AI & SOCIETY
https://doi.org/10.1007/s00146-020-01069-w

OPEN FORUM

Getting into the engine room: a blueprint to investigate the shadowy steps of AI ethics

Johan Rochel[1] · **Florian Evéquoz**[2]

Received: 30 December 2019 / Accepted: 31 August 2020

Abstract
Enacting an AI system typically requires three iterative phases where AI engineers are in command: selection and preparation of the data, selection and configuration of algorithmic tools, and fine-tuning of the different parameters on the basis of intermediate results. Our main hypothesis is that these phases involve practices with ethical questions. This paper maps these ethical questions and proposes a way to address them in light of a neo-republican understanding of freedom, defined as absence of domination. We thereby identify different types of responsibility held by AI engineers and link them to concrete suggestions on how to improve professional practices. This paper contributes to the literature on AI and ethics by focusing on the work necessary to configure AI systems, thereby offering an input to better practices and an input for societal debates.

Keywords Applied ethics · AI ethics · Data ethics · Data science · Responsible innovation

1 Introduction

The ethics of AI has given rise to an important body of the literature covering a wide range of issues (Müller 2020). Within this body of the literature, this paper focuses on the role played by individuals in the design, development and concrete use of AI systems. More specifically, we want to identify and conceptualize the ethical questions entailed by the apparently technical work necessary to configure AI systems for a specific task. We are convinced that the technical language in which this work is wrapped should not obscure the important decisions made by individuals. The stakes are high: it is not only about the responsibility of the AI engineers in their professional activities, but also about the public good impacted by their choices.

In this paper, we focus on AI systems that rely on machine learning algorithms, including deep neural network systems (Schmidhuber 2016) Enacting an AI system typically requires three iterative phases where human developers are in command. We call them "AI engineers" to underline the fact that they are practitioners programming and configuring computational operations. We map the relevant ethical questions along the main stages of the "Cross-Industry Standard Process for Data Mining" (see below). First, AI engineers prepare the data which will be used to achieve the objectives prescribed by the project leader. Second, AI engineers select and prepare the proper algorithmic tools used to analyse the data. Third, fine-tuning of the system is carried out to improve the intermediate results and to present these results to the project leader in a useful way. Our main claim is that these phases involve practices with ethical dimensions.

This mapping of these different ethical dimensions is carried out with the primary aim of identifying the ethical questions. This first step makes the presence of ethical questions more explicit. The key insight is not to bring ethics into AI, but to highlight the unavoidable presence of ethics in the way AI systems are configured by engineers. The secondary claim is clearly prescriptive in nature in that it links the identification of ethical questions with expected reactions by the AI engineers. The objective here is not only to map ethical issues, but also to provide guidelines on how AI engineers should act. For that purpose, we will assume a specific ethical approach and use it to work through tensions and decisions found in the practices of AI engineers.

To fully justify this prescriptive stance, we need to present our normative benchmark as explicitly as possible. We

✉ Johan Rochel
johan.rochel@gmail.com

Florian Evéquoz
florian.evequoz@hevs.ch

[1] Faculty of Law, University of Zürich, Zürich, Switzerland

[2] HES-SO University of Applied Sciences, Sierre, Switzerland

Published online: 17 September 2020

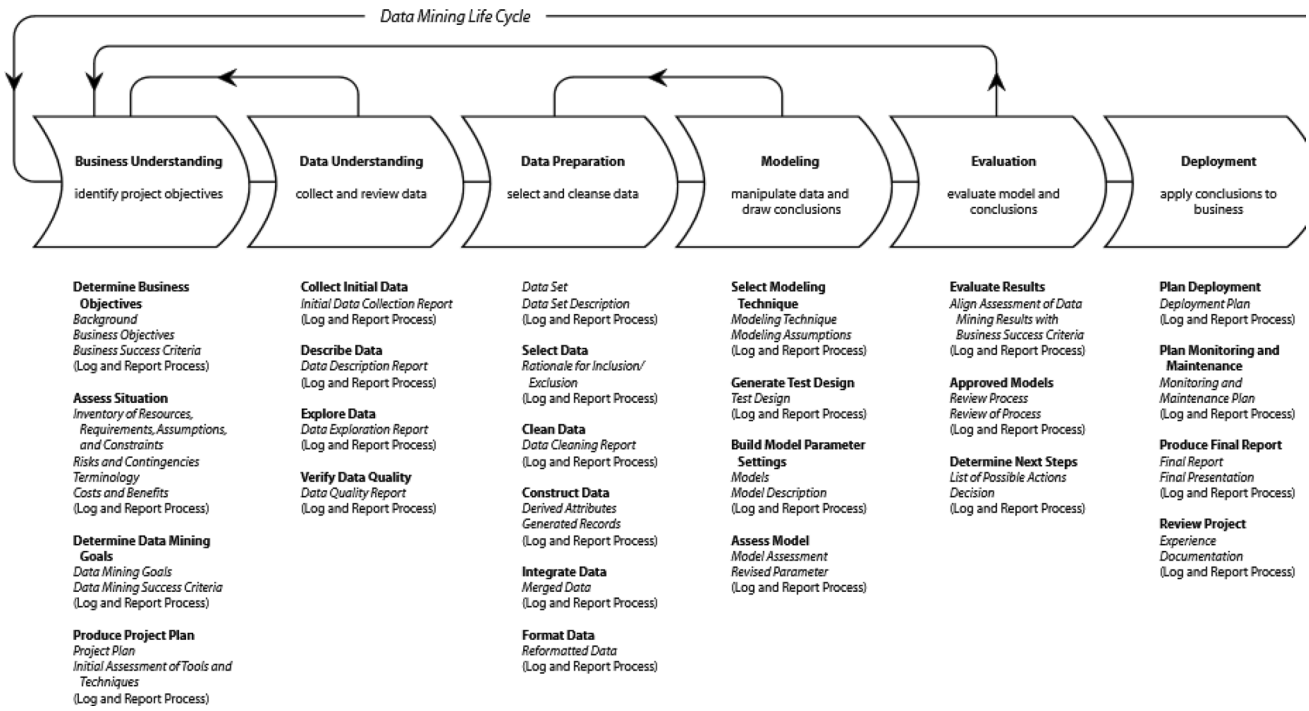

Fig. 1 Cross-industry standard process for data mining (CRISP-DM) (see footnote 4)

shall present the main elements of a republican approach to the responsibility of AI engineers, mainly based upon the idea of freedom defined as non-domination. Gathering these different elements, we will outline concrete recommendations for AI engineers. This ambition to provide concrete recommendations follows the call from Morley et al. to shift from "what" to "how" when it comes to the ethics of AI (Morley, Floridi and Cowls 2019) These recommendations should be taken not only as inputs for a conversation relevant among AI professionals, but also as part of a broader societal debate.

The contribution is organized in three steps. First, we define in detail the focus of investigation and locate it within the literature on the ethics of AI. Second, we make explicit our own theoretical background by defining concepts such as ethics, responsibility and freedom. Third, we present a step-by-step analysis of the different ethical questions identified. Finally, we conclude by identifying lessons learnt for further AI research.

2 The contribution to the literature

Taking the CRISP-DM process visualization shown in Fig. 1 as the basis for our analysis, we will focus on the main ethical questions raised at each step.[1]. As a general point, this visualization already helps to clarify where we claim to make a contribution to current literature.[2]

Overall, we would like to take a general view of the engineer-work done by AI specialists when it comes to enacting an AI system. Within the field which Floridi and Taddeo label "data ethics", they identify three important axes: the ethics of data, the ethics of algorithms and the ethics of practices (Floridi and Taddeo 2016: 1). Our piece is a contribution to the ethics of practices in the field of AI. We focus on the decisions made by AI engineers in configuring an algorithmic system and raise the question of their responsibility in doing so.[3] As proposed by Floridi and Taddeo, the goal is to "define an ethical framework to shape professional codes about responsible innovation, development and usage, which may ensure ethical practices fostering both the progress of data science and the protection of the rights of individuals and groups" (Floridi and Taddeo 2016: 3). As

[1] https://exde.files.wordpress.com/2009/03/crisp_visualguide.png

[2] For a slightly different mapping of the different steps, Morley, Floridi, Kinsey et Elhalal (2019).

[3] This explains why we do not limit the investigation to the strict definition of algorithms as mathematical constructs. As Mittelstadt et al., we also consider the specific configuration of the algorithm for a particular task. As they write, "the configuration of an algorithm to a specific task or data- set does not change its underlying mathematical representation or system implementation; it is rather a further tweaking of the algorithm's operation in relation to a specific case or problem". Mittelstadt, Allo, Taddeo, Wachter et Floridi (2016): 2.

proposed by Elish and Boyd, we focus on the "great deal of mundane work (which) underlies the practices of doing machine learning" (Elish and Boyd 2018: 13). This work includes, among others things, cleaning and/or curating data, managing training and validating data-sets, choosing or designing algorithms, and altering code based on outputs. In this sense, this paper elaborates on the goals defined by Floridi and Taddeo by proposing explicit recommendations for AI engineers for each stage of the CRISP-DM.

Seen from the perspective of the call formulated by Morley et al. to go from "what to how", this focus on practices aims to provide guidance to AI engineers in identifying and addressing ethical challenges. In their paper, Morley et al. examine the growing imbalance in the literature with regards to the attention that is paid to establishing principles and frameworks for AI (Floridi and Cowls 2019; Jobin et al. 2019; Whittlestone et al. 2019; Zeng et al. 2019; Fjeld et al. 2020) as opposed to operationalising these principles into practices and tools. In this paper, we share the authors' concerns and propose further steps towards bridging the gap between principles and practices by making explicit how AI engineers should address ethical challenges found in their practices. We do not provide technical means to do so, but focus on the "soft skills" required for AI engineers (Saltz and Dewar 2019: 202). As shown by the extensive review of resources carried out by Morley et al., there is a gap in the current literature when it comes to providing ethical recommendations which are not abstract principles, and which can be implemented by AI engineers (Morley et al. 2019: 16).[4] We address this gap by using a prescriptive approach (based upon the republican conception of freedom, as defined below), and by connecting this prescriptive stance to concrete recommendations for AI engineers.

In focusing on the process through which AI engineers configure their algorithmic system, we find inspiration in the field of designer ethics (Verbeek 2008). The different steps of the CRISP-DM can be looked at as design and development stages of a data-based AI application. To focus on the work of AI engineers is hence to raise the general question of their responsibility in this design process. As Floridi and Taddeo write, the ethics of practices (including professional ethics and deontology) addresses the pressing questions concerning the responsibilities and liabilities of people and organizations in charge of data processes" (Floridi and Taddeo 2016: 3). We will take as the standard case the situation of AI engineers working for a company/public institution and having to accomplish a data-based analysis for the sake of achieving objectives prescribed by their project leaders (who might be external clients). As we shall explain in the next section, we distinguish between different understandings of responsibility in order to enrich the discussion about what should be expected from AI engineers Poel and Sand (2018). In doing so, we try to connect the growing field of scholarship on "responsible innovation" with the AI ethics scholarship.[5] Specifically, we adapt the definition of "innovators" proposed by van de Poel and Sand and use it for our focus on AI engineers (van de Poel and Sand 2018). In the original contribution innovators are defined as agents who are involved in and shape the innovation process and the resulting innovative products and services. We focus here on AI engineers, raising the general question of their responsibility, with the pragmatic ambition of improving real data-based processes (de Hoop et al. 2016).

This focus means that we set aside two important dimensions of the AI ethics discussion. First, we set aside ethical questions around the data collection conditions. These are obviously crucial questions which are at the core of the data ethics scholarship. Interestingly, these issues are often connected to the responsibility of AI engineers in the sense of making sure that data and privacy (legal) compliance is fulfilled (Taylor and Purtova 2019). We complement this literature by focusing on other issues. In this sense, we will assume as a starting point that data has been collected in a legitimate way. We also assume that this data are stored in a way which does not raise ethical questions. It also implies that we set aside questions related to the environmental impacts of data storage or privacy-related concerns regarding the way these data are transmitted from the data providers to the company working with this data. We focus on the steps which build upon this data.

Second, at the other extreme of the CRISP-DM visualization, we set aside considerations about the broader justice impact of the deployment of AI systems. Most importantly, we set aside justice debates about the impact of AI as part of a deep automation movement potentially changing the way job markets are regulated and the way social policies are funded and organized. As we shall explain later, this means that a large part of the ethical questions linked to the deployment of a specific AI system, or by AI systems in general, are not part of this paper. We stop our analysis at the release of the product by AI engineers.

In short, we contribute to the current literature on the ethics of AI by taking a practice-oriented approach, focused on

[4] For a similar objective, see the "10 simple questions" identified by Zook, Barocas, boyd, Crawford, Keller, Gangadharan, Goodman, Hollander, Koenig, Metcalf, Narayanan, Nelson et Pasquale (2017). See also the questions prepared by the Open Data Institute, https://docs.google.com/document/d/1OXSrA2KDMVkHroxs_8SUoQZ5Uv0eRhtNNtIl9g_Q47M/edit

[5] On the definition of responsible innovation, see Blok and Lemmens (2015). For a critical analysis, Timmermans et Blok (2018). For a similar linkage between data-sciences and responsible innovation in the context of public research projects, Stahl et Wright (2018).

a specific category of individuals involved in the conception and creation of AI systems and we try to provide guidance in identifying and addressing ethical challenges.

3 Theoretical background

We start this investigation with a working definition of what an ethical question is. For the sake of this paper, an ethical question is mainly defined as a normative question, i.e. a question about the reasons an individual should be able to formulate to justify for his/her decision. This working definition might be unpacked in the following way. First of all, ethical questions arise where individuals have to make a decision. As soon as an individual has to choose an option among several possible options, we assume that a first evaluative judgment is necessary. This point is a conceptual point about the necessary steps at stake in making a decision. It does not mean that every individual—even that any individual—does actually decide in such a way. Many decisions we will address later are seen as implicit or purely technical decisions, thereby overlooking the evaluative step prior to any decision.

Second, if an individual has to make a decision, he/she might be required to provide a justification as to why a specific decision was made. This justification about the reasons for action is the core part of the normative question.

Third, this decision and its justification relate to the values which the individual performing the action thinks relevant. As written by Elish and Boyd, the numerous, apparently purely technical choices made by an engineer represent the "minutia where cultural values are embedded into systems. Every step requires countless decisions and trade-offs. In an imaginary if ideal world, code is bug-free, data are is straightforward, and algorithms are perfect fits for the desired task. Reality is much messier." (Elish and Boyd 2018: 13) We share this diagnosis and bring it one step further in trying to identify what AI engineers should do to address this messy reality. As formulated by Hagendorff, this means that the kind of AI ethics we are proposing should be able to "behave sensitively towards individual situations and specific technical assemblages". As he writes, we need an AI ethics which "deals less with AI as such, than with ways of deviation or distancing oneself from problematic routines of action, with uncovering blind spots in knowledge, and of gaining individual self- responsibility" (Hagendorff 2020: 114).

In this short contribution, we can only briefly explain the structure of the argument at stake and outline a substantial position. With respect to the structure, the argument is broadly consistency based with respect to the values of a liberal-democratic society.[6] The objective is to get a large consensus on the type of approach suggested to address the questions, although different justifications might be proposed. This approach recalls the "overlapping consensus" made popular by Rawls (1993). This part of the argument can be linked to the issue of trustworthy AI and, more generally, to the requirements formulated for AI systems, such as transparency, explicability, and accountability (Mittelstadt et al. 2016).

The justification we would like to outline is a (neo)republican approach to the responsibility of AI engineers. This approach is based upon the position developed by Philip Pettit in his work on freedom defined as non-domination. This outline of a republican approach must be read in light of the objective of this contribution: identifying ethical questions and proposing a way for AI engineers to address them. As expressed by Floridi, this objective is part of the formulation and justification of a "first-order framework of implicit expectations, attitudes, and practices that can facilitate and promote morally good decisions and actions" (Floridi 2013: 737). The republican approach seems particularly promising for two reasons. First, it is related to on-going discussions about the meaning of individual freedom with respect to specific expectations for AI-systems (such as explicability) (Floridi and Cowls 2019). In light of the fact that Pettit himself has framed his freedom as non-domination as an alternative to positive and negative freedom conceptions, republican freedom seems a good candidate to foster the freedom discussion in AI ethics. Second, the republican approach is best equipped to address new threats made possible by digital technologies.[7] As we explain below, the capacity of the republican conception to take potential interferences and power relations into account is crucial. Overall, it is important to highlight that this republican approach is but one possibility for identifying and justifying the recommendations. The methodology adopted does not require a claim that republicanism is the "best" approach. We claim that the republican approach is fruitful in justifying concrete recommendations for AI engineers. In Buchanan's words, the successful implementation of a prescriptive theory shall be synonymous with a "significant moral improvement over the status quo" (Buchanan 2004: 63).

To recall Pettit's original definition, domination is defined as arbitrary intereference. An interference is arbitrary if there is no mechanism that requires the interferer to track the relevant interests of the interferee (Pettit 1997: 52). For Pettit, the political ideal of non-domination is a permanent

[6] For a similar structure using the concept of sustainability as substantial value, Taylor et Purtova (2019). For a human rights approach, Andersen (2018): 34–35.

[7] With respect to surveillance in general, Hoye et Monaghan (2018). With respect to privacy, Roberts (2015); Newell (2014). In the context of workplace relations, see the ideas developed by Lazar-Gillard (2018).

effort to diminish arbitrary interferences and transform them into non-arbitrary interferences.[8] In a nutshell, individuals or institutions in a position to interfere with me should be forced to track and take into account my interests (Pettit 2008: 117).

Three main elements of this republican approach are relevant for our work on AI engineers. First, individual freedom is always to be conceived of as being within a social relationship (with other individuals or with institutions and political communities). In this context, the importance of a secured enjoyment of freedom defined as non-domination is particularly attractive as a relational account, that is, an account that considers the multiple patterns of influences that exist among individuals, companies, institutions or political communities (Young 2007: 39–58). It can also take into account the particular risks attached to the imbalances of power among different actors and the sometimes diffuse risks these relations can represent in terms of (potential) arbitrary interferences.

Second, within this relationship, some actors might exercise arbitrary interferences upon others. Even in the total absence of interference, individuals can be considered to be dominated if they are at the mercy of decisions made by others (Pettit 1997: 73 ff.; Bellamy 2011: 132). In a strong sense, individuals have to be empowered to be free or, as Valentini writes, have to enjoy freedom as a kind of "independence" (Valentini 2011: 162).

Third, to try to diminish this domination is about building procedural guarantees, which make sure that individuals and institutions, especially powerful ones, can be controlled. Measures can range from public mechanisms (constitutional guarantees, mechanisms forcing consideration of the interests of the individuals affected, and contestatory democracy) to private mechanisms (professional codex, measures enacted by a company for its employees).

The notion of responsibility is especially important in this context. It allows us to specify the link between the republican conception and the framework of responsible innovation (Pellizzoni 2019). We rely here on the conceptual framing proposed by van de Poel and Sand about the "variety of responsibilities" in the context of innovation (van de Poel and Sand 2018). According to them, we might first distinguish between a retrospective and a prospective concept of responsibility. While retrospective responsibility has to do with the question of the responsibility an actor bears for an action (or omission) in the past, prospective responsibility is about actions to be taken (or to be omitted) in the future in addressing a situation. Building upon this distinction, van de Poel and Sand map different understandings of responsibility. As parts of the retrospective responsibility, they list responsibility-as-blameworthiness, as-accountability, as-liability. As parts of the prospective responsibility, they list responsibility-as-obligation and responsibility-as-virtue. Issues of responsibility are usually addressed as blameworthiness and liability issues for past errors (Giuffrida 2019). In that sense, AI engineers might be held responsible for what they have done in the past.

Going beyond this classical focus of responsibility as liability, our goal is to enrich the discussion by focusing mainly on two aspects highlighted by van de Poel and Sand as being particularly relevant for actors involved in innovation processes (van de Poel and Sand 2018: 14–16). First, we will cast light on retrospective responsibility as accountability. Van de Poel and Sand define this accountability as having a "prescriptive dimension as it presumes the ability and willingness to account for one's actions and to justify them to others" (van de Poel and Sand 2018: 5). This understanding of responsibility is of direct interest to us, because it depends on the quality of the justification offered in the context of a specific community. This understanding of "justification" refers to our fundamental condition as human beings acting and reflecting upon reasons in a potentially conflicting situation. As Rawls explains, "justification as argument is addressed to those who disagree with us […]; being designed to reconcile by reason, justification proceeds from what all parties to the discussion hold in common" (Rawls 1971: 580). This understanding of justification refers to a deliberative exercise of critical and comparative arguments since it confronts rival normative propositions for a specific decision "against a background presumption of possible objection" (Simmons 2001: 124).

Second, we will focus on responsibility-as-virtue. This understanding is focused on certain character traits of the innovator: "this can be exemplified with an agent's disposition to assume or to take responsibility and an awareness of a range of relevant normative demands" (van de Poel and Sand 2018: 6). According to this understanding, responsibility-as-virtue is associated with due care to others. In the context of the investigation to come, the idea is to underline that AI engineers play a role with an impact which goes beyond their "technical" tasks. This brings us back to the contribution of this paper to the ethics of practices identified by Floridi and Taddeo. Responsibility-as-virtue underlines the requirement to foster the soft skills of AI engineers to empower them to see their activities as part of a community's life, with specific expectations in terms of justifications and values. They are part of a broader societal debate in which they are required to play an active role in integrating others' perspectives.

[8] As Pettit writes, "an act is arbitrary, in this usage, by virtue of the controls—specifically, the lack of controls— under which it materializes, not by virtue of the particular consequences to which it gives rise." Pettit (1997), 55. See also Pettit (2010): 75.

To summarize, we have outlined a theoretical framework relying upon a working definition of an "ethical question", a republican position on the definition of freedom as non-domination, and a particular understanding of the concept of responsibility.

4 Step-by-step analysis

We proceed with a step-by-step analysis along the CRISP-DM process reprinted above. We do not address in detail the "Business understanding" step. This step raises different ethical questions which pertain to the general normative quality of the objectives prescribed. To use a simple distinction, there are questions related to the project bearers (what if the mafia requires you to perform a data-mining project?) and to the project objectives (what if the data-mining project is conceived for an illegitimate purpose?). We broadly assume that both project bearers and objectives are legitimate.

4.1 Understanding and preparing the data

The first set of ethical questions relates to the selection and preparation of data for the sake of fulfilling the foreseen objectives.

4.1.1 Selecting the data

Ethical questions are raised by the different steps necessary to properly identify and select the data with respect to their adequacy for the objectives in question. The AI engineer has to make several decisions as to which parts of the data-set he/she should use for the purpose of achieving the objectives. This first normative moment relates to the adequacy of the data for the objectives. AI engineers do not look for any sort of data, they looks for data which allows the fulfilling of their objectives. This required adequacy brings us back to the understanding of the objectives. To identify the specific data required for the objectives, the AI engineer has to show an in-depth understanding of the objectives: not only a superficial understanding of what achieving these objectives means, but also an understanding of why these objectives should be achieved. Understanding the reasons and motivations behind the objective will enable the AI engineer to choose the most adequate data. We can subsume these reasons under the concept of the "rationale" of a specific AI project. To identify this rationale is not a decisive difficulty, but a good example of the information which an AI engineer needs to make an informed decision.

This in-depth understanding of the reasons behind the objectives is even more necessary when it comes to the construction of the data-set used. This step is not only about a selection among pre-existing data-sets, but about creating a data-set specifically for the project. The choices made in constructing the data-set are crucial for the responsibility of the AI engineer. His/her decisions are the ones which design the data-set. These decisions are often seen as the implication of the search for efficiency. However, as we will address below when dealing with the choice of algorithmic tools, there are a number of open questions around this idea of efficiency.

With respect to the construction of the data-set, a further point is interesting: the appearance of neutrality. Different data-sets exist, and the main job of the AI engineer is to select the most adequate one. But this view obscures the fact that existing data already relies upon certain presuppositions. Put in plain terms, there is no possible neutrality in the way we apprehend and classify the world (Bowker and Star 2000). The categories which we use reflect certain presuppositions about the world and the relative importance of potential perspectives on the world (Gebru et al. 2018). Most importantly for machine learning, the definition of categories raises the question of the choices made by AI engineers about the boundaries of these categories. As noted by Elish and Boyd, "for a machine learning system to work, data scientists must make choices about how to provide discrete labels and generate bounded categories for sensitive topics or in cases where such boundaries are far from solidified" (Elish and Boyd 2018: 14). For instance, we might relatively easily agree about the majority of cases to be included into the category "human being", but not on its specific boundaries (beginning or end of life, but also cyborgness). To include specific existing data-sets into a new data-set tests the responsibility of the AI engineers. By taking these data-sets as an integral part of his/her work, he/she integrates presuppositions made by others. In some cases—as in the examples of disputed boundaries—this integration comes with important ethical questions which need to be addressed by the AI engineer.

It is interesting to note, following Leonelli, that the same reflection applies to engineers in charge of designing large databases (e.g. for international research) (Leonelli 2016: 5). The terminology adopted when describing the data, and the type of software privileged for further research within the database can have significant impacts on the further use of the data in question. AI engineers designing such databases need a well-informed understanding of the objectives and their underlying reasons.

It is worth noting that attempting not to select any specific data among all the data available does not solve this difficulty. AI engineers could not select the data that seems the most appropriate, but instead feed the algorithm with the whole data at hand, trusting it to separate the signal from the noise. From an ethical perspective, this choice not to choose nonetheless bears an ethical value in itself. In the sense of responsibility as accountability outlined above,

the AI engineer owes his/her team (and the broader public) a justification for the trust put into the automated separation. Linked with the republican conception outlined above, the risk is to exercise a form of domination on the people affected by the AI systems, although these people might not have been identified or might not even be identifiable. The interference exercised through the algorithms is arbitrary if the interests of these potentially affected individuals are not somehow tracked and taken into account. To require a satisfactory justification from AI engineers is a way to institutionalize a requirement to take the interests of those affected into account.[9]

4.1.2 Preparing the data

Assuming the AI engineer can define and justify the use of specific categories as an integral part of the data-set, there is a further challenge in cleaning the data, most importantly with missing data or incomplete data-sets. This issue raises important challenges for the AI engineer. Should one try to correct these data-sets and, if yes, how? Depending upon the specific situation, distinct statistical methods are available to address this challenge. For example, one might infer the value of a missing data point by averaging the values of two adjacent ones. If too many data points are missing for a specific category, an option might be to leave out the category entirely, therefore, explicitly filtering out the existing data points in that particular category. The ethical point is about making the normative dimensions of these steps appear explicitly.

Two dimensions seem especially relevant. The first dimension pertains to the general objective pursued in replacing/generating data for missing data-sets. The normative decision is not the same if one pursues a general objective of efficiency (data is replaced for the sake of making the use of the algorithmic tool possible) or if issues related to representation are at stake (data are replaced for the sake of securing a balanced approach). This normative challenge brings us back to the understanding of the general objective being pursued (as part of the business understanding). The second dimension appears if one decides to generate artificial data to replace missing data. The assumptions which are used for this process of generation raise normative questions. These questions might be approached with a broad strategy to prevent unjustified distortion in the data-set. The strategy raises questions which we will address below on the quality of data. Replacing data might also be approached with a clear normative position: data is replaced with the goal of creating a more balanced data-set (defined along distinct possible benchmarks). In both situations, as noted by Elish and Boyd, AI engineers "must clean the training data to address weaknesses, while also assessing how constructed categories and data outliers might contort the model" (Elish and Boyd 2018: 14). These decisions are not purely technical. They involve ethical arguments which need to be made explicit.

The same considerations apply to the integration of distinct data schemas and to the handling of potentially conflicting schemas. To take a concrete example, AI engineers asked to integrate two data schemas might be required to verify information common to the two data-sets (e.g. name) before integrating them. The ethical question raised by this integration bears upon the degree of verification expected to secure the success of the operation. Two data-sets might contain the same name, but it might refer to two different people. Therefore joining data on common names might create wrong associations. The engineer is expected to perform a sufficient due-diligence test in assessing and verifying the information involved.

4.1.3 Assessing data quality

A key part of the preparation of data relates to assessing the quality of said data. The concept of quality might be used in a descriptive way, focusing mainly on a mathematical-statistical understanding of quality. This first understanding is very functional as it judges quality based on the requirements for achieving specific analytical goals. It is about data fitting predetermined analytical purposes. Using descriptive statistics, the AI engineer will generally inspect dimensions of the data to identify quality issues (e.g. "outliers" due to a misspelled locality name) with the purpose of correcting them, or to assess how realistic the data appears to be (compared to some predefined benchmark). The focus is here on correcting for "input errors".

A more normative understanding of data quality points out to a set of questions around the issue of bias. This issue is currently the topic of a rich interdisciplinary scholarship (Veale and Binns 2017). For our present purposes, the key point is again to underline the requirement to make explicit normative challenges to avoid potential domination. To assess whether data are "biased", we need to define a normative benchmark defining which statistical differences are treated in a specific way and "red-flagged" with respect to specific objectives. This requirement applies whether data-sets are systematically tested to make statistical peculiarities appear or whether data-sets are analysed in light of a specific potential bias (like a gender-bias). In the first case, the assessment of whether statistical peculiarities are normatively problematic also requires the identification of a normative benchmark. Obviously, the same is true for the

[9] This specific mechanism can be integrated into the feature of "anticipation" for responsible innovation. Stilgoe, Owen et Macnaghten (2013): 1572.

second approach where you need to identify a potential ground for bias in the first place.

This requirement provides an important link to the discussion about values outlined in the second section of this chapter. The normative benchmark identified should be aligned with the values found important in a specific society. To make this link explicit, we can formulate an argument linking values and principles (such as those found in a constitution) with the responsibility of AI engineers. The argument has the following structure: assuming that value *X* (e.g. gender equality) is an important value, data-sets should be assessed from the perspective of this value.

This identification step is the initial stage of the data quality investigation. The following step is about identifying the required actions. This is again a decisive step where the AI engineer takes a strong normative position on assessing the severity of the problem and the potential means to address it. The AI engineer might take note of existing bias in the data-set but does nothing to address it, for instance, because he/she thinks this bias irrelevant for the objective. Clearly enough, it is a very different decision if he/she decides to statistically "correct" the data-set to more fully respect the normative benchmark identified. The point is not to argue that this correction approach should be adopted in any situation of bias, but rather to argue for the requirement of making normative assumptions explicit.

There might indeed be situations in which the bias of the data-set is used as a crucial component of the data-driven analysis provided. If we take the example of an algorithmic system used for the sake of insurance calculation (say, car insurance), the potential bias found in the data-set (type of car accidents and type of people involved) might be extremely relevant, if not necessary, for achieving the objective (calculate how much people buying the insurance should pay). In such cases, the discussion about bias in the data-set raises a broader discussion about the implications of a normative benchmark for the overall objective. In the example mentioned, it might be argued that gender should not be allowed as legitimate grounds for differential treatment (as decided by the European Court of Justice in 2012[10]). It means that the calculation of car insurance costs shall no longer consider the differences between men and women. This legal limitation puts a limitation on the margin of appreciation which AI engineers have. By analogy, the same might be true for non-legally enforced discrimination grounds. A company might commit itself not to use specific grounds, and therefore address this point already in its assessment of the data quality. As explained by Floridi and Mittelstadt, an "ethics of care" might be relevant in addressing these potential discriminations against particular groups. Particular forms of practices or hypotheses could then be set aside as "off limits". However, they also note that "alternatively, it may be possible to conceive of privacy as a group-level concept and thus speak of 'group privacy rights' that could restrict the flow and acceptable uses of aggregated datasets and profiling" (Mittelstadt and Floridi 2016: 328).

Seen from the perspective of a republican definition of freedom, two important elements must be mentioned. On one hand, the value of equality among individuals is looming large in the bias discussion. Beyond fundamental moral equality among human beings defined as a premise of a legitimate political order, equality in its political dimension is key for republicanism. The idea of citizenship as realization of this political equality is a crucial idea for republicanism. As put by Bellamy, avoiding domination "implies a condition of equal respect among citizens, therefore, in which each can look the other in the eye through enjoying equality of status in the making of the collective decisions that govern their lives" (Bellamy 2019: 63). In this respect, the objective of ensuring non-domination is also about securing political equality among the members of a given political community (Besson and Martí 2009: 20–21). Data-based products and services should be checked from their potential negative impacts on the equality of individuals.

On the other hand, this concern about equality might be underlined using a freedom-based argument. This argument is then about preventing interferences from happening on specific grounds. Certain grounds—such as those linked to the identifying features of a person—are considered illegitimate grounds for public authority (and further private actors) to engage in these interferences. This reconstruction might be further refined by drawing upon the freedom-based account developed by Moreau (Moreau 2010, 2013). For her, a person has certain deliberative freedoms which should be protected. These freedoms should make sure that our decisions about how we live are protected against the effects of normatively extraneous features (Moreau 2010: 156). In other words, the features should not bear upon us as "costs" when making decisions about how we want to live. These freedoms are not the result of an interpersonal comparison in terms of opportunities or rights, but reflect what is due to the person in terms of recognising his/her entitlements. Identifying these entitlements requires developing a view of the human person and his/her protected features. Overall, both arguments make clear that republicanism puts a strong focus on preventing illegitimate bias from impacting people.

4.1.4 Recommendations

In light of this analysis and to prevent domination from happening (by him/her or through him/her), the AI engineer should:

[10] See the press information on the case, https://ec.europa.eu/commission/presscorner/detail/en/IP_12_1430

- Be fully informed about the objectives and the underlying reasons behind these objectives
- Use a precise and complete description of the data, making explicit the categories and concepts used for the description
- Clarify the normative presuppositions upon which datasets rely, especially for cases of disputed categories
- Make explicit the normative dimensions of the data preparation techniques, thereby considering the rationale of the project
- Give special attention to the case of missing/unusable data and the techniques used to correct them. Document these techniques.
- Clarify the normative benchmark used for the identification and assessment of data quality (mainly bias involved in the data-set)
- Give special attention to the normative implications of the different data quality correction techniques

4.2 Modelling the data

The second set of ethical questions relates to the stage that follows the data preparation stage. CRIPS-DM calls it the "modelling" step, focusing on the manipulation of prepared data with algorithmic tools. This second set of questions is centred on the way algorithmic tools are chosen, configured and used.

Similar to the questions identified above, the main challenges here relate to the requirement for AI engineers to make explicit their ethical decisions and the reasons underlying them. This measure is a mechanism meant to diminish risks of domination. First, this most clearly bears upon the choice of algorithmic tools for the sake of achieving specific purposes. The AI engineer has expertise in selecting the most suitable algorithmic tools for specific purposes. Because of this very instrumental relation—a tool for an objective—the normative dimension of this choice might easily be overlooked. The situation is addressed as a question of instrumental rationality, similar to the selection of datasets and categories briefly addressed above. This rationality is defined in terms of efficiency, similar to the Ockham's razor approach. However, as noted by Mittelstadt et al., "algorithms are inescapably value-laden" (Mittelstadt et al. 2016). This is relevant for the selection of the suitable tools, but also for the functioning of the tools themselves.

On one hand, the selection is a general question about this rationality as efficiency. The claim that AI engineers should select the most efficient algorithm might give the impression of normative neutrality, but this impression is misleading. Efficiency in the sense of "performance" is a normatively loaded concept. The idea that efficiency is only about the "good sense" of doing things with the least possible effort does not do justice to the richness of the concept (Schultz 2001). This issue has been a long-standing one for ethical debates in economics (Staveren 2007). For the purpose of this article, it is important to remain aware that efficiency does always rely, at the very least, upon a determination of the type of resources required for a specific task. If a specific algorithmic system is said to be "the most efficient in achieving an objective", this proposition relies upon a determination of the terms of efficiency. To make the normative dimension appear clearly, imagine this efficiency in terms of speed in achieving the result or in terms of impact on the use of energy and natural resources. In both cases, the tool is the "most efficient", but the terms of this efficiency are distinct.

On the other hand, algorithm selection is a question about the drawbacks of this rationality as efficiency.[11] The selection of the most suitable algorithmic tools does not answer the question of their potential problems. The choice of any tool raises questions about which elements become secondary or even hidden. As explained by Veale and Binns, "neural networks or random forests are more amenable to capturing synergy between variables than linear regression. Use of regression might omit important contextual variance, for example. Within a model family, further hyperparameters must be specified" (Veale and Binns 2017: 2–3). Hyperparameters include configuration of the models that cannot be "learned" from the data, e.g. the number of layers and neurons in a neural network model. Among other things, these parameters impact the prediction performance of the model, the generalizability of the predictions it is able to make, and the computational complexity of the model overall. The choice of the model and the setting of hyperparameters are decisions that are generally made based upon experience in similar contexts (scientific literature applied to the same domain) that defines some kind of a normative standard.

As stated above, efficiency of the model (or "performance") is a crucial element in the choice and parameterization of the model. In a typical project, AI engineers will test and compare different models and parameter settings. There are standard measures used to assess performance, defined as measurement of the quality of prediction. For example, ROC and the area below them (AUC) are widely used measures (Bradley 1997). However, specific projects might call for different measures. While the AUC computes the quality of predictions considering both false positives and false negatives with the same weight, there are cases where it is preferable to favor one over the other. For example, in the case of spam detection, it is preferable to minimize false positives (i.e. genuine email wrongly classified as spam), while some false negatives (i.e. some spam not correctly detected as such) are acceptable. Additionally,

[11] For an analysis of these drawbacks from the perspective of the trade-off between efficiency and equality. Jimenez-Buedo (2011).

there are different methods to perform these measurements, each with their specific advantages and disadvantages (e.g. cross-validation tests).

Linking these two questions makes clear that, even under the assumption of a clearly defined instrumental rationality, the selection of any algorithmic tool is actually the result of a balance of sometimes conflicting interests. Choosing one tool over others means identifying the potential negative points of this choice and balancing them with potential positive points. This situation might be exemplified by a classical trade-off in machine learning. Specific techniques might produce very good results, but come with a high level of complexity when it comes to explaining why or how a specific result was reached. To use such techniques is the result of a decision that balances the suitability of the techniques for the determined objectives and the difficulty in making the result transparent and understandable. In general, AI engineers can use techniques in "adversarial machine learning" to better identify vulnerabilities in the models (Papernot et al. 2017).

For all these questions, the republican conception of freedom justifies a requirement of justification as an anti-domination mechanism. Because AI engineers are required to make sure that they can justify their decisions, they are forced to consider potentially negative impacts on affected individuals. The technical features of the situation should not obscure the fact that the choices made by AI engineers have impacts on the result of the project and hence impacts on people. Especially with respect to the concept of "efficiency", there is an important public dimension in the justification required from AI engineers under the heading of their accountability. This dimension connects to the general ambition of illuminating the negative impacts of taking efficiency as an overall guiding principle for social interactions. There is on one hand the necessity to clearly define efficiency and show the inherent trade-offs which the concept has. On the other hand, it is also about making explicit the consequence of relying primarily or exclusively on efficiency to guide and organize human interactions on the basis of data-based tools.

4.2.1 Recommendations

In light of this analysis and to prevent domination from happening (by him/her or through him/her), the AI engineer should:

- Make explicit the terms of the efficiency as performance he/she is striving towards and their link to the project objectives
- List and evaluate the implications of the specific algorithmic tools chosen, especially its potential drawbacks
- Explicitly address trade-offs between conflicting legitimate objectives in selecting specific tools.

4.3 Evaluating the results

The third set of questions relates to the evaluation of the preliminary results and the fine-tuning iterative steps taken by AI engineers to improve these results. We will also consider ethical questions related to the final preparation of such an analysis, i.e. preparing it for submission to the project leader.

4.3.1 Interpretation

The first question to be addressed in this third stage concerns the benchmarks used to assess the intermediate results. In a situation with a pre-defined objective, we come back to the point mentioned at the beginning about the normative dimension of this objective. The AI engineer has to interpret the results in light of a normatively pre-defined benchmark. In a situation without a pre-defined objective, the AI engineer should propose an interpretation of the results without an external benchmark. As noted by Elish and Boyd, "because machine learning results can be difficult to interpret, there is a danger that data scientists might inappropriately use the results when converting them back into conceptual information for decision-making" (Elish and Boyd 2018: 15). Two specific dangers are well identified in the literature. On one hand, there is a danger of over-interpretation of correlations or suggested connections between data which have no real connection.[12] On the other hand, there is a danger of over-fitting, meaning the use of the model to explain the "noise" in the data rather than its substantial elements. Following Elish and Boyd, these two standard dangers require an attentive AI engineer to flag-up problematic situations.

In both cases (with or without pre-defined objectives), the context in which the AI engineer is asked to interpret the result is also decisive (Mittelstadt and Floridi 2016: 322–323). In light of our reflection on domination in the professional context in which the engineer works, pressures applied to him/her could distort the interpretation of the results. Issues such as stress, time pressure (implicit or explicit), or unrealistic expectations play a central role in securing for AI engineers the capacity to properly interpret the results of their analyses.

4.3.2 Fine tuning

Assuming a scenario in which the result is not suitable, we can raise a second set of questions focused on the efforts of the AI engineer to adapt the data-set or the algorithmic tool

[12] See e.g. Calude et Longo (2017). Mittelstadt et al. call this problem an "epistemic concern on inconclusive evidence". Mittelstadt, Allo, Taddeo, Wachter et Floridi, (2016), 6.

to obtain more suitable results. In these fine-tuning efforts to change the parameters of the AI system, the AI engineer will enact a number of decisions meant to improve said results. These decisions might be broadly ordered in three different categories.

First, it is possible to completely change the algorithmic tool used. In that case, as illustrated on the CRISP-DM visualization, the process goes back to the second stage where algorithmic tools are selected. Second, it is possible to adapt specific parameters of a given tool. These adaptations rely upon a renewed interpretation of the potential trade-offs identified in the second step. Third, the AI engineer might adapt the data-set he/she uses. He/she might statistically correct "errors" or improve the representation of specific categories. These manipulations raise the questions identified in the first stage.

With these types of measures, the AI engineer tries to come as close as possible to the objective identified. Having reached this point, the questions of deployment, understood as contact with the project leader, will be extremely relevant. As explained in the introduction, we do not consider here broad questions of deployment regarding impacts on society and the related justice questions. We focus on deployment questions immediately related to the work of the AI engineer. In this context, the question of the communication with the project leader is a key normative question.

An important part of the responsibility of the AI engineer relates to the communication about shortcomings and limitations of the tools used and the results achieved. There is a risk of domination by the AI engineer. He/she might have the capacity to arbitrarily interfere with the project leader's interests, i.e. without having to track his/her relevant interests. In other words, the project leader might be at the mercy of the AI engineer.

To prevent this domination, the AI engineer should clearly inform the project leader about elements which are of interest when it comes to the application of his/her results. This responsibility is especially important if the AI engineer has information about the final use of his/her analysis' result. To make a clear example, imagine that the engineer knows that his/her work will be used as a basis for fully automated decision-making tools directly impacting individuals. In that situation—meaning taking into account the relevance of his/her work and its future use—he/she has a strong responsibility to inform the clients about potential shortcomings of the work. These shortcomings might have a very important impact on the users of the algorithmic system and on those impacted by this same system. This responsibility is arguably less important if the work done by the AI designer is only used by professionals as a secondary means to support them in decision-making. In this case, there is also a responsibility to thoroughly inform, but a less crucial one with respect to potential impacts.

We might abstract from these two examples a rule stating that the degree of responsibility that an AI engineer has depends upon his/her own level of information about the future use of his/her work. To prevent his/her own potential for domination, the AI engineer should get into an in-depth dialog with the project leader. This in-depth dialog should not be an ad-hoc discussion upon completion of the project. Instead, it should form an integral part of project development and realization. It should be protected as an important space for exchanges aimed at preventing domination and arbitrary interference from happening. It is important to highlight that the risk of domination exists even without the ambition to harm the project leader. Information might not be transmitted because of external conditions (time limits, stress, no forums for exchange). In preparation for this exchange, the engineer should be able to assess the sensitivity of the use of his/her analysis. This assessment might take inspiration from approaches developed in trying to account for the level of awareness necessary for AI systems.[13] At stake here is the potential domination exercised by the AI engineers upon the project leader and, indirectly, upon users or persons affected by the AI system.

4.3.3 Recommendations

In light of this analysis and to prevent domination from happening (by him/her or through him/her), the AI engineer should:

- Make as explicit as possible the benchmark used to assess intermediate results (in light of the rationales of the project)
- Proceed with the steps identified above if he/she fine-tunes the data-set or the algorithmic tool (iteration)
- Organize a space for the discussion of shortcomings and limitations of his/her results with the project leader
- Communicate as transparently and explicitly as possible the shortcomings and limitations of these results with the project leader
- Inform the project leader in good faith about the use of the results.

[13] This functional approach is taken in the medical devices sector. These devices are assessed with respect to the functions they fulfil and their relevance for individuals. They are classified on a corresponding scale. See "Essential Principles of Safety and Performance of Medical Devices and IVD Medical Devices", 2018, International Medical Devices Regulators Forum. Available at: https://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-181031-grrp-essential-principles-n47.pdf

5 Conclusions: lessons learnt

The previous section has provided an overview of the normative questions raised by the different steps of a standard machine-learning product development process. We have tried to shed (more) light on the apparently "technical" steps of developing and configuring algorithmic systems. To conclude this piece, we formulate the following general lessons-learnt.

First, it appears to be clear that our preliminary working definition of an "ethical question" does fulfil its mission. The questions identified all refer to situations in which an AI engineer has to evaluate, decide and act in a specific way. We have characterized these situations as raising ethical questions, because we expect AI engineers to be able to make explicit their reasons for choosing option A over other existing options. Thanks to this analysis, we have created the conditions for an exercise of justification and we have identified the main questions for this justification.

Second, our analysis makes clear that this understanding of ethical questions as occasions for justification specifically calls for an engagement with values. To choose option A over other existing options, the AI engineer needs to make explicit—at least to him/herself—the standard he/she uses to decide between said options. This standard can refer not only to his/her own set of values, but also to the values endorsed by a company or, more broadly, by the society in which he/she lives. This directly interrogates the value-based framework within which the engineer acts. Of course, the engineer is found in several ethical frameworks (as an individual, as a citizen, as an employee) and these frameworks can conflict with each other. The first step to address this complex situation is to provide transparency about the different conflicting reasons and values that are at play.

Third, the step-by-step analysis also helps us to better specify the definition of responsibility at stake. In the second section of this paper, this was claimed to be our contribution to scholarship on responsible innovation, especially on the conceptualisation of "responsibility". We have seen different types of responsibility emerging in the case of AI engineers. There is first a traditional understanding of responsibility as being backward-looking, mainly about blameworthiness and liability. This is relevant for all cases in which an AI engineer makes a mistake that is incompatible with his/her expected level of expertise. Furthermore, as soon as non-supervised AI systems become widely used, we need to complement this traditional understanding of responsibility with a better view on its distribution among several actors. It is about distributed responsibility among human beings involved in the chain of creation and use of AI systems, but also with algorithmic entities (Leonelli 2016). Further research, not least with a sociological ambition, remains to be conducted before we will be able to understand how responsibility is distributed in specific real settings.

There is a second understanding of responsibility as accountability. The focus here is put on the requirement for AI engineers to make their choices explicit and to be in a position to justify these choices, at least potentially. The claim is not that every decision should be justified but rather that, theoretically, every decision is justifiable in light of accepted values (within a company/within a society). The capacity of AI engineers to be responsible in the sense of being "accountable" relies upon the quality of the justification offered. It also has a very important public dimension in making clear that the justification should be understandable for other members of a specific community. Third, the AI engineer might be responsible (in the sense of responsibility-as-virtue) in offering exemplarity. This exemplarity concerns AI engineers as groups of professionals, but also as members of a given society. In this perspective, AI engineers are judged according to their capacity to care for others or, in the terms of responsible innovation, to provide "collective stewardship of science and innovation in the present" (Stilgoe et al. 2013: 1570). Specific character traits such as attention to others, rigor, generosity are seen as important elements for AI engineers who want to be able to live up to their responsibility-as-virtue (van de Poel and Sand 2018: 16). This willingness and capacity to care for others is crucial for the capacity to integrate new demands which can arise during the process of data-based work, but also afterwards.

Fourth, our investigation has shown the relevance of freedom defined as non-domination and the importance of instituting anti-domination mechanisms meant to secure non-arbitrary interferences. On one side, the AI engineer is in a position to potentially exercise domination over others, mainly his/her project partners, but through them all of society. Because he/she has expertise in an area which could potentially impact a considerable number of people, he/she has a responsibility to try to protect values as sacred as individual freedom. In this first sense, freedom as non-domination is a public value which most AI engineers should be committed to if they work in a liberal-democratic context or if they personally endorse a liberal-democratic set of values. In that sense, freedom as non-domination might be operationalized through a professional codex which summarizes the terms of the AI engineer's responsibility. The specific sections of the "Ethically aligned design" by the professional association IEEE represent a good example (Systems 2016: 135 ff). The report by the IEEE mentions methods for operationalizing this responsibility (e.g. independent review organization, technical documentation, auditable processes). We have mentioned examples of the types of mechanisms that ought to be put in place as requirements to provide justification for one's choices.

In a second sense, the AI engineers might themselves be dominated by others, mainly the companies employing them. AI engineers might be forced to do projects which they cannot endorse because of the values they think important. In this second sense, freedom as non-domination is about protecting the capacity of AI engineers to do their job in acceptable conditions. Similarly, these reflections might also be integrated into professional codex, not to directly protect the broader public, but to protect AI engineers (and, thereby indirectly, the broader public). This constellation is by no means specific to AI engineers. Medical doctors are caught in the same double-sided situation: they might represent a threat to individual freedom, but they might also be subjected to domination, e.g. by an institution like a hospital. The point of this third lesson learnt is not to argue for the specificity of the AI engineers, but rather to make clear that their situation needs to be addressed in a similar way as in other sensitive professional occupations.

These two points also stress a point vividly discussed in public debates: the lack of diversity within teams of AI engineers.[14] Conceptually, this lack of diversity might negatively impact the requirement to be as explicit and transparent as possible about one's ethical decisions and standards. If the team is homogenous in terms of professional background, professional values, or even broader life curriculum, it is no surprise that the pressure (e.g. peer-to-peer pressure) to make one's positions more explicit is not strong enough. As soon as you bring diversity (and in that sense, heterogeneity) into a situation, the requirement to explicitly align team members on a compatible set of values becomes more important.

Funding Open access funding provided by University of Zurich.

[14] See for instance the work by the think-tank "Diversity in AI" https://diversity.ai

References

Andersen L (2018) Human rights in the age of artificial intelligence. AccessNow, 6–39.
Bellamy R (2011) Republicanism: Non Domination And The Free State. In: Delanty G, Turner SP (eds) Routledge Handbook of contemporary social and political theory. Routledge, UK, pp 130–139
Bellamy R (2019) A republican Europe of States: cosmopolitanism, intergovernmentalism and democracy in the EU. Cambridge University Press, Cambridge
Besson S, Martí JL (eds) (2009) Legal republicanism : national and international perspectives. Oxford University Press, Oxford
Bowker G, Star L (2000) Sorting things out: classification and its consequences. MIT Press, Cambridge
Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn 30(7):1145–1159
Buchanan A (2004) Justice, legitimacy and self-determination: moral foundations for international law. Oxford University Press, Oxford
Calude CS, Longo G (2017) The deluge of spurious correlations in big data. Found Sci 22(3):595–612
de Hoop E, Pols A et al (2016) Limits to responsible innovation. J ResponsInnov 3(2):110–134
Elish MC, Boyd D (2018) Situating methods in the magic of big data and AI. CommunMonogr 85(1):57–80
Fjeld J, Achten N, Hilligoss H, Nagy A, Srikumar M (2020) Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center for Internet & Society, Cambridge
Floridi L (2013) Distributed morality in an information society. Sci Eng Ethics 19(3):727–743
Floridi L, Cowls J (2019) A unified framework of five principles for AI in society. Harvard Data Sci Rev 1(1):1–15
Floridi L, Taddeo M (2016) What is data ethics? Philos Trans R Soc 374:1–5
Gebru T, Morgenstern J, et al (2018) Datasheets for datasets. arXiv, 1–17.
Giuffrida I (2019) Liability for AI decision-making: some legal and ethical considerations. Fordham Law Rev 88(2):439–456
Hagendorff T (2020) The ethics of AI ethics: an evaluation of guidelines. Mind Mach 30(1):99–120
Hoye JM, Monaghan J (2018) Surveillance, freedom and the Republic. Eur J Polit Theory 17(3):343–363
Jimenez-Buedo M (2011) The political uses of some economic ideas: the trade-off between efficiency and equality. Am J Econ Soc 70(4):1029–1052
Jobin A, Ienca M et al (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1(9):389–399
Lazar-Gillard O (2018) Work, domination, and contemporary republicanism, Thesis
Leonelli S (2016) Locating ethics in data science: responsibility and accountability in global and distributed knowledge production systems. Philos Trans R Soc 374:1–7
Mittelstadt BD, Floridi L (2016) The ethics of big data: current and foreseeable issues in biomedical contexts. Sci Eng Ethics 22(2):303–341
Mittelstadt BD, Allo P et al (2016) the ethics of algorithms: mapping the debate. Big Data Soc 3(2):1–21
Moreau S (2010) What Is Discrimination? Philos Public Aff 38(2):143–179
Moreau S (2013) In defense of a liberty-based account of discrimination. In: Hellman D, Moreau S (eds) Philosophical foundations of discrimination law. Oxford University Press, Oxford, pp 71–86
Morley J, Floridi L et al (2019) From what to how: an initial review of publicly available ai ethics tools, methods and research to translate

principles into practices. Sci Eng Ethics. https://doi.org/10.1007/s11948-019-00165-5
Müller VC (2020) Ethics of artificial intelligence and robotics. Stanford Encyclopedia of Philosophy
Newell BC (2014) Technopolicing, surveillance, and citizen oversight: a neorepublican theory of liberty and information control. Govern Inf Q 31(3):421–431
Papernot N, McDaniel P et al (2017) Practical black-box attacks against machine learning. ASIA CCS 2017 Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security
Pellizzoni L (2019) Reconfiguring non-domination: green politics from pre-emption to inoperosity. Crit Rev Int Soc Polit Philos 1–18
Pettit P (1997) Republicanism: a theory of freedom and government. Clarendon Press, Oxford
Pettit P (2008) Republican liberty: three axioms, four theorems. In: Laborde C, Maynor J (eds) Republicanism and political theory. Blackwell, New Jersey, pp 102–132
Pettit P (2010) A republican law of peoples. Eur J Polit Theory 9(1):70–94
Rawls J (1971) A theory of justice. Belknap Press of Harvard University Press, Cambridge
Rawls J (1993) Political liberalism. Columbia University Press, New York
Roberts A (2015) A republican account of the value of privacy. Eur J Polit Theory 14(3):320–344
Saltz JS, Dewar N (2019) Data science ethical considerations: a systematic literature review and proposed project framework. Ethics InfTechnol 21(3):197–208
Schmidhuber J (2016) Deep learning in neural networks: an overview. Neural Netw 61:85–117
Schultz W (2001) The moral conditions of economic efficiency. Cambridge University Press, Cambridge
Simmons AJ (2001) Justification and legitimacy: essays on rights and obligations. Cambridge University Press, Cambridge
Stahl BC, Wright D (2018) Ethics and privacy in AI and big data: implementing responsible research and innovation. IEEE Secur Priv 16(3):26–33
Staveren Iv (2007) The ethics of efficiency. SCEME Working Papers: Advances in Economic Methodology 18
Stilgoe J, Owen R et al (2013) Developing a framework for responsible innovation. Res Policy Elsevier 42(9):1568–1580
The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2016), Ethically aligned design
Taylor L, Purtova N (2019) What is responsible and sustainable data science? Big Data Soc 6(2):1–6
Timmermans J, Blok V (2018) A critical hermeneutic reflection on the paradigm-level assumptions underlying responsible innovation. Synthese 1–32
Valentini L (2011) Justice in a globalized world: a normative framework. Oxford University Press, Oxford
van de Poel I, Sand M (2018) Varieties of responsibility: two problems of responsible innovation. Synthese 1–19
Veale M, Binns R (2017) Fairer machine learning in the real world: mitigating discrimination without collecting sensitive data. Big Data Soc 4(2):1–17
Verbeek P (2008) Morality in design: design ethics and the morality of technological artifacts. In: Vermaas PE, Kroes P, Light A, Moore SA (eds) Philosophy and design. Springer, Berlin, pp 91–104
Whittlestone J, Nyrup R et al (2019) The role and limits of principles in ai ethics: towards a focus on tensions. Proceedings of the 2nd AAAI/ACM Conference on AI, Ethics, and Society, pp 195–200
Young IM (2007) Global challenges: war determination and responsibility for justice. Polity Press, Cambridge
Zeng Y, Enmeng L et al (2019) Linking artificial intelligence principles. Proceedings of the AAAI Workshop on Artificial Intelligence Safety, pp 1–4
Zook M, Barocas S et al (2017) Ten simple rules for responsible big data research. PLoS Comput Biol 13(3):1–10