AI·PROFICIENT

Artificial intelligence
for improved production efficiency,
quality and maintenance

Deliverable

D1.2: Legal and ethical requirements for human-machine interaction

WP 1: Pilot site characterization, requirements and system architecture

T1.2: Human-machine interaction, legal and ethical issues

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

Illustration Index

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 program under grant agreement No 957391.

Title: D1.2 Legal and ethical requirements for human-machine interaction

Lead Beneficiary	UL
Due Date	April, 2021
Submission Date	April 28th, 2021
Status	Preliminary
Description	Deliverable 1.2
Authors	Marc Anderson and Karën Fort
Type	Report
Review Status	~~¨ Draft~~ ¨ ~~WP Leader accepted~~ ¨ PC + TL accepted
Action Requested	~~¨ To be revised by partners~~ ~~¨ For approval by the WP leader~~ ¨ ~~For approval by the PMT~~ ¨ For acknowledgement by partners

VERSION	ACTION	OWNER	DATE
0.2	Adding text (MA) in template (KF)	KF	March 23rd, 2021
0.3	Taking into account feedback from UL	KF	March 31st, 2021
0.4	Taking into account other feedback	KF	April 16th, 2021
0.5	Review by WP leader	PA	April 17th, 2021
0.6	Review by PMT	NT	April 26th, 2021
1.0	Final Deliverable	KF	April 28th, 2021

Executive Summary

The Deliverable D1.2 is a public document of AI-PROFICIENT project delivered in the context of WP1, Task N1.2: Human-machine interaction, legal and ethical issues. This version of the deliverable will be updated to provide a finalized version at M9, in particular to deal with the legal part of it.

It contains an overview of the integration of ethics into the first six months of the AI-PROFICIENT project. The overview includes a literature review of the current and past work upon Industrial AI ethics, a survey of related guides and principles, and observations regarding the state of the discipline and its particular character relative to other fields of AI ethics

It outlines the methodology followed by the authors and then presents a general overview, along with two actual Use Cases from AI-Proficient project – anonymized for confidentiality – in order to show how the methodology has resulted in specific recommendations to the AI-Proficient partners.

A section on legal issues – to be filled out further in the version at M9 – outlines some of the relevant laws, international standards, and potential legal issues.

The document ends with concluding observations and the proposed strategy for embedding ethical considerations in the AI-Proficient project going forward.

1. Introduction

This report is prepared in the spirit of a work undergoing continued modification and improvement.

It has been prepared by the Ethics Team which includes Karën Fort and Marc Anderson. We have had the help of the Extended Ethics Team, which includes other members of the Université de Lorraine partner and also the ethics contacts designated by the other partners of the AI-Proficient project.

The Ethics Team is not a partner in the normal sense of 'partner' in the project. Nonetheless it is in another sense a 'transversal partner' to every aspect of the project, since ethical concerns run all across the project without respecting the usual boundaries. Our effort to bring others into the ethics discussion as part of the extended ethics team noted above, has been an attempt to recognize and apply this transversal aspect practically.

The Covid-19 pandemic has forced us to work remotely through video conferencing and electronic communications in most cases, but we have adapted as well as possible to the circumstances.

It might be helpful to offer, here at the outset, a definition of ethics. Ethics has been the subject of many definitions, definitions not always consistent among themselves. Given that, we can only make our best effort at a definition.

Assuming we can broadly define *awareness* as the response of processes in the world to other processes, then ***Ethics*** might be defined as: *the consistent integration of what aware processes in the world are able to actuate as value insofar as that value does not conflict with what other aware processes actuate as value* (Anderson, 2019).

The realization of the above definition would eventually result in a situation where the maximal potential value of the whole process we find ourselves in, 'the world', would be actualized by its sub-processes – including our human selves – as a continual creative harmonisation of action relative to all other processes.

But we begin in a world which includes disharmony of processes, including processes formed by the results of our past actions and the past actions of others. And ethics is practical according to the above definition, it is a matter of *action*, mentally and physically.

Recognizing this, the consistent approach for ethics is to work from the bottom up, recognizing the potential harmonisations of value in our physical contexts while expanding their applicability with principles in our mental contexts – a co-evolving effort of drawing principles from more specific experiences.

The ***Ethics of AI*** might then be defined more specifically as: *an application of the above definition to human created entities which display mental like tendencies in 'relative' freedom from human manipulation*. Practically this definition becomes a matter of constantly remembering that AI is a human product of our past efforts, that it will have consequences for the non-human processes in the world, that it is flawed insofar as we are flawed, that it is best developed together with an understanding and referral to the physical and mental contexts that it will be used in rather than separately from them, that we cannot expect it solve problems that we are not addressing without it, and (Jasanoff, 2021) that we should strive to integrate into it some of our most ethically related human tendencies which do not fall under the notion of 'intelligence', e.g. judgment, wisdom, experience, and tacit knowledge (and, we will add: consciousness, patience, and empathy).

All of these considerations would tend to harmonise it with regard to other value actuating aspects of our lives and make it an ethical AI.

More specific still, the ***Ethics of AI for Industry*** might be defined as: *an application of the above definitions to the human work context of industrial processes, i.e. the production or manufacture of predominantly physical products, with or without machinery*.

Practically this last definition is fulfilled by taking into consideration the *human relations* of the industrial employee in the work context, the particular *physical human created work surroundings* of the employee, the *stresses and hierarchical prescriptions belonging to work environments*, and the *effect* *that industrial AI integration will have upon the process of the products being created in that work context (since those products always stand in relation to their human creators)*.

And all of these in a flexible manner, since – if our definitions are helpful – the AI implementation for any given industrial context is dependent on many ethical choices that have come before and their ethical consequences radiate out far beyond the workplace, to the product as potentially harmonising with or polluting the greater environment, to the communities which the employee and employer belong to, and to the larger society.

Possible definitions notwithstanding however, the ethics of AI for industrial applications is, as we will argue, in its infancy still. This makes it important to be able to adapt to the context of the problems which the research of AI-Proficient is addressing. As the title of the deliverable highlights, *human-machine interaction* – and prominent within it, *human-machine interfaces*, as exemplified in our Use Case example #1 - is the central issue which our ethical effort must tackle. The central issue then is human grounded from the beginning, in the very physical industrial context of the worker manipulating machinery, rather than beginning in the realm of data relatively stripped of its human character. This *human starting point* – appropriately for the beginning of the project - guides our discussion and suggestions which follow. After a brief discussion of the current State of the Art in AI Ethics for Industry for purposes of comparison, we therefore begin with a review of the Methodology adopted so far, a review which includes the changes in method which came about through various realizations regarding the human or machine elements of each Use Case. This includes a brief discussion of the strategy for addressing human-machine interaction in AI-Proficient from the ethical side, and of the aspects of AI Ethics that we have come to view as particular to ethical Industrial AI development.

We then proceed to an Identification of the Ethical issues. This begins with a general anonymized overview of all the Use Cases and their context in the AI-Proficient project. It is followed by a review of two of the actual Use Cases which includes: a generalized description, identification of ethical issues, and finally recommendations (which were or will be given to the industrial partners) relative to the identified ethical issues.

An Identification of Legal Issues, uncovered relative to each Use Case follows. It includes a review of the main relevant legal instruments, an outline of the relevant international standards, and a discussion of potential legal issues. [add note on our limitations regarding legal issues in the first draft]

Finally, we conclude with a review of the Preliminary Results of adopting our method, relative to the choice of Use Cases, and a brief discussion of the Strategy Going Forward over the coming months in developing the ethical aspect of the project, rounds out the report.

It should be remembered that within the timeline of the project, the first draft of Deliverable 1.2 has been created before the AI/tech partners have fully specified their proposed contributions to the project. Accordingly, in its ethical review and recommendations, Deliverable 1.2 is heavily skewed toward the immediate context of the Use Cases as presented by the industrial partners. This is both necessary, because of time constraints, and appropriate, because timely ethical reviews of the immediate industrial contexts are essential, both for selecting the Use Cases to develop and for starting the development of the selected Use Cases off on the right track ethically, according to Ethics by Design.

Accordingly, as the project advances, the reviews and recommendations of future ethical elements of the project will become more balanced, in considering the proposed AI structures themselves, the development of the human-machine interfaces, transparency, explainability, and other particularly AI/tech related concerns.

AI ethics for industrial applications, such as the applications in the AI-Proficient project, is an applied ethics necessarily. This makes it important to make Deliverable 1.2, as much as possible, a tool to be consulted and applied by the Industrial and AI/tech partners of the project, as they develop AI services for the chosen Use Cases. We hope that it – complemented by further more detailed private level reviews and recommendations to the project partners – will be helpful in that respect.

It should be remembered however that the deliverable is still secondary to the actual ongoing discussion, analysis, recommendations, and implementation of the recommendations during the project. Whatever we may write about that process is always viewed after the fact. The real ethical task is to *guide the development of the AI services to be ethically sound in practice from the very beginning,* wherever we take that beginning to be (in the case of the AI-Proficient project this is the original choice even at the proposal stage to include an ethical component and take it seriously). That is Ethics by Design.

2. Current State of the Art in industrial AI Ethics

2.1. At the crossroad of many disciplines

The Ethics of Industrial AI is difficult to place precisely. AI is widely applicable, and each area of application raises quite different ethical issues, even within what could be reasonably categorized as industrial applications. Moreover, it is, as we argue in our discussion on the current state of the art, a new and developing sub-discipline, where nothing is quite obvious or exactly where we might want it to be. This is compounded by our view that an Ethics by Design should not be an exercise in ticking off checklists, even though we can make some use of such lists. Ticking off checklists can lead to a state of mind where the minimum is attempted to satisfy the list item – ‘following the letter of the law; -, whereas Ethics by Design, if practiced as a continuous collaboration with developers/designers tends to promote a more concrete training in ethical reflection which can then be applied to later stages of development and also future projects, i.e. a teaching and learning process.

Our suggestion then is that the Ethics of Industrial AI is a process of exploration and discovery of ethical issues, somewhere between the overlapping disciplines of production systems, machine learning, occupational health and safety, and human machine interfaces (see Figure 1), and must take all of these into account.

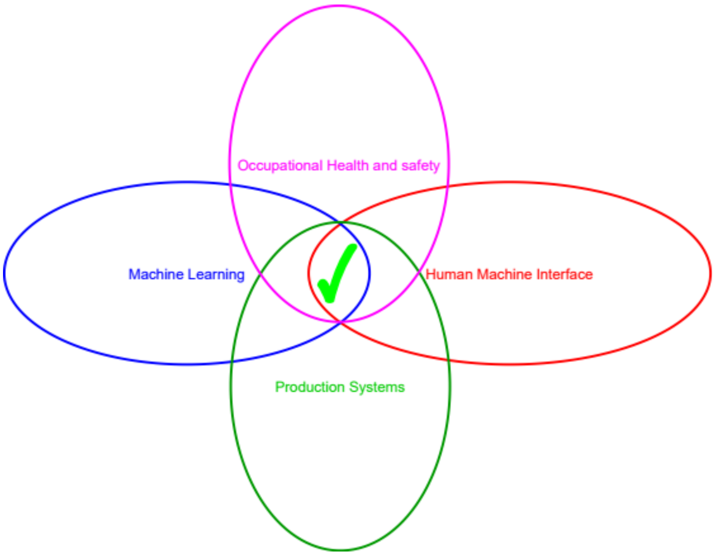

Illustration 1: At the crossroads of many disciplines.

2.2. Existing Checklists

A number of checklists, lists of key principles, and related tools, have been devised to categorize the issues around the ethics of AI development and deployment. Although, the checklist and principle approach is not our approach for the most part, for reasons noted just above and in our discussion of the State of the Art to follow, nevertheless we highlight here some of the main ones, in order to let the reader make a comparison with the approach we suggest.

Ethics Guidelines for Trustworthy AI (2019)

Developed by the EU Commission High-Level Expert Group on Artificial Intelligence. A guide created by 50 experts on AI after open consultations, presenting seven key guidelines that trustworthy AI should meet: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; and accountability. *Note that under the initial AI-Proficient project plan, Task 6.4 (M23-M36) envisions the development of ethical recommendations and practical principles specific to AI service development in manufacturing. These recommendations for manufacturing, growing out of the current Deliverable, will redefine at a more concrete level, the current High-Level Expert Group guidelines – which are very abstract – and in some cases complement them*.

https://ec.europa.eu/futurium/en/ai-alliance-consultation

Altai (2020)

An online assessment tool which takes the user through a series of increasingly specific questions regarding the context of an AI development and use. The tool is based upon the High-Level Expert Group (HLEG) on AI guidelines mentioned above.

https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence

Pour un développement des IAs respectueux de la vie privée dès la conception - Projet OLKi (Pégny, 2021)

A checklist of principles, accompanying and following from a much larger work laying out the reasoning behind the principles, which focuses particularly on ethical practices for gathering and using personal data.

https://hal.univ-lorraine.fr/IMPACT-OLKI/hal-03104692v1

Microsoft AI Fairness Checklist (2020)

A checklist developed on the basis of other checklists and through consultation with machine learning developers. Part of an article exploring the best ways of developing checklists as such.

https://www.microsoft.com/en-us/research/project/ai-fairness-checklist/

Deon

A command line tool to be added to data science projects to give guidance and a constant reminder (including a 'badge') to software developers to review their work.

https://deon.drivendata.org

Industry 4.0 Systems Framework and Analysis Methodology (Neumann et al., 2020)

A fairly comprehensive fillable worksheet to be used to parse out the human impacts (physical particularly) of Industry 4.0 integrations, including AI integrations.

https://www.sciencedirect.com/science/article/pii/S0925527320303418?via=ihub

Allistene-CERNA List of General and Machine Learning Recommendations (2018)

A list of very general principles of what researchers should considers when engaging in ML related research.

https://www.allistene.fr/files/2019/05/54730_cerna_2017_machine_learning.pdf

Montreal Declaration for a Responsible Development of Artificial Intelligence (2018)

A list of extremely broad principles to be followed in the development of AI.

https://www.montrealdeclaration-responsibleai.com

IBM Everyday Ethics for AI

A set of five ethical focus areas accompanied by recommended actions, questions for developers, and related issues to consider, along with an example illustrating the implementation of each principle.

https://www.ibm.com/design/ai/ethics/everyday-ethics/

IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, General Principles

A list of four very high level ethical principles modeled on documents such as the Universal Declaration of Human Rights.

https://ethicsinaction.ieee.org

UNESCO Preliminary Study on the Ethics of Artificial Intelligence

A study which at the end outlines a generic list of principles according to which AI should be developed, but more importantly expands on a further list of high level ethical concerns related to potential areas of application of AI, e.g. education, culture, peace, etc.

https://unesdoc.unesco.org/ark:/48223/pf0000367823

2.3. Discussion of Current State of the Art in Industrial AI Ethics

Work on AI Ethics for Industry is limited. Industry 4.0, a development which is open to multiple definitions (Tay et al., 2018), is currently a hot topic. But even though it coalesces somewhere around the concept of industry as Internet connected and AI supported, using autonomous Cyber Physical systems and networked sensors, and combining emerging technologies in various fields, it has a tendency toward being vacuous, and a real danger of being reduced to: the industry of ‘whatever is the newest thing.’

Hence, besides having multiple and vague components of which AI is only one, it is moreover a very future oriented concept, whose proponents sometimes appear to forget in their eagerness, that a great deal of heavy industry retains a very physical and mechanical character, certainly the character of industry 3.0 (and perhaps even that of Industry 2.0 depending on the country in question). It is so future oriented that without being clarified it has already spawned talk of Industry 5.0 (Kadir et al., 2019), which is envisioned as a post-fossil fuel, and biological based industry, scenarios which assume in advance successful solutions to some of our most difficult global problems. This oversight regarding typical working conditions in manufacturing and heavy industry perhaps explains the relative lack of direct industrial AI Ethics research in this area.

The research of (Trentesaux et al., 2017) and also (Trentesaux et al., 2021) – who themselves admit the scarcity of such research in the scientific literature – should be noted, dealing with the ethics of cyber-physical or autonomous-cyber physical human systems. This research, the latter for example, offers Industry 4.0 related case studies modeled with digital twins as proof of concept to develop design guidance for an ethical controller to be embedded in autonomous cyber physical human systems.

But again, this research is extremely theoretical, dealing with quite advanced systems of an order which are assumed to be capable of, e.g. not merely recognizing images, or correlating data, but both recognizing ethical categories of human behaviour and responding to those categories with AI based complex ethical behaviours. It is fascinating, it may be useful in designing the eagerly anticipated future, but it is – in the opinion of the authors – very much not the present need, at least in industrial contexts such as those considered in the AI-Proficient project. The industrial conditions addressed in the Use Cases of the present project, for example, conditions which we suspect – but cannot objectively confirm

– are the average for heavy industry located in manufacturing plants, when considered on one current scale for rating the autonomy of manufacturing plants (Gamer et al., 2019) would rate as 1 (perhaps in some aspects 2) on a scale of 0 to 5, where 0 is defined as complete lack of autonomy and 5 is defined as fully autonomous and completely absent of humans.

The present need is to ensure that the developers of AI services take into account what are still largely industry 3.0 conditions (physical, work time driven, dirty, loud, dangerous, sometime unreliable, high pressure environments) in developing what are essentially context based predictive or problem-solving AI services. These can be defined as AI services which have the potential to simplify or complement – but also to complicate and disrupt, hence the need for ethical consideration – certain human industrial tasks, or to correlate data from manufacturing processes whose variables are too complex to allow for easy human understanding.

If, as we suggest, there is such a lack of direct industrial AI ethics related research, the next best thing, in order to get a sense of the ethical state of affairs, is to fall back on research dealing with Industry 4.0 (and even Industry 3.0 as needed) but not specifically with AI – of which there is some –, and on research dealing with AI ethics of which there is much, but of very varying quality and practicality.

Until the *hoped* for Industry 4.0 transformation is complete – so that presumably the human element is taken out of the context of industry and the exploration of the ethical treatment of machines themselves comes into view – ethics must come down to the humans in the picture. If we are to be practical in the industrial context – and if we are not then we have no basis to speak of ‘industrial ethics’ – then its ethics begins first in the context of the workers in the manufacturing line, and the managers and supervisors who guide them, from where it expands outward to include other parties.

Of research which recognizes human centrality in considering Industry 4.0, the most thorough by far, is Industry 4.0 and the human factor – A systems framework and analysis methodology for successful development (Neumann et al., 2020). They rightly highlight the lack of attention to humans in consensus priorities of Industry 4.0 development, criticize the lack of human context in the emerging term Operator 4.0, and advances a clear framework which takes into account most elements of the human environment, including the physical.

From another angle, (Kinzel, 2017), in highlighting the lack of consideration of humans in Industry 4.0 research and discussion, adds to this a more elusive for quantification, but no less essential element in considering the ethics of Industry 4.0 work contexts: the human need to be included and feel purposeful, to have self-esteem, to self-actualize, while suggesting that mediators might play a role in this respect, a suggestion which runs close to our own as will be shown.

AI ethics in general on the other hand covers a broad field and focuses on many issues which are not directly relevant to the industrial context. AI ethics is heavily biased toward the ‘data viewpoint,’ at the expense of the human viewpoint, even though data as such can very much be said to help or harm human interests, depending on the use it is put to The objective and neutral ‘view from nowhere’, a product of the thinking of 17th century modern philosophy has come to ground the working approaches of tech fields such as data science and engineering, as Birhane notes (Birhane, 2021), and this in turn has focused the issues for AI ethics in general away from practical relevance to the industrial context.

But if the mechanistic philosophical worldview has given rise to the very machines which create the problem at issue, then on the other hand it is not surprising that mechanistic ethical constructions are created first in the attempt to solve the problem. Thus, as Mittelstadt notes, lists of principles mistakenly based on ethical solutions in a very different domain – that of medicine where human well-being has been the historical center of development – are rapidly proliferating as foundational AI ethics (Mittelstadt, 2019), despite being so general as to be practically useless, or being in fundamental contradiction to the actual assumptions driving much of AI development. An example of the latter can be found in (Pégny, 2021), in suggesting that machine learning (ML) models be made publicly available when possible, in order to test their security properties, whereas it seems likely that some of the most used ML models will be developed under proprietary conditions for profit, thus possibly rendering the principle moot from the beginning.

Principles if well founded, may give guidance, but they need interpretation to give it. They cannot work in the face of an unwillingness, or an inability, to do the work of interpreting and applying them, just as codes of ethics have been shown in quantitative research to have no effect in software development when their interpretation is left to software developers (McNamara et al., 2018).

The work ahead for AI ethics in general, and for industrial AI ethics in particular, is twofold: first, to leave aside the nice formulations of principles – we have enough of those – unless they are immediately practical ground level principles (an example of which will be given in what follows), and secondly (Morley et al., 2020) to move from principles to practical implementations.

In short, what we suggest is that, largely: there is no clear State of the Art yet for industrial AI ethics. It remains to be created. It must gain balance by taking into account the human element, and – contrary to the trend toward Operator 4.0 which tends to view the human as just another factor to be 'integrated' into Smart Factories (Gazzaneo et al., 2019) – it must pause to consider the human worker's point of view and wishes (Wioland et al., 2019). It must go still further and re-discover the human element in what has hitherto been regarded as the pure 'realm of data' in which software developers play; software developers are human and software, algorithms, etc. are human creations.

It must become practical, which includes addressing the changing human process and not simply the rationalized and categorized 'human as object,' the so called 'human resource.' Even Neumann's framework (Neumann et al., 2020), which is quite thorough as far as it goes, cannot address human sourced but emergent characteristics introduced with AI in the human context of its use and development.

Moreover, as far as we know, there is no research addressing potential uses of relational tending logic, i.e. dynamic logic applied to processes – such as the process of a manufacturing line – and thus departing from the perhaps too neat categorizations of logic (as well as reasoning) offered by Birhane.

Dynamic logical thinking involves thinking in terms of tendencies in processes rather than in terms of objects (the human in a workplace too easily becomes an object rather than a process). It particularly addresses temporal relations.

An example of this will be seen in Use Case #1, where the envisioned AI human collaboration can be viewed as a process expanding the time for an action (insofar as the human operator must interpret potential AI errors), even while it is alternately viewed as a process shrinking the time of the action in question (recognizing and delivering the product label text to a central computer with the help of the operator).

The two *tendencies* are then in contradiction logically, which leads to the practical, but dynamic, questions: which tendency predominates? and at what point does implementing the AI service become a practical contradiction in terms of the goal? The latter is a question whose context is *fluidly expansive or constrictive* depending on how wide or local you selectively view the issue: immediate surroundings of the operator, operator team, section of the industrial plant, etc.

The suggestion from what follows is that an industrial AI ethics – and potentially other branches of AI ethics – must be a customized effort, rather than one of consultation of principles and checking of lists. It has been suggested (Morley et al., 2020) that Ethics by Design operates by constraints and should be replaced with pro-ethical design which leaves open the choice for the agent to choose the un-ethical. But this well-meaning suggestion forgets that for any given context there are a multitude of ethical and un-ethical factors at play beyond the local context, often unexamined social constraints, which 'game the system,' and push agents in the local context toward the un-ethical.

So, to take an example relative to industrial AI ethics: the shop-floor worker is subject to a work hierarchy, with an immediate superior, as well as a work contract, both of which are higher level constraints (relative to pay, productivity, etc.) which will tend to sway mere free choice toward the un-ethical, even to the point (Lefeuvre-Halftmeyer et al., 2016) where the worker will accept considerable physical injury in order to satisfy external expectations of productivity. To take another example relative to AI ethics in general: the software developer is usually under deadline and under contract, and has not the time to apply ethics in such conditions, without help.

Better then to offer constraint as practice – the practical – within a range, where a range of choices, all ethical, can be selected from. But opening up such ranges requires time, and wise, skilled, and experienced interpretation by human elements dedicated to the local context, i.e. ethicists who can stay with a project through its development, immersing themselves in the process of the context, offering ethical alternatives at each step, and creating a full and creative Ethics by Design.

Ethicists may be in the context of industrial AI ethics, the very mediators – or at least one type of them – which (Kinzel, 2017) argues for. Certainly, they must counterbalance the contemporary taste for mere analysis in ethics with the more ancient and venerable philosophical view of ethics as action, as (Ocone, 2020) notes.

In short, practicing ethicists must descend from the realm of principles in the sky buried in journal articles and get their hands dirty in the soil of the industrial process. This approach is what we are arguing for here.

3. Methodology

Our methodology has included a number of connected tools and practices, detailed as follows.

3.1. Tools

The main working platform used by the AI-Proficient project partners for document sharing, meetings between partners, and general collaboration, has been the Microsoft Teams platform. Thus we have created an AI-Proficient Ethics channel on Teams, as a space to organize ethics related meetings with partners and post ethics related resources. At last count 18 members of the AI-Proficient project, spanning all partners had signed up to the channel.

An AI-Proficient ethics mailing list was established as well, along with a specific request to other partners in the project to designate an ethical contact with whom we could discuss ethical issues. All partners provided designated contacts, who were then added to the ethics channel and mailing list along with other interested project individuals. Our ethics mailing list currently has 13 members.

Video conferences in various formats, often accompanied by PowerPoint presentations with detailed pictures of the work environments, have helped us understand and discuss the ethical issues with the industrial partners.

The Covid-19 situation and subsequent cancellation of real-life visits has made the in-situ analysis of the work environments more difficult, but a real-life tour made by the senior member of the ethics team at the beginning of the project and a live virtual tour of one of the industrial plants at the opening kickoff were very helpful in getting a sense of the work environment.

The Deliverable 1.2 itself, in its development as an internal report has acted as also served as a tool to clarify and get feedback from members of the extended ethics team.

3.2. Practices

All AI-Proficient meetings which were open to us up to this point, were attended by at least one ethics team member, including kickoff meetings, working group meetings, and Q&A sessions, some 25 meetings so far since the project began, ranging from hour to day long meetings. In this way we have tried to alert the other project partners in advance to the ethical issues, challenges, and discussions that would be needed as the project develops.

We have also attended numerous meetings internal to the UL team and organized ethics specific meetings. Finally, the members of the ethics team have met weekly since the beginning of the project to discuss all related ethics issues including the means of evaluating the performance of the AI, establishing a baseline context for each Use Case, developing typologies for the Use Cases, etc.

Meetings with all the partners of the AI-Proficient project have been carried out remotely by video conference, due to the limitations imposed by Covid-19. We have been able to have several of the meetings internal to the UL team as well as weekly meetings of the ethics team in real life, although the majority of internal meetings have also been remote.

Extensive notes have been taken in all meetings, in order to get an immersive sense of all sides of the Use Case contexts as well as the points of view of the project partners, in order to be able to generate and focus on the important questions to ask as we go along.

Considerable research of the available scientific literature on industrial AI ethics – selectively cited above – has been carried out in order to establish the current State of the Art in this area, and the ethics team is co-writing an article to be submitted to one of the major AI ethics conferences of 2021 to highlight the insights gathered specifically from the AI-Proficient project. To these research efforts has been added meetings and discussion with experts outside the project in closely related fields, e.g. from members of the INRS (Institut national de recherche et de sécurité pour la prévention des accidents du travail et des maladies professionnelles).

This latter in particular has been part of our strategy for addressing the ethical issues of human-machine interactions, i.e. to recognize that there are complimentary research fields, e.g. worker safety, studying human-machine interactions very carefully beyond -but overlapping with – the field of AI research, fields which we can profitably consult with. As we go forward we hope to engage in many discussions with other such researchers so as to integrate their insights into addressing the issues we uncover in industrial AI ethics.

3.3. Current Work Situation as a Baseline to Consider Future Changes

In the original AI-Proficient proposal, the human role in human-machine interactions, is envisioned as being categorized under three terms, namely: human-in-the-loop, human-in-control, and human-in-command.

The term human-in-the-loop predates its current use relative to AI development. It can be defined as a human interacting with a simulated model, usually in order to use the outcomes of the interaction to uncover problems, or test new procedures for processes simulated by the model. The simulated models are not necessarily computer related. In the context of AI the term has come to mean human interaction with an AI model which trains the AI model, by validating or correcting some of its outputs. It is one form of machine learning.

Human-on-the-loop can be defined as the possibility of human intervention in the initial design and ongoing monitoring and supervision of an AI system.

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.

All three terms – except the first, in its original non-AI meaning – are relatively new. All are somewhat inconsistently used in the current literature, the latter two being only substantially defined in the High-Level Expert Group guidelines, and present evidence of being buzzwords rather than relatively stable concepts. The definitions of all three terms display a tendency to view the human as a complementary component – an extra – inserted into a system which is otherwise presumed to be intrinsically human free ideally.

Moreover, their tentative definitions are too broad to catch the subtleties in the industrial context (and perhaps in other contexts as well). Consider for example the human-in-command categorization. First, the very notion of being in command covers a range which is not easily pinned down until some *localization of context is deliberately selected*. In the context of AI integration for autonomous vehicles that context is quite localized: a driver is in command of the vehicle insofar as the driver is free to adjust the speed, direction, etc. of the car. In that localized context the command is very strong indeed, to the point where we easily speak of the driver having full command over the AI, but still only up to a limit.

The command of the driver over the AI cannot be greater than that which the driver has over the context of driving in the broader sense however. The deliberately set limits imposed by vehicle manufacturer and government relative to maximum engine output/speed, automatic braking, etc. are a form of command beyond the command capabilities of the driver, regardless of whether they are active in real time vis-à-vis the manufacturer or government. Moreover, if the selected context becomes less local –

i.e. broader –, the driver is even less in command, e.g. instructions from highway patrols to pull over, speed limits, the actions of other drivers, and the very design of the highway itself exercise a higher level of command over the driver.

But if the driver, already considered to enjoy a very high level of command, can be shown to have limits to that command, then the factory operator in the industrial context is even more limited. Taken at the level of the factory operator as *employee of a company*, the factory operator works under at least three significant constraints which affect the ability to be in command: the financial motives of the organization, the motives of production speed and efficiency, and a work hierarchy. All of these chip away at the strength of ‘being in command’.

Thus the characterizations of human-in-command, etc. are insufficiently flexible, particularly when combined with a view of the work context – all too easily adopted – as a set of abstract objects, which include the human as object, engaged in simple, tidy, and straightforward interactions with one another, rather than a complex ongoing process which easily changes depending on the differing viewpoints, motives, time ranges, etc., being considered. To take a blunt example: the shop-floor operator may be in *presumptive* command in interactions with the AI, until the boss says to the operator: “we need to get x done today, let the AI do it.”

It may be objected that the messy world of human hierarchies, etc. is the problem and issue here, not potential implementations of an AI, which can readily and straightforwardly be configured for the human-in-command characterization, e.g. with a ‘kill switch.’

But that is just the point. AI tends to be envisioned as being developed in the tidy world above and beyond messy human life (and in fact AI development has its own version of the above messiness at the development level, since programmers are also employees), and then be applied in the expectation that it can be integrated easily with that human messiness, without a careful and slow consideration of the challenges of the latter. Our suggestion is that it can’t – if it is to be robustly ethical –, and that too simple characterizations like human-in command, etc., used as labels to be stuck on after the fact, are misleading and give a false sense of safety.

The slower and more deliberate consideration of the context of application – here the industrial work context – where the degree of localization is selected, but also widened or restricted at need is the path to follow in our opinion

And that path begins evidently with an exploration and characterization of the current state of the work context: ‘what is happening now before we begin?’

This is how we see an Ethics by Design process: once we understand the current state of affairs we can go on to consider what changes are envisioned, in order to then consider the ethical issues around those changes and give ongoing recommendations while the AI services are being developed. Those changes are relative to the current work situation for the people involved in each Use Case.

Taking this into account, we identified three different types of Use Cases:

1. The Use Case is a new task for the operator (no baseline).

2. The Use Case modified the task of the operator (baseline to be established)

3. The Use Case does not modify the task of the operator or makes it disappear (no baseline needed)

To better understand the Use Case work context, we developed an initial list of baseline questions. (see below)

The questions were adapted and re-worked for each Use Case, as various hidden aspects of a particular Use Case come to light (some aspects of a Use Case context only come to light through sustained discussion). So, for example, relative to Question #5, the subjective metrics of the Use Cases have been illusive partly due to the time constraints, but also due to the considerable preparation involved in accessing the employees of the manufacturing partners, and even in knowing how to approach the employees (but we hope to accomplish this over the coming months).

Moreover, questions have had to be adapted when it was discovered that more people than the just operators would be affected on the fringes of the Use Cases, e.g. maintenance workers, technicians, etc.

Taking expansive notes on the context and on issues which cannot clearly be compartmentalized by the questions have helped fill out the pictures of each Use Case, including in particular for an industrial context, the physical and temporal environment of each Use Case. Gaining a good understanding of the environment of each Use Cases is yet another part of our strategy for addressing human-machine interactions, which are very much bound up with time, space, and the human body acting within them. This speaks to the need for a customized ethics, in which the process of assessment by the ethical specialist is immersive and ongoing rather at this level than a mere quick acceptance of the context as initially described by the manufacturing partner.

It will be seen then, that the baseline questions are very much centered on the operator in physical context now. To balance this, a new set of future looking questions relative to the technical development and related and changing wishes – and uncovered limitations relative to those wishes – of the manufacturing partners, will have to be created.

Once a preliminary picture has been formed of the Use Case context and the issues particular to it, we have gone on, after a period of reflection using logical and critical analysis, to make preliminary recommendations, both in outline form, and in some cases more in depth in order to explain the reasoning behind the recommendations. This step is important even at the early stages of the project, in order to help the tech and manufacturing partners get a feel for the ethical issues to be solved, and also to help in selecting the most ethical promising Use Cases for advancement.

As the project advances, we hope, based on our ongoing ethical discussions with the partners, to be able to develop and offer potential alternatives or extensions to the terms human-in-the-loop, etc. noted above, flexible terms linked to a more personalized assessment of the particular issues of a given work context.

3.4. Baseline Questions

#1 Is the task already done by the operator or is it a new service to be implemented?

#2 What is the actual chain of responsibilities?

#3 What are the present interactions between humans? Are there hierarchical relations? Who is making the decision(s)?

#4 What are the existing tools? & Objective metrics: What is the quality obtained today? Speed? Number of issues (human error, machine breakdown, etc.) per month/year? Time lost?

#5 Subjective metrics: How does the operator feel about the present process?

#6 How many operators are involved for each Use Case and how much do they work together as a group/team?

#7 Are the operators changing regularly for each Use Case?

#8 When a task in the Use Case is done always by the same operator: how long has the operator been doing that task?

#9 What do the existing user interfaces look like (if any)?

#10 How much can actions/processes in the Use Case be traced presently?

#11 What is the success rate for all Use Case at the present time?

#12 What is the minimum success rate threshold to be reached by the service in order to allow it to be deployed?

#13 What are the cost of errors/scrapping production, etc. for each Use Case?

#14 Does the operator have to move in the context of the Use Case?

#15 In which environment is the operator moving and how often?

3.5. Identification of Ethical Issues

General Overview of Use Cases

The AI-Proficient project is centered around three European manufacturing plants belonging to two global scale industrial companies and a number of European AI/Tech partners, led by the Université de Lorraine.

Managers from the three manufacturing plants presented a total of thirteen potential Use Cases for consideration with the understanding that a number of these Use Cases would be chosen after preliminary consideration of technical, ethical, and legal issues relative to each. Each Use Case was chosen on the basis of some problem that the respective plant managers wish to resolve with the aid of AI services.

The Use Cases are thus all cases intended to solve some existing – in some cases long standing – problem, which has eluded the efforts of the industrial partner employees – ranging from plant engineers to shop-floor level operators- to solve. This is important to point out because in this sense the Use Cases are firmly embedded in a well-established context, very often a physical context, and their development and integration will have definite and specific effects on operators and related workers, on engineers, and on lower level management.

The Use Cases cover a range of contexts. We have made a preliminary categorization of the cases by type: the Use Case creates a new task for the operator (no baseline), the Use Case modifies the task of the operator (baseline to be established), or the Use Case eliminates or does not modify the operator’s task (no baseline needed)

Beyond the broad categorization each Use Case has its own particular challenges and problems:

- Some of the Use Cases are more data oriented, so that the intention is to use AI in order to uncover correlations in the data gathered relative to various manufacturing processes. These correlations will then be used either to optimize the process or correct some issue which human understanding, or human capacity to make adjustments, has so far failed to correct.
- Other Use Cases have a more physical context, where the use of AI will potentially change the way a human operator interacts with the manufacturing machinery in a physical way, including interactions within operator teams and beyond them.

Human-Machine interfaces, an issue which we have not discussed so far, are central to a number of the Use Cases, as will be seen in our first Use Case example. This has raised the specific issue of adapting interfaces to industrial conditions, e.g. with operators wearing gloves or other safety clothing, operating in busy and dangerous environments, etc.

Some of the Use Cases address issues where preventive maintenance guided by AI is needed, which involves parties beyond the shop-floor operators, e.g. die-makers and maintenance technicians.

All of the Use Cases are part of manufacturing processes where time is crucial, and production is ongoing and consequently – particularly for financial reasons – cannot normally be halted for extended periods in order to take decisions or make changes. The issue of time thus becomes central to all of the Use Cases, time for the operators/engineers/managers to react, time to understand the results of potential AI interventions and guidance, etc.

New sensors systems, and new collection of data will be needed in some of the cases, other cases can be addressed by correlating historic data. The data for all cases is not public data, but proprietary data, which makes for a shift of focus from the usual data concerns of AI ethics. Moreover, this is not personal or sensitive data, insofar as the traceability of the operator's decisions – if traceable – are allowed for in the context of being an employee under contract. This does not mean of course that the data could not be used to trace the operator's actions (errors or omissions); in many cases it might. But the ethical response to such possibilities will have to *complement* the legal framework and government regulations under which the work contract is drawn up, e.g. by recommending a formal clarification of that framework and contract to clarify to what degree employee data is personal or sensitive, and thus avoid such issues.

Finally, all of the thirteen Use Cases are more or less directed toward finding solutions which will decrease waste, and increase productivity, product quality, and consequently profitability for the manufacturing plant. This, coupled with a definite hierarchy in the general work context, has implications for the issue of responsibility of, e.g. operators, but also of the managers, since reaction to the proposed AI services will be made according to differing motives.

4. Two Examples of Use Cases: description, ethical issues, recommendations

4.1. Use Case Example #1: Image Recognition of Labels

Description

The first Use Case example is centred around a problem regarding the labels on large bags of certain industrial products which are added to the manufacturing process in one of the participating manufacturing plants. There are a number of platforms feeding into hoppers (feeders), each of which are numbered and monitored by a console operator working at a central control board in the plant, who can tell e.g. the weight of the bag contents upon the platform.

Large bags of product are brought to each platform as needed by a loading operator and lifted onto the platform with a hoist. The label of each bag is then physically removed and brought to the control board room by the loading operator, where the console operator – or sometimes the loading operator who brings the label – then inputs the label number manually into the central control system through the interface of a standard mouse and keyboard.

The central control system then checks the lotnumber of the bag label against known recorded lotnumbers of the needed product to ensure that the bag is the correct product needed for the particular manufacturing process desired. It notifies the console operator with an alarm if the label is incorrect.

Sometimes, however, the loading operator has to do another job before bringing the bag label to the control room, so that perhaps 30 minutes have already passed before the number is entered into the system and – if the bag is the wrong bag – the wrong bag product has already begun to be added to the manufacturing process. Sometimes an incorrect bag is hoisted onto the right feeder, or alternately, a correct bag is hoisted onto a wrong feeder. In the latter case the console operator usually notices the error through monitoring the weight of the bags on each feeder and seeing significant weight where there should be none. The problem to be addressed is the need to speed up the input of labels into the central control system in order to detect earlier when an incorrect product is introduced into the desired manufacturing process.

The goal for the Use Case is to introduce an AI safeguard incorporating text recognition, wherein the AI will scan the label of the bag of product for the name and lotnumber, and the number of the feeder which the bag is on, and show these text recognitions to the loading operator. The operator – using a tablet for the new task – will then visually confirm the accuracy of the text recognition and press a submit button, which will send the label data directly to the control system, which will then give a further 'green light' to the console operator or loading operator if the bag product is ok to use. The hope is to have a text recognition success rate of 90 to 95%. If the success rate is below this range the management expects that the loading operators will abandon the text recognition AI tool as unworkable.

Ethical Issues

1 - It should be remembered that the central concern is one of time. i.e. the loading operator either forgets, or delays, going to the control room in a timely manner. Errors in the control system check of the bag product information, once input, were not mentioned in discussion sessions. Thus, resolving the time delay would resolve the problem presented in the Use Case. So the question to bear in mind overall will be: will the AI image recognition service shorten or remove the time delay between the bag being placed upon the feeder platform and its being cleared by the control system?

Insofar as it cannot do this it will be tend to be un-ethical in adding needless work and stress to the job of the loading operator, and also adds further complexity – open to errors – into the overall system.

2 - In order to shorten the time delay, several conditions will have to be met.

The interface must be reliable and practically useable by the loading operator, since the loading operator has to read it and also manipulate it physically. In this case this means it must be easily readable under the conditions in the bag loading area, including particular lighting conditions, and perhaps dusty conditions. It must be easily graspable under the working conditions in question – which include wearing work gloves, according to the Q & A sessions – and provide ease of input for any input of information.

3 - The reliability must extend to the AI's success in image recognition. The stated minimal requirements for accuracy were 90 to 95%. This raises a number of issues:

First, how and by whom will the accuracy rate be checked? The operator should not be expected to have to check the accuracy unless this is specifically added as a new task.

Second, assuming the image recognition accuracy rate reaches 90 to 95%, there are still 5 to 10% errors to account for. In those 5 to 10% of error situations, will the operator be expected to have the AI re-scan the bag label hoping the AI gets it right eventually, and/or will he be expected to manually correct/supplement the AI's effort? If the latter, time (and stress) are added to the loading operator's job; a shift of responsibility is made (i.e. the operator now has the potential to input his own mistake in correcting the AI); and the complexity of the input interface under working conditions will also become a more significant factor (i.e. it will have to be more than a single 'submit' button)

Third, what will the protocol be in case the loading operator makes a mistake and presses submit on an inaccurate text reading by the AI? (habit forming and consequent response errors in repetitive tasks where errors are rare is a known factor to be considered)

4 - All time delays caused by errors by the AI, potential corrections needed to be made by the operator, etc. (as noted above) will offset the time gained in sending the recognized text directly to the control system. This has to be considered if the goal of the AI integration is to be practical.

5 - In having to visually check the text output of the AI's image recognition, the loading operator is informally being designated as the 'safeguard for the AI,' which gives rise to issues of who is in control formally, and where responsibility lies for errors.

6 - The console operator is being taken out of the process as a safeguard insofar as the loading operator has become responsible for checking the AI results. One level of human safeguard on the production process will thus be lost.

7 - Hypothetical Scenario: a text recognition error has been made by the AI (the 5 to 10% inaccuracy) but not caught by the loading operator, who has pressed submit. The text is sent quickly and the control system gives an alarm (or no green light) to the console operator. How does the console operator know

whether the error is a rare AI related error or an operator bag placement error, also rare? The console operator no longer has the bag label to check where the error lies.

8 - An added responsibility for error is thus transferred onto the loading operator by implication. Whereas before, the operator was responsible for a mistaken bag placement, now the operator is also potentially responsible for failing to accurately spot an AI text recognition error.

9 - It is unclear so far in the Use Case proposal who is getting the green light, and subsequently where the responsibility for releasing the product bag into the manufacturing process lies, or if the responsibility for this is changing.

10 - The physical activity of the loading operator is undergoing a change, as there will be no more trips to the control room in this context.

11 - The social interaction aspect of the operator team is undergoing a change, since the loading operator and console operator will no longer meet in the context of the bag label being delivered. It may have an effect on team cohesion (an effect shown in related research studies)

Recommendations

1 - The technical partner developing the tablet/tablet interface should work with the operators to design and then test it (as a mock-up before AI integration) in the actual work conditions, and then adapt it based on operator suggestions to insure that it works in the particular lighting, in dusty conditions, with work gloves, whether it is an appropriate size, shape, and weight for the work conditions, and whether it is rugged enough to stand dropping, bumping, etc.

2 - A simple practical holster should be provided for the loading operator to carry the tablet when not using it.

3 - It should be formally stated which parties are involved in testing the accuracy of the AI text recognition, e.g. to what extent the loading operator is involved in this.

4 - It should be formally stated by the technical partners, whether the loading operators will have to correct AI text recognition errors. (It was not envisioned in the preliminary Use Case presentation to have the loading operator press more than a ‘submit’ button.)

5 – If the loading operators will have to also correct AI text recognitions, the technical partner should work with the loading operators to design an appropriate input system in the interface for such more complex inputs, and test it in the actual work conditions, e.g. the keys/buttons must be designed to be practical for gloves, large enough, operable in dusty conditions, and clearly visible under the work area lighting.

6 - Develop 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.

7 - Consider explicitly whether the time delays possibly introduced by the AI service (errors, etc.) will nullify the time gains made in sending the bag label information directly to the control system. If they do, then consider whether the actual integration of the AI service is practically worthwhile, or whether a non-AI change could resolve the problem, e.g. the loading operator takes a simpler digital picture of the bag labels and feeder number and sends them to the console operator to check as a timely safeguard against bag misplacement (but also brings the label to the control room to be inputted manually just as before)

8 - Formally make clear to the console operator how their job changes relative to the new AI-integrated context, e.g. will the console operator now review in any way the name and label of the product bag which has been inputted automatically into the control system?

9 – We recommend that the layer of safeguard of the console operator is not removed (or do not remove it until the AI has been through a long trial period), e.g. when the recognized text for each bag and

feeder is sent, send also to the console operator a regular picture taken simultaneously during text recognition scan, so that the console operator inputs it as before – but from the picture instead of the physical label – and set the control system to wait for both inputs (recognized text and console operator manual input from the digital picture) before giving the green light.

In other words: use the AI eventually, after a trial period, as a secondary safeguard to address the central issue of the loading operator time delay in carrying the label to the control room, rather than using the loading operator as a safeguard against AI error.

10 - Clarify formally who is getting the green light to release the product bag into the manufacturing process. If both console and loading operator are getting the green light and are able to release the bag product, then formally clarify which of them has the ultimate responsibility for the action.

11- Take into consideration whether the change of physical activity in not going to the control room affects operator wellbeing. Ask the operator team.

12 - Monitor regularly whether the cohesion of the operator team changes for the worse due to the change in social interaction of the team members. Ask the operator team members.

4.2. Use Case Example #2: Preventative Material Preparation Guidance from AI at the beginning of a Production Line

Description

Our second Use Case example is centered around material at the beginning of a production line being fed into a hopper for further processing. The materials are brought on a pallet to a feeding conveyor, where they are then moved along to a hopper to be mixed. The feeding process at the entrance of the hopper is monitored by an operator in order to prevent jams in the hopper and consistency of the subsequent product.

The incoming material is normally a continuous piece of material, but occasionally there are missing sections and also ends which have to be joined together when a new pallet of material is introduced, which causes sections with surplus material.

The monitoring operator must make quick adjustments, e.g. joining ends of material, or building up extra material in advance, in order to prevent the hopper from jamming or leaving gaps in the material which influence later parts of the processing of the product.

The operator works manually with a knife, cutting and joining or adding, as necessary, sometimes switching the machinery to manual mode or adjusting the rpms of some components of the machinery as necessary to do this.

The goal is to integrate the AI to give guidance to the operator (or to take action directly) in order to ensure the consistency of the incoming material, e.g. to tell the operator where and when to add material on the conveyor. This needs some vision sensors directed at the material on the incoming conveyor or some other form of measurement.

Ethical Issues

1 - The heart of this Use Case is with time management and a predictive guidance by the AI to make up for a lack of time for the operator to react. The question then becomes whether the AI could in theory solve the time problem.

The reaction time is very short. The question regarding how much time the operator has to make adjustments was asked. The answer was 30-40s, sometimes as little as 10s. The ensuing discussion around whether it seemed a very short time for the operator to get guidance from the AI and use it to react, given the lack of time, indicated that the hope was for the AI to be configured to make the adjustments automatically somehow. This is very vague in terms of establishing control and responsibility in the context.

Operators can be late or not fast enough to compensate for parameters which must be kept constant at the entrance to the hopper, so time is already an issue here, and it will be more an issue if the operator must add consulting the AI in that short time. The added stress on the operator is an ethical issue here, but it comes from a practical issue related to the value of the overall system.

There is an ambiguity here about the source of the problem: the problem may be that: #1 the operator does not have enough time now to react because the adjustments are very physical adjustments often (adding or reducing material to buffer the input to the hopper), or it may be #2 the operator does not know when to make adjustments, or #3 the operator does not know how much to adjust.

If it is only or significantly #1 then AI is not the solution to the problem. If it is #2 and/or #3 then AI may be able to solve the problem, but only if the feeding procedure can be clarified and standardized better (e.g. if the AI proposes “add more” material it will have to tell the operator how much to add, which needs some standard measure of weight; and if it proposes ‘add material now’ then ‘now’ cannot be instantaneous for the operator, so the feeding conveyor will have to be segmented somehow to convert ‘time’ into ‘place’ on the conveyor.) And if the problem is a mixture of #1, #2, and #3, then problem #1 still has to be solved first.

It is necessary to figure out where the problem is with some trials, perhaps some modeling, which will probably go beyond the timeline of this project.

2 - The selected key performance indicator (KPI) here is the quality of the product issuing from the larger part of the production line in which the feeding conveyor/hopper is the beginning, but that quality is measured far away from the action of this Use Case, namely at the end of the process. There are too many variables influencing the chosen KPI for it to be used as a success benchmark for this Use Case. With this KPI it will be difficult or impossible to separate how much the operator in question is influencing the solution, or trace the chain of action (an issue of traceability and responsibility).

3 - There is an ambiguity at the level of the stated goal, relative to whether the operator is truly in command or whether the AI will make adjustments automatically. If the UC is chosen this needs to be resolved, e.g. by giving a definite trial period for operator in command only, and another for AI auto adjusting.

Recommendations

1 - Measure KPI relative to average frequency of hopper jams or production scrapping to get a minimum success rate to measure whether integrating the AI is worthwhile here.

2 - It was suggested that this Use Case is central to the production line because it is the beginning, so that an improvement here will benefit overall quality of the product.

If so, then perhaps it is worth considering adding a ‘geared’ or looping feed conveyor (an engineering problem) to extend the time in which the material travels between the pallet and the hopper in order to give the operator more time in the feeding process to consult the AI and react to its proposals. Perhaps it is also worth considering addressing this problem in the stage immediately prior to the Use Case context, in the area where the material is initially prepared. Perhaps the AI could be better used there somehow.

3 We recommend that you delay this Use Case until you understand the problem better and the operator’s role in it

5. Identification of Legal Issues

5.1. Main Legal Instruments

Aspects of the legal framework include data protection, cybersecurity, safety, liability, and accessibility.

A large part of the focus of legal efforts addressing AI development in the EU will be on updating, adjusting, and clarifying existing laws relative to AI characteristics, transparency in particular, because opaqueness of AI services do not allow breaches of law relative to existing laws, to be noticed. (EU Commission Whitepaper on AI, 14)

https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

We will carry out further research to try to understand the *default* legal position on data created by employees or generated by employees in cases where data ownership is not specified in a work contract, i.e. does such data generated during working hours relative to work tasks, remain the property of the employer by default? The answer to this question will have a direct bearing on all of the Use Cases in the project, with regard to GDPR regulations, etc.

Some insights from the Rendez-vous IEEE Chapitre 28 Français SMC: Éthique & IA (Professeur Linda Arcelin - Faculté De Droit, Université de la Rochelle), March 28, 2021

There is no univocal response from jurists regarding the legal issues around AI yet. Technology is moving faster than legislators can respond.

So far *Responsibility* is the first point of contact for jurists and AI. But there is a difficulty. Only 2 types of persons are recognized in law: moral and physical persons, so it is difficult for jurists to categorize AI.

Responsibility falls under three categories for French law: penal, civil, and administrative. So far the concentration of jurists is upon civil responsibilities, under the categories of personal fault and defective products.

In 2021 the EU will create criteria for autonomy relative to AI. AI will be classified as High Risk vs Other AI.For High risk AI defective product rules will apply. For Other AI personal fault rules will apply.

Under Administrative Responsibility laws of competition will apply, e.g. when algorithms decide market prices, if deep learning causes prices to come together contrary to market regulations. There may be a category of ‘facilitateur d’entent,’ e.g. a programmer who facilitates the breaking of competition laws by a company through an algorithm. Such a programmer would be punished under the law.

Laws relevant to industrial AI integration include the following.

GDPR regulations

OECD Recommendation of the Council on Artificial Intelligence (2019)

https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

P9_TA-PROV(2021)0009

Artificial intelligence: questions of interpretation and application of international law (EU parliament resolution of January 2021)

https://www.europarl.europa.eu/doceo/document/TA-9-2021-0009_EN.pdf

NIS Directive

https://digital-strategy.ec.europa.eu/en/policies/nis-directive

(other regulations to be added)

5.2. Relevant International Standards

We are planning a meeting with members of INRS to discuss relevant international standards for the AI development of AI-proficient.

Meanwhile our research has suggested that the development of AI standards is in very early stages. Nonetheless, the following International standards under development, are potentially relevant to the development of AI services in the AI-Proficient project.

IEEE standards (in various stages of development):

(IEEE) The Ethics Certification Program for Autonomous and Intelligent Systems
It focuses on development of metrics to certify AI related products as 'trusted' with regard to Transparency, Accountability, and Algorithmic Bias
https://standards.ieee.org/industry-connections/ecpais.html
(in development)

P2894 - Guide for an Architectural Framework for Explainable Artificial Intelligence (working group stage)
https://standards.ieee.org/project/2894.html

IEEE P7000 - IEEE Draft Model Process for Addressing Ethical Concerns During System Design (draft stage)
https://standards.ieee.org/project/7000.html#Standard

IEEE 7010-2020™ - IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-being (published)
https://standards.ieee.org/content/ieee-standards/en/standard/7010-2020.html

ISO standards (all relevant standards are still in preparatory/pre-draft stage):

ISO/IEC AWI TR 5469
Artificial intelligence — Functional safety and AI systems

ISO/IEC AWI TS 6254
Information technology — Artificial intelligence — Objectives and methods for explainability of ML models and AI systems

ISO/IEC AWI 25059
Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI-based systems

https://www.iso.org/committee/6794475/x/catalogue/p/0/u/1/w/0/d/0

(possibly further standards to be added here for the final draft)

5.3. Discussion of potential legal issues

(material to be added after further discussion and consultation)

6. Conclusion

6.1. Preliminary Results

A preliminary ethics summary, based on a template developed by WP1 leaders, was sent to the industrial and tech partners, rating each Use Case in outline format, with regard to potential ethical difficulties.

The two Use Case examples in question were rated as having high ethical difficulties (on a simple scale of low, medium, high).

For Use Case example #2, recommendations and the reasoning (i.e. ethical issues) behind them were also sent to the responsible persons for the respective industrial partner. One of the recommendations (see above) was that Use Case example #2 be delayed until the key issue to be addressed by the Use Case was clarified.

The tech and industrial partner leaders then discussed the Use Cases, together and internally, in order to make a selection from among the Use Cases offered.

As a result of this discussion, Use Case #2 was not chosen to develop in the project (among others), despite it being initially designated as one of the most important by the respective industrial partner. The serious consideration of our recommendations by both industrial and tech partner leaders was much appreciated. Going forward we hope to get informal feedback on how our ethical reasoning and recommendations are used in other development decisions within the project. This will help us improve out working method.

We have also observed that the *tech partners* do indirectly have a feel for the ethical issues*, even before seeing our ethical recommendations*. They have a sense of the difficulties involved in the human aspect when, e.g. the role of the operator in the Use Case is not well clarified. If they have a choice, they tend to prefer Use Cases where the role of the humans is – or seems – clearer, or where human involvement is minimal (the exception is where the tech partner is specifically contributing unavoidably human related technologies, such as human-machine interfaces).

This observation was based on the fact that *before* the ethics team had submitted our own summary or recommendations to be considered, the Use Cases in which most tech partners indicated the least interest in their outline submissions of interest, were those that the ethics team subsequently highlighted as having the highest ethical difficulties.

If this informal observation bears out, then in one sense this is not a good thing because it is a negative reaction, despite being a quite natural one, i.e. extensive human involvement in a work context makes it more complex, potentially dangerous, and difficult to address easily and efficiently from a technical perspective.

In another sense it is promising insofar as it leaves an opening for better ethical discussion with the tech partners on the basis of a sense of the ethical issues at stake being a-priori recognized as essentially human centred. If we can bring into the discussion the question: “how does the human aspect in this context make a technical development and integration of the AI service more difficult for you as tech partner to carry out?” then we can go on to suggest ways to address those difficulties, rather than simply pushing on with deliberate blindness to the fact that in fact humans are *always* involved and that technological developments are always human creations first.

6.2. Human-machine interaction, AI-Proficient Strategy for Going Forward

The Covid-19 pandemic has had some negative impact upon the efforts of the ethics team, e.g. a number of expected in-situ visits to experience the physical work environment and current human-machine interfaces of the operators in the Use Cases have not been possible for the most part. The restrictions of lockdowns and general misunderstandings caused by lack of in person meetings also makes ethics discussions somewhat more difficult. We have done our best under the circumstances.

The AI-Proficient project is an ongoing process with different partners engaging in different parts of the process. The engagement of the UL ethics teams is no different. The ethics contribution is an ongoing, evolving, and dynamic process of which the Deliverable 1.2 can only be a momentary ‘snapshot,’ and this should be remembered.

Our strategy going forward will be to continue applying our own understanding of Ethics by Design, as a continual and evolving process of discussion, questioning, reflection, recommendations to guide development, and assessment of results.

Detailed written ethical recommendations were made – or are in process of being made – to the manufacturing partners, as private documents, for every Use Case. The latter could not be presented to the same level of detail in Deliverable 1.2, which is a public deliverable, due to legal and privacy constraints. As choices are made regarding the development of the desired AI integrations, we will continue to make further private recommendations, or modify those already made, both to the manufacturing partners and to the tech partners.

The focus will turn much more toward the design choices of the tech partners as they develop the AI services needed for each Use Case. We intend to initiate special Q&A and discussion sessions with those partners, to attend any design related meetings that are open to us, to make recommendations to them regarding human friendly interface choices, explainability, transparency, etc.

Further research in the current scientific literature on the more programming oriented ethical aspects of the project’s AI development are envisioned.

Another part of our strategy going forward will be to try to better establish the wishes of all parties, which includes in particular understanding the viewpoint of the employees at the ‘shop-floor level.’ The external input from members of INRS mentioned above have given us insights into how to approach these employees in order to address their concerns, e.g. in questionnaire design, which the timeline of WP 1.2 did not allow for.

Acknowledgements

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

Bibliography

Anderson, Marc. (2019). Hyperthematics: The Logic of Value. New York. SUNY Press.

Birhane, Abeba. (2021). Algorithmic injustice: a relational ethics approach. Patterns. 2. 100205. 10.1016/j.patter.2021.100205.

Demir, Kadir & Doven, Gozde & Sezen, Bulent. (2019). Industry 5.0 and Human-Robot Co-working. Procedia Computer Science. 158. 688-695. 10.1016/j.procs.2019.09.104.

Gamer, Thomas & Hoernicke, Mario & Klöpper, Benjamin & Bauer, Reinhard & Isaksson, Alf. (2019). The Autonomous Industrial Plant -Future of Process Engineering, Operations and Maintenance. IFAC-PapersOnLine. 52. 454-460. 10.1016/j.ifacol.2019.06.104.

Gazzaneo, Lucia & Padovano, Antonio & Umbrello, Steven. (2020). Designing Smart Operator 4.0 for Human Values: A Value Sensitive Design Approach. Procedia Manufacturing. 42. 10.1016/j.promfg.2020.02.073.

Demir, Kadir & Doven, Gozde & Sezen, Bulent. (2019). Industry 5.0 and Human-Robot Co-working. Procedia Computer Science. 158. 688-695. 10.1016/j.procs.2019.09.104.

Jasanoff, Sheila. (2021). Cosmopolitan Ethics and the Challenges of Global Governance of AI. UNESCO Third Roundtable on Ethics of AI: Shaping the Future of AI through Cultural Diversity. March 26th. Paris https://www.youtube.com/watch?v=Rdp6hQXVpqM

Kinzel, Holger. (2016). Industry 4.0 – Where does this leave the Human Factor?. Journal of Urban Culture Research. 15.

Lefeuvre-Halftermeyer, Anaïs, & Govaere, Virginie, & Antoine, Jean-Yves & Allegre, Willy, & Pouplin, Samuel, & et al.. (2016). Typologie des risques pour une analyse éthique de l'impact des technologies du TAL. *Traitement Automatique des Langues*. ATALA. TAL et éthique. 57 (2). 47-71.

McNamara, Andrew & Smith, Justin & Murphy-Hill, Emerson. (2018). Does ACM’s code of ethics change ethical decision making in software development?. 729-733. 10.1145/3236024.3264833.

Mittelstadt, Brent. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence. 1. 10.1038/s42256-019-0114-4.

Morley, Jessica & Floridi, Luciano & Kinsey, Libby & Elhalal, Anat. (2019). From What to How: An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices.

Neumann, W. Patrick & Winkelhaus, Sven & Grosse, Eric H. & Glock, Christoph H. (2021) Industry 4.0 and the human factor – A systems framework and analysis methodology for successful development. International Journal of Production Economics. 233. 10.1016/j.ijpe.2020.107992.

Ocone, Raffaella. (2020). Ethics in Engineering and the Role of Responsible Technology. Energy and AI. 2. 100019. 10.1016/j.egyai.2020.100019.

Pégny, Maël. (2021) Pour un développement des IAs respectueux de la vie privée dès la conception. hal-03104692

Tay, Shu & Te Chuan, Lee & Aziati, A. & Ahmad, Ahmad Nur Aizat. (2018). An Overview of Industry 4.0: Definition, Components, and Government Initiatives. Journal of Advanced Research in Dynamical and Control Systems. 10. 14.

Trentesaux, Damien & Rault, Raphaël. (2017). Designing Ethical Cyber-Physical Industrial Systems. IFAC-PapersOnLine. 50. 14934-14939. 10.1016/j.ifacol.2017.08.2543.

Trentesaux, Damien & Karnouskos, Stamatis (2021). Engineering ethical behaviors in autonomous industrial cyber-physical human systems. Cognitionm Technology, and *Work*. 10.1007/s10111-020-00657-6

Wioland, L. & Debay, L. & Atain-Kouadio, J. (2019). Processus d'acceptabilité et d'acceptation des exosquelettes : évaluation par questionnaire. *Références en santé au travail*, Institut national de recherche et de sécurité pour la prévention des accidents du travail et des maladies professionnelles. 160. 49-76.