the main focus of a practical application (Dewey and Tufts, 1932). In this project that relation is recreated at the level of the work organization. The 'individual' is the industrial worker or the individual software engineer and the 'society' is the group of industrial and tech company project partners. This is very fertile ground upon which to operationalize ethics. It can also frame the questions to be asked going forward.

Consequently, research 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 also be centered around 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.

For AI ethics in the heavy industry context and work context more generally, there is also an opportunity to study how an applied ethics approach 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 which could bring this contradiction out in the open of a public discussion seem to be especially called for.

Further, the fact that the partner assessment of results, arguably did not differ by a huge margin from the ethics team assessment – particularly since many upgrades were in a few categories and by a specific partner – also presents an opportunity. A good line of further studies here would be how much parallel ethical assessments and the possibility of peer review of ethical assessments at operational level, influence how tech company partners or industrial partners rate their own performance. AI Ethics in the industrial context could serve practically to integrate such studies with psychological and other related research in order to uncover positive best practices to be applied.

A final question is whether a ground level ethics operationalizing approach works as well in a more limited time frame, e.g. less than a year, or, conversely, does it get better results over a period longer than three years?

7 Conclusions

In this paper the results of a bottom up approach to AI ethics as applied in an EU Horizons 2020 project, which is now in its third year, have been presented. These results show how many of the specific ethical recommendations, given to project partners to date, were implemented, either fully, partially, or remain un-implemented. The results are described first as a percentage of total recommendations made, then as sorted under fifteen categories, and finally as sorted by both category and implementing partner. The process of categorizing recommendations is also described.

The ground up approach in question has been based upon carefully considering context, making recommendations, and monitoring and evaluating results. This paper is thus the more quantitative result of a larger attempt to operationalize AI ethics in Industry 4.0. From this more quantitative result, it was argued that

the approach works and that it discloses two difficulties in operationalizing ethics in this context: the tendency of tech companies to reject the possibility of not using the technology in most cases and the tendency of new technological integrations to add new work covertly which offsets and contradicts the stated aims behind the new technology additions.

This paper also discloses several interesting and less obvious results which deserve further study, namely that both the individual cultures of organizations like those participating in the project and particular individuals within such organizations, have a major influence on ethical operationalization. With these leads the authors are confident that future research can work to better understand how ground level operationalization of AI ethics in industrial settings can be advanced.

Acknowledgements

This research was funded by AI-PROFICIENT which has received funding from the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 957391. The authors would like to thank Bertrand Remy at LORIA for generating the heat maps of Figs. 7 and 8.

Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Data Availability Statement

All data generated or analysed during this study are included in this published article [and its supplementary information files].

References

Anderson, M. M., & Fort, K. (2023). From the ground up: developing a practical ethical methodology for integrating AI into industry. *AI & SOCIETY, 38*(2), 631-645.

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational linguistics, 34*(4), 555-596.

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

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.

Ciobanu, A.C. and Meșniță, G. (2022) AI Ethics of Industry 5.0 – From Principles to Practice. *CEUR Workshop Proceedings*. Vol. 3214.

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

Hagendorff, T. (2022). Blind spots in AI ethics. *AI and Ethics, 2*(4), 851-867.

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.

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

Singh, A., Dwivedi, A., Agrawal, D., & Singh, D. (2023). Identifying issues in adoption of AI practices in construction supply chains: towards managing sustainability. *Operations Management Research*, 1-17.

Tahaei, M., Constantinides, M., & Quercia, D. (2023). Toward Human-Centered Responsible Artificial Intelligence: A Review of CHI Research and Industry Toolkits. arXiv preprint arXiv:2302.05284.

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