From the Ground Up: Developing a Practical Ethical Methodology for Integrating AI into Industry

Marc M Anderson, Karën Fort

From the Ground Up: Developing a Practical Ethical Methodology for Integrating AI into Industry

Marc M Anderson[1] Karën Fort[2]

Abstract

In this article we present a new approach to practical artificial intelligence (AI) ethics in heavy industry, which was developed in the context of an EU Horizons 2020 multi partner project. We begin with a review of the concept of Industry 4.0, discussing the limitations of the concept, and of iterative categorization of heavy industry generally, for a practical human centered ethical approach. We then proceed to an overview of actual and potential AI ethics approaches to heavy industry, suggesting that current approaches with their emphasis on broad high-level principles are not well suited to AI ethics for industry. From there we outline our own approach in two sections. The first suggests tailoring ethics to the time and space situation of the shop floor level worker from the ground up, including giving specific and evolving ethical recommendations. The second describes the ethicist's role as an ethical supervisor immersed in the development process and interpreting between industrial and technological (tech) development partners. In presenting our approach we draw heavily on our own experiences in applying the method in the Use Cases of our project, as examples of what can be done.

Keywords

artificial intelligence, tailored ethics, industry 4.0, workplace, ethical supervisor

Introduction

The pace of AI integration in heavy industry is accelerating, and a number of authors, e.g. (Pacaux-Lemoine and Trentesaux, 2019), have recently drawn attention to a variety of ethical risks associated with such integrations. Most approaches to AI ethics - including industrial AI ethics - are general. The current trend is to work at the level of broad ethical principles, generating sets of principles from expert working groups (Jobin et al., 2019), such as the Ethics Guidelines for Trustworthy AI (EU High-Level Expert Group on Artificial Intelligence, 2019) – HLEG –, or collating such principles together under more general categories (Zhou et al., 2020). The application of these principles is typically developed into 'tools' or 'frameworks,' which are offered not as a series of definite suggestions, but merely as efforts to promote reflection in designers and developers.

In this paper we will argue for a tailored and bottom up approach to ethics in industrial AI, particularly heavy industry, anchored in the terrain of actual work realities. Whether this approach is useful for other areas of AI ethics we leave open, but have little doubt that it would be useful and can be adapted to other areas. The

- [1] marc.anderson@inria.fr

  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

foundation of our approach comes from our ongoing participation in an EU Horizons 2020 project, which brings together multiple technology development partners, several corporations in heavy industry, and one university, in exploring the real-world application of AI systems to improve manufacturing lines and industrial control systems in a number of manufacturing plants located in three countries. This paper presents our method and some of our results when engaging a mix of different developing Use Cases.

We will begin with a review of the current state of ethics related to industrial applications of AI. This includes an attempt at defining the elusive concept of Industry 4.0, as well as some of the particular difficulties in categorizing industry prior to ethical engagement and in engaging it by means of current approaches in ethics. From there we will go on to offer two main arguments. First, ethics can practically be tailored to the work context of the worker, in time, space, and other realities such as the official – and unstated – hierarchical relations of the workplace. Second, the ethicist can act at ground level, in an intensive way, as an *ethical supervisor*, directly and continually engaging and guiding the developing AI integration while watching out for workplace realities which would otherwise remain hidden if the situation were only approached through high level principles.

Current Ethical Approaches to Industry 4.0

The Difficulty of Defining Industry 4.0

Retrospectively, eras of industry have been defined and categorized down to the beginning of the industrial age. Industry 1.0 is taken to be the era of mechanical production combined with water and steam power; industry 2.0 is characterized as industry involving division of labor, mass production and electrical energy; industry 3.0 meanwhile is defined as industry integrating electronics, IT systems, and automated production. (Demir, et al., 2019). Industry 4.0 however is open to multiple definitions (Tay et al., 2018), and is very broad: “Industry 4.0, also called digital manufacturing or Industrial Internet of Things, appears to be an umbrella term that encompasses all digital technologies that are transforming the production processes worldwide … based on the Internet, networks, and Cyber-Physical Systems.” (Satyro et al., 2022)

Besides covering multiple technologies of which AI is only one, Industry 4.0 is a very *future oriented* concept, characterized by an “aspirational nature” (Neumann et al. 2020). Speculations – already, before Industry 4.0 is even well advanced – about Industry 5.0 which is envisioned as a post-fossil fuel, and biological based industry (Demir et al., 2019), indicate just how future oriented the industrial iteration classification is.

Industry 4.0 is not so much something to be defined then, as it is *to be achieved*. Looking through the lens of this aspirational view, the proponents of Industry 4.0 sometimes appear to forget that a great deal of heavy industry retains a very physical and mechanical character. Viewed broadly, industry as such is not a solid block (completely homogeneous) – and hence the approach which we outline in this article emphasizes the ethical engagement of *heavy industry* as its context. You *might* be able to select a country, a region, a type of production or manufacturing, a particular factory, a (work section/group), etc. and be able to say – tenuously – that it is now in the Industry 4.0 stage. More likely is that any reasonably broad selection you make is more aptly characterized as Industry 3.0 mostly, with some Industry 4.0 elements, and remnants of Industry 2.0. And, remembering the tendrils which reach out to make up a globalized industry, in some parts of the world Industry 2.0 might easily still be the norm. The conditions of our own project - which may be 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 or 2 (perhaps in some aspects 3) 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. To give an example from our own *in-situ* visits to our partner manufacturing plants, in the midst of high-tech processes which rate as Industry 3.0 or Industry 4.0 we found shop floor operators regularly using simple knives to cut and shape products, or carrying leftover material on their shoulders. The named iterations of Industry 3.0, 4.0, and so forth, are focused upon the ‘machinery’ additions to human labor of course, but we think it would be a mistake to separate the human process from these mechanical aspects. At least it would be a mistake in an ethical sense if we want human centered design.

The argument can also be made based on the outcome of the industrial process however. Does it make sense to label a context as Industry 4.0 if it uses parts sourced from an Industry 3.0 or even industry 2.0 context, or if the lifecycle of its products or product components, pass – perhaps repeatedly – through an industry 2.0 stage? Take for example the rush to carry the Industry 4.0 label to the maritime shipping industry (Stanic et al., 2018) even though the heavy industry of building ships is largely Industry 2.0 based, and the dismantling of large ships – whose components will be resold and re-used – in countries such as India and Bangladesh, is quite frankly at an Industry 1.0 or lower stage. This problem of non-homogeneity becomes even more evident when the human aspect of industry is considered. The human issues, e.g. hierarchical relations, or the ability of the human body – which is not a 'body 4.0,' 'body 3.0,' etc. in its physical form – to adapt to human-machine interfaces, are most often those which are stuck at a so called 'earlier' stage.

Thus, the very *4.0* itself of the term *Industry 4.0* is problematic. Labeling by versions was developed in the early 2000s based on the software industry practice of labelling software versions. Versions of software have led to versions of products. Ethics in industry is not merely ethics in products however. Such labeling loses the sense of building upon the history of the development of industry in two ways. On the one hand, it can lead to viewing industry as having definitely *passed* certain stages, a doubtful position. On the other hand, the labelling can lead to viewing that history as incremental improvement on a finer scale and characterizes the new stage of industry where the problems of an earlier stage are cast off *homogeneously*. Industry is not homogeneous in its advancement, as our own project experience has shown us. It is a process, which depends upon the human within it as a process, an element which renders iterative labelling problematic.

If the definition of Industry 4.0 is still vague there is a real danger of it amounting to: the industry of 'whatever is the newest thing,' an outlook which, if not compensated for, is not promising for ethical engagement.

Is there an Industry 4.0 AI Ethics?

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 – can be mentioned in the first place. That research deals with the ethics of cyber-physical or autonomous-cyber physical human systems. The latter article, 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. This research is extremely theoretical however, dealing with quite advanced systems 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. We think that greater emphasis on the ground level industrial context is needed however, where the shop floor worker is very much surrounded by problematic issues of tasks to be done in limited times and in limited spaces with elements leftover from older modes of human work. These 'older' industrial conditions addressed in the Use Cases of our project, zones could be classed as 0 and 1 on the autonomy scale noted above, even though 'nested' within a larger context of highly autonomous machinery. They call for a rethought ethical approach which can complement the type of research just mentioned.

Can we fall back on Industrial Ethics or more general AI Ethics?

If, as we suggest, there is such a lack of direct industrial AI ethics related research, which engages the non-homogeneity of heavy industry, then 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 practicality.

Of research which recognizes human centrality in considering Industry 4.0, the most thorough by far, is (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 advance 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: 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. The latter suggestion runs close to our own view of the ethicist's role.

If we consider AI ethics in general, it covers a broad field, and focuses on many issues which are sometimes not directly relevant to the industrial context, e.g. social cohesion, diversity in the field of AI, public awareness of

AI, and field specific deliberations (health, military, etc.), among others (Hagendorff, 2020). In AI ethics, a supposedly objective and neutral view, 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). This in turn has focused the issues for general AI ethics away from what would be practically relevant to the industrial context, by evolving AI ethics in the direction of broad principles, which tend to be addressed to "multiple stakeholder groups" (Jobin, et al., 2019).

If, arguably, a mechanistic philosophical worldview generated in 17th century philosophical thinking has given rise to the very machines – our computers – which enhance the ethical problems at issue, then on the other hand it is not surprising that mechanistic ethical constructions – the multitudes of 'universal' ethical principles, guidelines and checklists – 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 contradiction, practically speaking, to the actual assumptions driving much of AI development. Suggesting that machine learning (ML) models be made publicly available when possible in order to test their security properties, (Pégny, 2021) is a good idea and yet it runs up against the practical constraint that some of the most used ML models will be developed under proprietary conditions. Such limitations echo others which we have experienced in our own project within industry and will be discussed further on.

Going beyond Principles, Guidelines … and yet more Principles

Industrial and general AI ethics in their current state are of little help then, because they focus too much upon general principles. Virginia Dignum has recently highlighted the fact that work on principles and guidelines is extremely abstract and extremely high level, and that even though principles and guidelines are needed, "we are not moving to action just by endorsing principles," (Dignum, April 20, 2021). Our suggestion is that this is particularly the case for AI ethics in heavy industry. Principles 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 industrial AI ethics in particular (and perhaps also AI ethics in general), is twofold: first, to leave aside formulations of principles unless they are immediately practical ground level principles (examples of which will be given in what follows), and second (Morley et al., 2020) to move from principles to practical implementations. What we suggest is that, largely: there is no truly practical AI ethics yet which can be called upon for industry generally or for Industry 4.0. It remains to be created. It can gain balance by taking into account the humans involved, 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 would be helpful to pause to consider the human worker's point of view and wishes (Wioland et al., 2019). Going still further it can re-discover the humanity in what has hitherto been regarded as a pure 'realm of data' in which software developers play; software developers are human, and software, algorithms, etc. are human creations. We think an AI ethics for industry can become practical, which includes addressing the changing human process and not simply the rationalized and categorized 'human as object,' the so-called 'human resource.' Practicality can be drawn out of the industrial context itself.

Tailoring Ethics to the Real Situation of the Worker: Time, Space, and other Workplace Realities

A tailored approach is a matter of making specific recommendations and thus building a good foundation for low level practical principles – often physical related but not always – from the specific situations and recognizing after considering those situations which higher level principles may not apply, or apply only weakly. This is not therefore an abandonment of principles. It means not reaching for broad ethical principles until the situation on the ground in which the principles are to be applied is expansively understood and engaged by the ethicist by developing a baseline of what is happening initially. It also means adopting principles of an appropriate scope and being a pathfinder in the matter of principles. Sometimes, since the application of principles – including which principles to apply – is not immediately evident, this will mean *developing low-level principles directly from a selected and examined context*, low-level principles which are needed in order to subsequently develop or apply broader principles.

We began our effort to tailor ethical engagement in our project to the actual situations of the workers by developing a list of questions which we would gradually answer in order to fill in a picture of the unique baseline situation for each Use Case, i.e. the situation as it was before any project modifications are applied. An anonymized sample of the questions we asked as our Baseline Questions is included in Online Resource 1.[3] From these we developed as detailed a picture as possible of the worker's physical environment, the tools being used, the tasks carried out by the worker and the time required for those tasks, the human work relations, etc. This process of questioning was ongoing and proportional to our access to discussion sessions with the plant engineers, and discussion with and observation of the workers.

From there we went on to consider what was envisioned in integrating AI in each Use Case. In some cases, the plan was vague, in others more definite. Here the expectations of the manufacturer and engineers in terms of KPIs or 'key performance indicators', i.e. *indicators of quality improvement in some part of the process*, were considered, along with the first technical suggestions from discussions with tech partners as to how AI would be used to solve the problems of the various Use Cases. Having now an idea of the baseline situation as well as the proposed changes, we could then consider the changes carefully and dig out any preliminary ethical issues. We then drafted a first set of very specific ethical recommendations for each Use Case to be used by the tech and industrial partners if feasible. Online Resource 2[4] presents an anonymized sample of recommendations for one Use Case.

In our opinion this approach worked well in uncovering ethical issues in the real situation of the worker, where some issues can be spotted very quickly. For example, in one of the initially proposed Use Cases, the request was to have the AI give suggestions to an operator who was working to feed a continuous ribbon of material from a pallet onto a conveyor, leading to a hopper where it was to be mixed. The operator had to oversee several feeding stations, among other duties. The AI was to be connected to a camera to scan the ribbon of material and warn of twists or inconsistencies in advance and alert the operator. In initial discussions we asked the time and space parameters of this process: "how long does the operator have to make adjustments?" and "where does he have to move to between feeders?" The answer was he had about 20 seconds, to make an adjustment, and had to move back and forth between 4 feeding stations. Our recommendation was that it would be very problematic to introduce another task – checking the AI suggestions – into this context without overwhelming the operator, unless the feed could be slowed to give the operator more time. After discussion between the tech and industrial partners around this point, the Use Case was dropped completely as part of the Use Cases to go ahead with. Later in a plant visit we saw that our recommendation made on the basis of a conceptual analysis of the situation was proven correct when viewed in the factory floor context. We saw that the operator in question was in constant motion between feeders, forklift, and control panel, and moved up and down nearby stairs to check other machine areas. He might easily have been swamped by the additional task of checking the AI, beyond any benefit that the predictive solution of the AI could have brought. Sometimes *not* implementing a technology is the best way out of a related ethical dilemma.

The baseline questions themselves need not be set in stone, but rather drafted according to an initial appraisal of the working environment. Going forward they can be flexible, adaptable, and disposable as well if they have no evident bearing on the situation. They and the next stage of clarifying what is planned in the AI integration can then be used to draw out the ethical problems in a situation, rather than rigidly leaned on as a crutch - even at a very specific level - for assuming that certain ethical problems can be found, generically, in every situation.[5] The latter is not the case in our experience, and we give examples below of this and of the kinds of ground level ethical issues which we uncovered using this method.

Example 1: initial KPI does not consider Physical Time and Space of the Worker's Environment in relation to the AI integration

An issue in one of the Use Cases was that the chosen KPI, here the quality of the product at the *end* of that section of the line, is very far – both literally and figuratively – from the main context of the Use Case, i.e. where the human interaction with AI and machinery would take place. In a production line, one aspect of the

[3] https://gitlab.inria.fr/kfort/ICTProject/-/blob/main/DataForEthicsInIndustry4Paper/ESM_1.pdf

[4] https://gitlab.inria.fr/kfort/ICTProject/-/blob/main/DataForEthicsInIndustry4Paper/ESM_2.pdf

[5] One of the problems with focusing on high level charters, frameworks, etc.

production process usually follows another in steps. In this case the distance between the proposed KPI and the action meant that the success of the AI integration could not easily be measured since many variables – variables being changed by AI integration in other parts of the line – would be in play at once. There was thus no clear way of knowing whether the AI integration was worthwhile (or simply needlessly increasing complexity which would stress the operator), or how much the operator's actions when coupled with the AI guidance would be traceable relative to the KPI, in terms of separation of who did what, or who made mistakes (either human or AI).

The respective recommendation was to change the KPI, i.e. to *choose some suitable KPI, among several available options which were nearer the human operator context*. This recommendation can then be developed into a higher level but still *specific and practical principle* to be applied to *production line manufacturing contexts if useful*: ***in order to practically measure the success of an AI service integration for a particular industrial context, set the KPI for that context as near as practically possible to the main area of human activity in that context***. This, then, illustrates the tailored ethical approach of moving from the particular situation of the worker to a higher-level principle.

Example 2: AI integration in real work situations is driven by Specific Goals and sometimes unstated interests which when discovered and understood can help in making the best practical use of Ethical Guidance

In various question and answer sessions, and technical meetings of the project, which were called in order to get an overview of the Use Case contexts, it became clear that not all Use Cases were of equal importance to the industrial partners. For some Use Cases the notion of 'quick wins' in terms of efficiency and quality were mentioned. Notwithstanding the fact that the notion of quick and easy alterations of an industrial process is ethically problematic in itself – since the time for considering consequences, trials, etc. is reduced – there is a larger relevance for a tailored ethical approach.

First, - and this ties in with the notion of continuous guidance below – the interests of industrial partners do not become evident without the bottom up approach. It takes sustained discussion to understand the relative interest which managers and workers have in each of the options open to them (here in our project, for example, in terms of which Use Cases to choose to develop). This is because some interests cannot be expressed in initial public planning stages – or they are deliberately omitted for privacy reasons – but also because the interests of the tech and industrial partners evolve as choices are made. The attentive ethicist can often 'pick up' on these unstated points which often have considerable ethical consequences and which come out naturally in the give and take of more private internal discussions. The ethicist can then sometimes address such ethical issues in a way which steers a course between the private interests of the partners and a more ethical implementation of proposed AI integrations.

Second, and perhaps more importantly an understanding of relative interests derived from the real work situation allows us to tailor our ethical response in terms of where the greater effort, time, and resources need to be put by the ethicist. Efforts can be tailored accordingly, since you are not likely to get recommendations adopted if you cannot be flexible. A context where 'quick wins' are desired, may need one approach, but larger more long-term quality issues, or problems which are very exploratory, may need a very different approach.

An example of the above occurred in one of our in-person visits to one of the manufacturing plants of our industrial partners, where we had opportunities to talk in person to process engineers inside the manufacturing plant – rather than by video conference – and then later to talk further both formally and informally. In these discussions we heard things like "we heard a lot about AI in other industries [other than heavy industry] and we wondered if we should try it to improve some operations of the plant." From this we noted that the introduction of AI had an exploratory aspect for them. We also heard things like "AI seems like a magic box to us that can solve anything … but we're a little bit anxious about using AI to solve our *x* problem." When we asked why, the answer was "well we've tried to solve that problem a dozen times, so if the AI can fix it, we'll feel a bit stupid … so in one way we'd like to have the problem solved … but in another way we'd like to see the AI not be able to fix it." Needless to say, this type of ambiguous thinking was not an ideal starting point to achieve an ethical outcome. Accordingly, we made several recommendations including: to explicitly recognize the existence of this tension of interests in the approach to the Use Case, to formally clarify what was going to be attempted, and for the process engineers themselves (and the operators) to engage with the developers to develop the AI solution explicitly as a human centered process using the AI as a tool so as to avoid a 'magic box' attitude, i.e. take the attitude that *we the humans are fixing the problem x… with the help of the AI*.

Example 3: Tailoring Explainability in AI integrations to the Baseline Working Situation

"Explainability concerns the ability to explain both the technical processes of an AI system and the related human decisions (e.g. application areas of a system)" (HLEG, 2019). As a sub-category of the higher-level principle of Transparency, the suggestion to strive for explainability is good, but needs *building up* and *selective application* according to the needs of the ground level situation. One of the Use Cases in our project illustrates this. That Use Case proposes to integrate the AI in order to stabilize an industrial reaction in which a number of compounds are combined. Operators at a control board watch over this process and are able to adjust various parameters, but it happens nonetheless that the reaction process degrades toward sub-nominal conditions from time to time. One of the tech partners is tasked with developing explainable AI for the project, including the Use Case in question.

In establishing this Use case baseline in Q&A sessions, we asked, from various angles, what was the current state of understanding of the operators and managers of the reaction process relative to the sub-nominal conditions which are being addressed. The response was that they were not clear at all on why the process degraded. But on the other hand, the respective project tech partner insisted that they would strive for explainable AI in the case and that the operator would be an integral part of training the AI. Clearly there was a dilemma between satisfying the demands of explainability (and AI feedback) and the baseline situation of the knowledge of the operators (and also managers). It was fair to ask how operators who had no clear understanding of the process before AI implementation could meaningfully contribute AI feedback or could gain anything by AI explainability unless some additional steps were to be taken.

Our recommendation was that this be formally categorized as an *Exploratory* implementation of the AI, thus clarifying from the beginning what could be reasonably expected from the operator in terms of feedback, and further that the exploration should proceed in *clearly separated stages* so as to build a foundation for any explainability that could be achieved. It was further recommended, to the same end, that since the reaction process itself had definite stages, the exploration begin by eliminating the simplest and least complex stages first as the cause of the sub-nominal state. These are the types of logical considerations which we suggest help clarify and make possible a practical implementation of what might otherwise be a hopelessly vague principle of Explainability.

One could then go on to generalize the ground level case into a more general, but still practical principle of **Parsimony for Explainable AI**: ***when an AI service is giving guidance on a process issue which is unclear to the human operators involved, then XAI should only be considered as a practical use of resources to the extent that the human operators can work with the developing AI in stages, and in a stepwise fashion, learning/teaching themselves by following the developing AI as it is used to explore the process.***

Floridi's Method of Levels of Abstraction (Floridi, 2008) could be used in such a situation to create a Gradient of Abstraction between various Level of Abstractions (LoAs) chosen to engage the relevant observables of operator aspects of the Use Case and link them to relevant observables of AI output aspects of the Use Case. Thus, in the operator range of the gradient, hybrid LoAs might consist of *Conceptual Elements for Satisfying Explanation* including the types (assert, halt, cause/effect, sustain, increase, decrease, etc.) and *Possible Control Panel Inputs*. While LoAs in the AI range of the gradient might include *Chemical Process Stages*, *Available Explainability Techniques*, and *Individual Explainability Technique Explanation Parameters*. The advantage of Floridi's Method is that in avoiding the ontological approach, and allowing a *pragmatic choice* relative to the need at hand – here in the case of Explainability as an ethical need – it might help avoid the ethical danger of 'forgetting the human as human,' or again 'turning the human into numbers,' i.e. haphazardly translating the human engagement of the system in the direction of the *concrete* (in Floridi's sense of the term) into a complexity which then cannot be reversed to that level of phenomenal *abstraction* (again Floridi's sense of the term) relative to *an understanding of the control of the chemical process* which may be *satisfying* to human explanatory needs at the control panel. Being able to mark the path back to explanation in humanly satisfying terms, a path which the developer could thus not easily overlook because it is marked out, would be a gain. In this particular Use Case it would commit a developer acting in good faith to build up an explanation *with* the operator, which was the intent of our recommendation, while providing a flexible tool to do so.

The Ethicist as an Ethical Supervisor: Direct and Continual involvement in the Development Process

‘Ethics by Committee’ and the role of the Ethicist

In order to explore and understand the real situation of the worker as discussed above, a change has to be made in the way ethicists approach their work. We noted earlier the current trend toward high level principles and guidelines. That trend is arguably in part a byproduct of the tendency of modern European philosophy to strive to universalize but it has meant that contemporary ethics tends to work predominantly ‘in the clouds,’ trying to argue for and support very general principles, leaving others to apply them somehow.

The current focus of many ethicists in artificial intelligence centers on building consensus. This is not unique to ethics in the field of AI. The Centre for Data Ethics and Innovation (CDEI), for example, discusses some of the issues in developing AI assurance.[6] In detailing the various user roles and their needs as separate groups, from government policy makers and regulators, to executives and affected individuals, it stresses this need to build consensus. It becomes clear however that satisfying the needs of the many user groups means taking into account interests that are often deeply at odds with each other. The result so far, where the field of ethics participates in this consensus building approach, has been what we might expect, impossibly broad ‘feel good’ ethical principles. We recognize the reasons for such developments around consensus, in the fact that – as the CDEI shows – AI systems are components of very complex ‘sociotechnical’ systems. We think it is time to complement consensus-based development by turning to the ground level of experience however. Part of the ethicists role may be ‘ethics by committee,’[7] but if the role remains no more than this, if the ethicist does not draw from real world cases and situations, then that role is in danger of melting into a vague hybrid of ‘what we can all agree upon,’ a situation where the ethicist is bound to be lose out because the ground level experience of the other members of the sociotechnical complex *is* being drawn upon. Governments look to votes and economic growth for example, executives seek profit, developers seek clients, and so forth. To match this the ethicist can look to their own equivalent of the sources of experience to nourish their efforts. Thus the ‘ethics by committee’ role of the ethicist needs to be balanced by an ‘ethics by experience,’ whose proper ground level field of action is *the ethical situation of a human individual acting in some particular context*.

In other words, we consider that it is better to help improve experience ethically for one person, or a few, in some definite part of life, than to aim only for some globally applicable principle or principles which everyone may ignore in practice, either because they do not know how to apply them, or because they are so broad as to be evaded practically by those who might not want to apply them, e.g. corporate interests. The improvements for the few will hopefully add up, and form a model for engaging improvements for another few, and so forth...i.e. ethics from the bottom up. The latter is an alternative and poses no danger to the broad principles approach.

A New Role for the Ethicist

To achieve a balance the role of the ethicist needs to evolve. What we are suggesting is that the ethicist become, when possible, a sort of ethical supervisor at ground level. There are two aspects to this supervision. First, engaging the industrial context in which the AI service is to be integrated in detail, including gaining an understanding of the physical environment, the temporal constraints, the particular tasks of the workers in question, the industrial process in question, and the work hierarchy. Secondly, accompanying the AI integration from the planning stages, through development, to and beyond testing and adoption. This latter involves working directly with tech developers and manufacturing engineers as they develop AI services, and gaining as immersive an understanding as possible of the proposed AI service as it develops.

We have already noted the beginning of our own method of achieving the above. We began by using baseline questions flexibly to form as clear a picture as possible of the human work environment and situation for each Use Case. We then compared this with what was intended in the new AI integration, in the ways the AI would presumably affect or change the worker’s tasks, physically, temporally, in relation to the work hierarchy or group cohesion of the workers, etc. We then went on to give a first set of recommendations, based upon our preliminary understanding of how the tech partners in our project would tackle the AI integrations for each Use Case. The recommendations were as specific as possible to be practical, as seen in Online Resource 2.[8] We then followed the development of our recommendations through the pilot demonstration scenarios phase of the project. Here we reviewed the tech partner responses to our recommendations, considering: did they respond fully? or partially? or not at all? did their responses or newly evolved parts of the pilot scenarios raise further ethical issues? We then drafted/evolved new sets of recommendations specific to the pilot demonstration

[6] https://cdei.blog.gov.uk/2021/04/16/user-needs-for-ai-assurance/

[7] The ‘committee’ here can include the group of ethicists themselves in some cases.

[8] https://gitlab.inria.fr/kfort/ICTProject/-/blob/main/DataForEthicsInIndustry4Paper/ESM_2.pdf

scenario phase (see Online Resource 3)[9], incorporating and modifying the old recommendations as needed, or indicating clearly whether we thought the recommendations had been fulfilled, or not. Recognizing our missteps and learning from this process is also very much a part of our approach. If we learn that some of our recommendations are not being taken into account, then we may not have expressed or communicated them well, or they may have been impractical. We have to take responsibility for this and modify them and learn better for future efforts.

Our intention is to carry out this process throughout the whole length of the project, reviewing and considering new phases of the project as they come up, drafting new and evolving sets of specific recommendations which pass from a focus on the industrial partner side of things, to the tech developer side of things, but always with a sustained attention to the effects of the AI integration on the real situation of the worker. As the project proceeds, some Use Cases will receive less attention, as ethical issues become addressed, or as changing tech developer plans or industrial partner interests render those issues moot. By analogy, the project can be thought of as a tree growing from seed, with the ethicist as a gardener: tending, watering, pruning, shaping, and generally watching over the tree as it grows.

This approach is intensive. It involves the ethicist being constantly available – 'on call' – to the tech and industrial partners in order to resolve ethical issues participating in technical discussions between industrial and tech partners, making efforts to understand the technical aspects at the maximum that a layman can – e.g. explainable AI options, platform architecture, project issues related to the Semantic Web – and listening closely so as to draw out and remind the partners of the human side of the situation during such discussions.

To give an example of outcomes of immersion in the technical aspects, which may seem obvious but wasn't in practice, we were engaged in discussions with one tech partner about a VR headset – a commercial product – they had planned to use as a human-machine interface for text recognition of some labels in one of the partner plants. Before the first meeting on the issue we took the time to read the product safety brochures – readily available online – where we discovered a number of warnings, most importantly that the product was not ATEX certified for use in potentially explosive atmospheres. We pointed this out in the meeting, whereupon the process engineer in charge of the Use Case agreed that the device could *definitely not be used* in the plant. Significant changes were made to the development plan for the Use Case as a result, including later finding an ATEX certified VR headset.

The intensive approach also involves constant updates in the form of evolving recommendations. It involves bringing in outside experts, such as work safety experts, to network and discuss specific issues with the tech and industrial partners. – We have made *in-situ* analyses of the working environments in question as well and have had direct discussions with the workers. Incorporating logical and flexible thinking regarding what is affecting what in a production line, and carrying out sustained reflection on what is the real goal of the AI implementation and whether and how it can best be modified so as to cause the least ethical issues for the worker (and managers), have also played a major part.

Some Practical Results of Continuous Guidance

In our experience continuous guidance has produced results. We have influenced *which of the initial Use Cases were chosen to be included in our project, which group (workers or managers) would be responsible for providing AI training feedback*, and *the choice of human-machine interfaces to be used by the workers*, among others. We have also helped the project partners make their specifications more precise – qualitatively and quantitatively – including at early stages of the project, e.g. in clarifying *what they are looking to achieve*, *how long some new task will take*, *what would be an acceptable error rate for AI suggestions*, etc. The approach has also uncovered ethical issues which only an immersive and intensive approach could uncover, of which we give several examples below.

Example 1: Uncovering Ambiguities of Responsibility

In several of the Use Case contexts we encountered situations where a console operator was making decisions regarding changes of multiple parameters as part of a larger continual manufacturing process. The parameters were also watched over – irregularly and usually at a distance – by process engineers, who can call the console operator and give instructions. In subsequent discussions a baseline state of affairs was disclosed in which the process engineers give suggestions sometimes, but the responsibility for carrying them out rests always with the

[9] https://gitlab.inria.fr/kfort/ICTProject/-/blob/main/DataForEthicsInIndustry4Paper/ESM_3.pdf

operator. As they put it: the process engineers 'never touch the controls themselves; they let the operator adjust the parameter.'

The upshot was that the responsibility for adjustments was unclear and irregular. Is an operator responsible simply because he pushed a button, following instructions of a superior, particularly when instructions from the superior are irregular? Of course, this vagueness and irregularity would easily carry over into the stage when the AI service would be used to make some or all parameter adjustments to varying degrees. The working culture of the industrial partner, at issue here, is very much an aspect which is difficult to compartmentalize under the Industry 4.0 label and through broad principles. Human relations – particularly hierarchical relations – in this sense are ancient and not advancing in an iterative manner by any means.

If we generalize from the various situations in the project where this issue arose – again passing from the specific context to a *more* general view –, then we can say that there is a hierarchy in this context but no clearly defined chain of responsibility, i.e. the hierarchy is not strict for certain tasks, thus responsibility cannot be easily imputed. The recommendation which was offered, follows evidently: *formally clarify who is responsible in situations where instructions are given by the process engineer, and record all interventions of the process engineer in adjustments of parameters, before the AI service begins to be used to suggest or carry out adjustments of parameters*.

The specific principle follows likewise – tying back into our suggestions above regarding the tailored approach – **Preliminary Consideration of Ambiguities in Workplace Responsibility:** ***prior to Industrial AI service integrations, examine the proposed work contexts for vagueness and irregularity in the work hierarchy relative to specific tasks, and formally clarify responsibility and keep a record of irregular interventions by superiors in any tasks where the above applies***. In HLEG terms – going up yet another level in terms of generality of principles – this leads to the principle of *Accountability*, as an overarching principle. But you cannot reach broad *Accountability* in terms of AI until, with the more specific principle above, among others, you can ensure that the work context has had its chains of responsibility clarified prior to adding AI.

In other words, *viewed in terms of principles guiding action within experience, ethics is a series of principles overlapping at different levels of generality, and only the continuous involvement and guidance of the ethicist in the real working situation gives access to the specific principles at lower levels of generality which will give good practical results*. AI integration is as subject to this as to any other ethical situation, and even more so in an industrial context. *The human in the context has a specific baseline situation with its own ethical character before any AI enters the picture. It is wise to take this into consideration if the subsequent AI integrated process is to be ethical.*

Example 2: Uncovering Irregularities in Work Tasks

In one Use Case in which the context was the integration of an AI service in order to predict wear in certain parts and optimal time to change those parts in a portion of the manufacturing line, we found through immersive discussion, that some of the operators related to the Use Case check the part wear, while others don't bother to. We also found in the above-mentioned Use Case where bag labels were to be read by AI based text recognition to speed up a process, that in the current context, bag labels are manually input into a central computer by a control board operator, but sometimes, irregularly, the loading operator inputs the label information.

In each of these cases, some or all of the tasks in question may be taken over completely by the AI service. But even if they are, the ethics of the transition phase at least – one can hardly imagine that an AI service will work at 100% immediately without trials and modifications – will depend upon making clear *who is supposed to be doing what, when,* and implementing a protocol to record or deal with irregularities in specific job tasks. All of this can only be done by first uncovering the issues at their level, through an immersive participation of the ethicist.

On site visits to the manufacturing plant help a lot in getting a sense of these actual work conditions. Discussion with the process engineers is good, but – through no fault of the engineers – descriptions in writing or online meetings, of the interactions in the environment, are sometimes limited to abstractions. Plant visits open up a new level of understanding of the interactions of operator, engineer, and machine. They give a sense of time and space limits and of what it might feel like to be *this or that operator*, or be near *this machine*, hot, noisy, and smelly, or have to monitor *these eight control panels* regularly while watching for AI suggestions.

Another example of this came home to us in one of our plant visits where the prior abstract description of the control board operators, their shifts, and their relations, was hugely enhanced by seeing the control room. There we found that the control room occupants were not limited to only the two control board operators who must be there, as had been described to us, but that other operators would wander in and that as many a half dozen operators were in the room discussing operational issues at once. Moreover, we got a sense of the companionship of the operators, the sense of ease that they had with working and interacting with one another. We remarked on this, and the process engineer replied that they hadn't thought of this aspect exactly, "but yes this plant is very much like a family, everyone knows one another really well." Our related recommendation – in line with and complementary to the research of outside work safety experts we talked to on the potential harm to cohesion that introducing technology such as AI causes – was for them to make sure not to lose this strong sense of cohesion by implementing a formal periodic review of worker cohesion and satisfaction. Frequently, human centered ethical concerns are not obvious until you 'go and see for yourself.'

Is the Ethicist as Guiding Supervisor Feasible?

Practically speaking there are a number of problems – costs, limitations, or tradeoffs – with carrying out the above suggestions, that we will note. We do not have space here to respond fully to these, but will lay them out as an avenue for future consideration, along with some brief indications of the ways these problems could be approached. These problems include financial and administrative costs, non or disingenuous participation of industrial partners or technology developers, the strait jacket imposed by non-disclosure agreements and the culture of corporate competition, and the loss of conceptual freedom and shouldering of responsibility for the ethicist.

Added Financial and Administrative Costs

The most obvious of these problems perhaps are the added costs of the direct engagement of the ethicist. To hire an ethicist full-time for such direct engagement would probably necessitate at least a regular salary, or a pro-rata compensation, which would have to be found somewhere. In addition, there would be the administrative costs of travel expenses and hosting the ethicist for regular *in-situ* factory visits or short stays, integrating the ethicist into the industrial or developer team, and preparing security and safety clearances/training and other legal clearances. In the case of our own project these costs have been met by the European project, and the small ethics team of two (a post-doc contributing full time and an associate professor contributing part time) has been able to more or less keep up with the needs of engaging with several industrial partners and the group of tech partners. To cover industry on a large scale would require many such teams, and more extensive funding.

Financial and administrative costs are thus a definite limitation, but we do not think they are an unsurmountable issue. They might be met, for example, by reframing the study of ethics – all branches of ethics – at the doctoral and post-doctorate level generally, to include opportunities to participate directly in whatever other community the graduate or post-doc is most nearly related to. So, a doctoral student or post-doc in medical ethics would include a period of direct interaction with medical practitioners in practice, in animal ethics a period of interaction with nature conservationists, or at farms, or at animal shelters, and so forth. In terms of ethics of technology and industry this could mean voluntarily forming teams in which an industrial partner (e.g. a manufacturing plant) joins with a university department to fund a direct engagement aspect of the ethics doctorate or post-doc. The benefits of this would be various. The initial study of ethics in the doctorate would be reinforced and shaped by the practical engagement with the external partner and since problems are often very context specific, that engagement would provide the 'grist for the mill' of further research. This process could then be repeated in the post-doctorate at a higher level. The industrial partner would benefit by having created the link with the doctoral student or post-doc, gaining another point of view, possibly improving their industrial operation, and having one or more ethicists that they might later turn to – having already built the connections – for either permanent work or temporary consultations. Even a cursory look at numbers of philosophy graduates – of which ethics graduates make up no small part – shows that only a small percentage ever attain permanent academic posts, while many languish hand-to-mouth in part time teaching or other jobs. Thus, such a link between industry and ethicists would be no drain on academia. It would instead be a help on the way to permanent posts, as well as a help to those who are unable to find permanent posts in the usual sense.

It can also be argued however that ethics is a discipline whose financial benefits are indirect and difficult to assess but quite definite enough to deserve funding, even by commercial corporations themselves. To make this case to industry would be to encourage the above suggested partnerships. Finally, it could be argued, similarly to Adam Smith's arguments regarding the military – but perhaps with a stronger claim to the argument – that ethics is a profession which uses resources without increasing productivity, but is nonetheless necessary: ethics does not make money. Such an argument would add weight to calls for government funding to supplement partnerships for the type of ethical engagement envisioned.

Industry Unwilling to Participate

Another issue which could limit the results and application of our ethical approach is the non or disingenuous participation of industrial partners or technology developers. If the industrial partner or tech developer is unwilling to engage in good faith, or at all, with the ethicist, the results are bound to be disappointing. In our own project, the requirement for an ethical overview is imposed at a higher EU level. While the industrial partners and technology developers have participated willingly for the most part, they have sometimes shown much less interest in the ethical aspect than in other aspects of the project. Those who participate less tend to tell us that they know little or nothing about ethics, and sometimes participation seems to be conditional to the resolution of technological development and solutions in terms of commercial interests, so that participating in the ethics of the project is reduced to a matter of getting ahead in the project itself. A related limitation is that we have been able to observe the workers going about their jobs directly, but talking directly to them is more circumscribed by the industrial partners. We also note that we are working within a research project, even though it is in partnership with industry, so it is possible that our results are colored by the willingness of the industrial partners to join the project in the first place, i.e. they are by default open to ethical review, because in some sense they have to be, which might not be the case in other contexts.

Mediocre participation thus presents very serious limitations to the type of engagement we propose. Our response can only be that ethical supervision is fundamentally a teaching and learning moment, an opportunity to form communities striving toward higher types of value. It is best carried out with industrial and technology partners who are voluntarily engaged in the community of practical ethical embedding. But, fundamentally it also requires from the ethicist clarity, continuity, specificity, and giving the reasons behind recommendations, all of which we have tried to bring into our approach. It also involves discussion with other partners at levels which can generate interest, i.e. at the ground level of the specific technological solutions which they are already interested in developing*, but deeper discussion in connection with the human consequences of those technological solutions*. In the end this may be the most difficult limitation to overcome. It is the challenge of the ethicist as interpreter and teacher, but here directed at industrial and tech partners rather than at university students. One might call this ‘teaching by doing,’[10] i.e. being a model by allowing others to see – including other ethicists - how we have gone about the nuts and bolts of engaging particular work contexts. We think this above all is what differentiates our approach from a principles-based approach, and that the two approaches should complement one another.

Corporate Competition and the Need to Anonymize

The ‘strait jacket’ imposed by the context of corporate competition and the need to anonymize, present very different type of limitations than those described above. It revolves not upon getting ethics practically embedded in the integration of artificial intelligence and related technologies into industry, but upon *being able to disseminate the details and results of that effort*, so as to be a help to others who may wish to adopt the method.

The requirement to anonymize – as we were required to do in this article – the specifics of the work contexts and materials, the problems being addressed, and the names of the industrial partners and products, is a limitation. In our project we have also been required to have articles vetted before submission by a project committee with members from all partners. These requirements are common to EU level projects however, including those not in partnership with industry.

Alongside this, business competition also plays a role in what can be disclosed. Industrial partners do not want to let competitors know about any problems, or current practices or changes in technological or organizational setups, which might give their competitors an advantage. Such restrictions are necessary given that business competition is a fact in industry currently. These restrictions affect our ability as ethicists to fully discuss the issues, both with other ethicists and with interested parties in other disciplines, in the really open way we would like. To be able to give details about what is being manufactured and how it is being manufactured would add ‘color’ and interest to ethical issues around the Use Cases. Generating this interest through the ‘human at their work’ details is very much in the spirit of the ground level and immersive ethical approach we are advocating. There is thus a very definite tradeoff between getting things done ethically in industry and being able to show

[10] In the same way that a medical intern might well learn more from following a senior doctor on her rounds and participating in the diagnosis and treatment of actual hospital cases, than by reading and memorizing ‘principles’ from medical textbooks.

others and interest them in what you are doing.

Our response here is to accept the tradeoff, insofar as we have to. We think it is better to ethically safeguard the lot of, e.g. the shop floor worker, as much as we can – even if we can't give all the details – than to remain at a level way above, where vague ethical principles can be discussed easily with our peers.

Loss of Conceptual Freedom and Increased Responsibility

Finally, the potential loss of conceptual freedom and increased shouldering of responsibility for the ethicist is *a problem which effects the ethicist themselves*, rather than the industrial or technology partners, or the larger community of ethicists. What we mean here is that having to directly engage the worker or engineer in the industrial context, or the technology developer in the context of the planning and development of tech solutions can be viewed as a form of external pressure upon the ethicist as a philosopher. This might include, when being funded by industrial or technology partners, the pressure to conform to their expectations and interests, i.e. to 'not bite the hand that feeds' – an ethical problem in its own right –; but the problem is bigger than this. There have been trends in philosophy which advocated philosophy as a practice embedded in context – at least as early as Socrates – but there have been equally strong trends which find the value of philosophy precisely in its ability to break free of the messy world of context embedded experience, and merely *speculate*, including ethically. This freedom is given up in a bigger way than usual in the approach advocated above, not least in getting mixed up in legal requirements, non-disclosure agreements, government policy issues, industrial and technology standards, work safety issues, and others which move beyond the usual boundaries of ethics. In the same vein and because of these interdisciplinary moves, the ethicist also begins to take on responsibility themselves insofar as recommendations become more specific. To formulate broad ethical principles 'which we can all agree on' is to remain in a comforting zone of vague mutual responsibility. On the other hand, recommending *this* definite action/organization/or approach, in *this* space and time context of *these* particular people (e.g. workers), carrying out *this* task, for *these* exact reasons, makes the ethicist individually responsible for the outcome of the recommendation to a much higher degree.

One response to this problem for the ethicist is to suggest that only those who are comfortable with this sort of practical engagement and responsibility undertake it. There has to be an interest in engaging the specific context, in order to proceed with the methods that we have detailed. Our reading of the current literature in AI ethics for industry and in ethics of AI and technology in general indicates that many ethicists are already uncomfortable and disappointed with the current approach of high-level principles. It is to those people that the suggestions of our approach may be most useful to build upon.

Conclusion

Industry 4.0 too easily becomes a buzz-word in which the outlook of successive versions in the realm of software development is overlaid upon more physical based practices and relationships which are not easily amenable to actual practical changes which have ethical consequences. Meanwhile, this, combined with a contemporary emphasis on high level principles in AI ethics has resulted in a shortage of real practical results in ethical engagement of AI issues for heavy industry. In order to address these issues, the main aim of this article has been to outline a new approach for AI ethics in heavy industry, the approach we have taken as the ethics team in an ongoing EU Horizons 2020 project. The two aspects of our method, developed within our project, include: *tailoring ethics to the real time and space situation of the worker*, and *re-envisioning the ethicist as a guiding supervisor continually involved in the development process*.

For us, the first aspect has been to build a baseline view of the real context of an AI service integration before appealing to high-level guidelines or principles. We try to get as full a sense as possible of what is going on before bringing in ethics. We have done this by intense discussion sessions with industrial and technology partners in the project, sometimes all together, sometimes in one on one sessions. We have reviewed the use cases and asked multiple questions from many angles about the workers and engineers, the jobs being done, the environment, the goals and intentions of the industrial and tech partners, the specifics of time and space of the problems and the intended AI solutions. From there we have gone on to make very specific ethical recommendations about which avenues of research are more promising when viewed ethically, which interfaces should or should not be used, which people in the industrial context are best placed to undertake a task, which order the solutions proceed in, how the workers can best be included (e.g. in the context of developing AI Explainability), and others. The developmental choices are then reviewed with the industrial and tech partners as they are carried out and the recommendations are evolved or adapted, or new recommendations are made. The recommendations are also generalized at a low level, if useful to other such work situations, and perhaps linked to higher level principles and guidelines as well. But the emphasis on *practicality and implementation* come first, rather than trying to fit a unique situation into or under a broad ethical principle conceptualized in advance.

The second aspect of our method has been to deliberately implement an intensive participation of the ethicist in the ongoing industrial AI research and integration of our project. To do this we have carried out *in-situ* visits of the manufacturing plants involved, personally evaluated proposed Human-Machine interfaces, participated in technical meetings, studied relevant facets of the technical side of the project (e.g. explainable AI options, platform architecture, and project issues related to the Semantic Web), brought together outside experts such as work safety experts to discuss issues with the other project members, and basically been 'on call' continually to discuss ethical issues as they arise. Another part of our intensive participation has meant being persistently and actively involved through the various stages of AI integration, from planning through technical development and trials. All of this was accompanied by a deliberate effort to gain a sense of: what is really important to shop-floor workers and management, the company culture and actual working practices, and the outlooks and predispositions of developers, in order to uncover ethical issues which would otherwise be overlooked.

We have achieved practical results within the bounds of the project. So far these results include: getting proposed human-machine interfaces changed or abandoned; modifications to the user interfaces of those HMIs; getting certain Use Cases abandoned altogether or heavily modified (including, e.g. changes to who – floor operator or engineer – would be responsible for AI training feedback); specific quantitative estimates from tech partners for changes brought about by introducing AI (e.g. extra time for task, number of images to be analyzed for AI training, etc.); getting commitments to introduce AI in stages with training periods during the implementation phase; streamlining and simplifying AI suggestions to the operator; error protocols and formal clarifications as to responsibility in relation to AI related errors; changes to the way – e.g. de-anthropomorphizing the AI – use case solutions are written up or conceptualized; and getting the developers to formally conceptualize the employees involved on the industrial side in terms of dynamic processes in order to include the human as part of project platform design. Perhaps most importantly, we have gotten the industrial and tech partners to seek out our opinion on issues that they are uncertain about, without prompting, and to deliberately design training data apps with ethics in mind.

We believe these are strong preliminary results. If they are representative of what can be achieved, and subject to the limitations acknowledged above, then the *implications* of our method are that *industry and technology, together and as members of partnerships with other parties, should hire more supervisory ethicists* instead of consulting checklists of principles – which too easily turns into ethics washing –, ethicists meanwhile should take up the job that engineers and technology developers are usually not trained to do which is to *get deeply involved in ethical issues at the technical, organizational, and individual human (worker) level, so as to make specific ethical recommendations which have real impact in industry.* If we truly want a viable industry 4.0 we will need to engage aspects of industry which are now usually ignored in the real working situation of the humans involved. We are confident that the above suggestions toward a bottom up and intensive ethics of AI in industry, are one way to achieve really practical ethical results.

Declarations

Funding: this research was funded by the Université de Lorraine
Conflicts of interest/Competing interests: the authors declare no conflict of interest
Availability of data and material: Not Applicable
Code availability: Not Applicable
Authors' contributions: Not Applicable
Ethics approval: Not Applicable
Consent to participate: Not Applicable
Consent for publication: all authors give consent for publication

Reference List

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

Ghazi Ahamat, Madeleine Chang and Christopher Thomas. “User Needs for AI Assurance,” Centre for Data Ethics and Innovaton blog post, April 16, 2021. https://cdei.blog.gov.uk/2021/04/16/user-needs-for-ai-assurance/ (Accessed 11/01/2022)

Demir, Kadir Alplasan & 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.

Dignum, Virginia. (2021). Responsible AI. Ethics in the Age of Smart Systems, 10th Annual Symposium. April 20th.

Floridi, L. (2008). The Method of Levels of Abstraction. Minds & Machines 18, 303–329. https://doi.org/10.1007/s11023-008-9113-7

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.

Goode, J. Paul. (2020). Artificial intelligence and the future of Nationalism. Nations and Nationalism. 27. 2. 363-376. https://doi.org/10.1111/nana.12684

EU High-Level Expert Group on Artificial Intelligence. (2019). Ethics Guidelines for Trustworthy AI.

Hagendorff, Thilo (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines 30 (1):99-120. 10.1007/s11023-020-09517-8

Jobin, Anna & Ienca, Marcello & Vayena, Effy. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence. 1. 10.1038/s42256-019-0088-2.

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

Lauer, Dave. (2021). You cannot have AI ethics without ethics. AI and Ethics. 1. 21-25. 10.1007/s43681-020-00013-4

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). AI Ethics – Too Principled to Fail?. SSRN Electronic Journal. 10.2139/ssrn.3391293.

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.

Pacaux-Lemoine, Marie-Pierre & Trentesaux, Damien. (2019). ETHICAL RISKS OF HUMAN-MACHINE SYMBIOSIS IN INDUSTRY 4.0: INSIGHTS FROM THE HUMAN-MACHINE COOPERATION APPROACH. IFAC-PapersOnLine. 52. 19-24. 10.1016/j.ifacol.2019.12.077.

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

Satyro, Walter Cardoso & Maria Villas Bôas de Almeida, Cecilia & José A. Pinto, Marcos & Celso Contador, José & F. Giannetti, Biagio & Anderson, Ferreira de Lima & Aurelio Fragomeni, Marco. (2022). Industry 4.0 implementation: The relevance of sustainability and the potential social impact in a developing country. Journal of Cleaner Production. 2022. https://doi.org/10.1016/j.jclepro.2022.130456.

Shilton, Katie & Heidenblad, Donal, & Porter, Adam & Winter, Susan & Kendig, Mary. (2020). Role-Playing Computer Ethics: Designing and Evaluating the Privacy by Design (PbD) Simulation. Science and Engineering Ethics. 26. 2911-2926.

Stanic, Venesa & Hadkina, Marko & Fafandjel, Nikša & Matulja, Tin. (2018). Toward shipbuilding 4.0-an industry 4.0 changing the face of the shipbuilding industry. Brodogradnja. 69. 3. 10.21278/brod69307

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. Cognition 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.

Zhou, Jianlong & Chen, Fang & Berry, Adam & Reed, Mike & Zhang, Shujia & Savage, Siobhan. (2020). A Survey on Ethical Principles of AI and Implementations. 3010-3017. 10.1109/SSCI47803.2020.9308437.