Digital Society (2022) 1: 6
https://doi.org/10.1007/s44206-022-00009-z

ORIGINAL PAPER

Assigning Obligations in AI Regulation: A Discussion of Two Frameworks Proposed By the European Commission

Mattis Jacobs[1] · **Judith Simon**[2]

Received: 15 October 2021 / Accepted: 29 June 2022 / Published online: 30 July 2022

Abstract

The emergence and increasing prevalence of Artificial Intelligence (AI) systems in a growing number of application areas brings about opportunities but also risks for individuals and society as a whole. To minimize the risks associated with AI systems and to mitigate potential harm caused by them, recent policy papers and regulatory proposals discuss obliging developers, deployers, and operators of these systems to avoid certain types of use and features in their design. However, most AI systems are complex socio-technical systems in which control over the system is extensively distributed. In many cases, a multitude of different actors is involved in the purpose setting, data management and data preparation, model development, as well as deployment, use, and refinement of such systems. Therefore, determining sensible addressees for the respective obligations is all but trivial. This article discusses two frameworks for assigning obligations that have been proposed in the European Commission's whitepaper *On Artificial Intelligence—A European approach to excellence and trust* and the proposal for the *Artificial Intelligence Act* respectively. The focus is on whether the frameworks adequately account for the complex constellations of actors that are present in many AI systems and how the various tasks in the process of developing, deploying, and using AI systems, in which threats can arise, are distributed among these actors.

Keywords Artificial Intelligence · Algorithmic decision-making · Governance · Regulation · Artificial Intelligence Act

✉ Mattis Jacobs
jacobs@tu-berlin.de

Judith Simon
judith.simon@uni-hamburg.de

[1] Technische Universität Berlin, Berlin, Germany

[2] Universität Hamburg, Hamburg, Germany

1 Introduction

The emergence and increasing prevalence of AI (Artificial Intelligence) systems in a growing number of application areas brings about opportunities but also risks for individuals and society as a whole. To minimize the risks associated with AI systems and to mitigate potential harm caused by them, recent policy papers and regulatory proposals discuss obliging developers, deployers, and operators of these systems to avoid certain types of use and features in their design that bring about "risks or negative consequences for individuals or the society" (European Commission, 2021c, p. 1) by threatening the realization of ethical values, the consideration of ethical principles, and fundamental rights (Datenethikkommission, 2019; European Commission, 2019, 2020, 2021c; HLEG-AI, 2019).

However, most AI systems are complex socio-technical systems in which control over the system is extensively distributed. In many cases, a multitude of different actors is involved in the purpose setting, data management and data preparation, model development, as well as deployment, use, and refinement of such systems. And, as Barocas and Selbst (2016), Danks and London (2017), and others demonstrate, threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights can manifest during all these tasks. Therefore, determining sensible addressees for the respective obligations is all but trivial.

This article discusses two frameworks for assigning obligations that have been proposed in the European Commission's (EC) 2020 whitepaper *On Artificial Intelligence—A European approach to excellence and trust* (European Commission, 2020) and the EC's proposal for the *Artificial Intelligence Act* (AI Act) (European Commission, 2021c) respectively. The EC's whitepaper *On Artificial Intelligence* proposes a capability-based approach for assigning obligations arguing that "the actor(s) who is (are) best placed to address" the respective issue should be obliged to do so (European Commission, 2020). On the contrary, the AI Act argues that the "majority of all obligations" should fall on the person or body "placing [the AI system] on the market or putting it into service under its own name or trademark" (Veale & Zuiderveen Borgesius, 2021) and thus focuses on rather fixed addressees.

While the two proposals argue that their respective framework for assigning obligations is appropriate (European Commission, 2020, p. 22, 2021c, p. 31), neither of them engages in a comparative analysis or in-depth discussion of both frameworks and their respective advantages and disadvantages. Therefore, the rationale behind the shift from the capability-based reasoning of the EC's whitepaper *On Artificial Intelligence* to the reasoning based on fixed addressees in the AI Act is neither readily evident, nor does it follow from one of the proposals.[1]

Therefore, this article attempts to evaluate both frameworks to assess if the EC's shift from one proposal to the other is normatively appropriate. The focus is on whether the respective frameworks adequately account for the complex

[1] The AI Act explicitly builds on the whitepaper *On Artificial Intelligence* in several passages (European Commission, 2021c, pp. 1, 5, 7–9). However, it does not address the different frameworks regarding assigning obligations.

constellations of actors that are present in many AI systems and how the various tasks in the process of developing, deploying, and operating AI systems, in which threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights can arise, are distributed among these actors.

To do so, Sect. 2 provides an overview of the different tasks that exist in the process of developing, deploying, and operating AI systems and the actors involved in performing these tasks. Section 3 sets out the two frameworks proposed by the EC for assigning obligations to these actors in more detail. Section 4 links the threats posed by AI systems to the various tasks in the process of developing, deploying, and operating AI systems. Based on these analyses, Sect. 5 discusses the merit of the shift from the capability-based framework outlined in the EC's whitepaper *On Artificial Intelligence* to the framework based on fixed addressees outlined in the AI Act. Section 6 concludes by summarizing the article's results and outlining the unresolved challenges. Furthermore, it sets out how further regulation and future research can support addressing these challenges.

2 AI Systems as Complex Socio-technical Systems

The EC defines AI systems in a very broad sense with a relative openness regarding the technical approach of the system and its application context (European Commission, 2020, p. 2, 2021a, p. 1, 2021c, p. 39). However, most of the threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights as well as technical features discussed in current policy papers and regulatory proposals concern machine learning-based AI systems with a narrow scope of application.[2] Here, especially systems that make decisions or provide the basis for decisions (for instance, in the form of predictions or recommendations) that concern individuals or groups are regarded as ethically relevant.[3]

AI systems based on machine learning that make decisions or provide the basis for decisions consist of several components. As Krafft et al. (2020) note, especially algorithms of two types are involved: one algorithm that infers "decision rules from data" and another one that "merely uses these decision rules to score or classify cases." The algorithm of the first type, "the learning method," and "the decision rules generated from it" constitute the core of such ADM systems, whereas the "scoring or classification algorithm, in contrast, is usually rather simple as it merely applies the trained statistical model" (Krafft et al., 2020, p. 3).

To "learn from experience" (Russell & Norvig, 1995, p. 518) requires capabilities to "collect, store, and process digital data […] and to utilize vast data sets to train

[2] Such AI systems based on machine learning need to be distinguished from knowledge-based "expert systems," in which problem-specific knowledge of experts is formalized, allowing to automate rule-based decisions "on narrowly defined tasks" (Russell & Norvig, 1995, p. 255). The statements made in this article about the constellation of actors involved in the development, deployment, and use of AI systems based on machine learning do not necessarily apply to expert systems.

[3] Many arguments and direct quotations in this article refer to these systems exclusively and therefore use the term "algorithmic decision-making system" or "ADM system" instead of "AI system.".

and feed machine learning algorithms that rely upon feedback loops to improve their own performance" (Yeung, 2019, p. 21). Indeed, most of the recent resurge in AI is based not on the novelty of theoretical models—many of which were "theorized and developed decades […] ago"—but on the availability of data and the capability to handle and process it at scale (Keller et al., 2018, pp. 7–8). For instance, as Quan and Sanderson (2018) note, natural language processing (NLP) would "not be possible without millions of human speech samplings, recorded and broken down," provided as training data. The capabilities to handle and process data as necessary became available, not least due to the proliferation of more potent hardware. Keller et al. (2018, pp. 7–8) elaborate that "storage technology is now mature enough to store and shift vast amounts of training data [and that] the development of GPUs for graphics and gaming applications have made massive parallelized computing significantly cheaper than when neural networks were invented."

While some dominant actors, such as Amazon, Google, or Microsoft, and some public actors as intelligence agencies have the capacities to build large-scale AI systems entirely on their own, i.e., without purchasing external expertise, pre-trained models, data, or hardware resources, most actors do not. However, cloud computing platforms make those resources accessible and affordable to the many. Furthermore, as Keller et al. (2018, p. 8) note, "access to open source tools and frameworks for creating AI systems also play a part in the current wave of excitement. Tensorflow, Torch and Spark are examples of open source software libraries which […] have made the creation of AI systems – especially during research and development – significantly easier." Besides offering tools to develop and train AI systems, some providers of machine learning infrastructure also offer pre-trained machine learning models that can be incorporated into applications allowing to "score and classify new content right away" available (Microsoft, 2018). While some actors offer a multitude of services in the fields of hardware access, data preparation, model building, and production, there are also large numbers of specialized actors who only offer one or few services in one of these domains (Dhinakaran, 2020).

Thus, AI systems based on machine learning are often not monolithic applications developed by one actor or group of actors. Instead, they are complex socio-technical systems consisting of various technical components that are potentially developed,[4] managed, and operated by various independent actors or groups of actors. For instance, in the case of an AI system that aims to identify risk factors in patients' health records by detecting patterns learned from patient data, several tasks need to be considered: developing the underlying NLP capabilities, providing and preparing health records as training data, building a model that recognizes patterns based on this data, and using the system to classify or score unknown health records. All these tasks might be addressed by different actors. Moreover, due to business interests or strict data protection regulations related to health records, involved actors could be inclined not to share relevant information about the respective components

[4] Or, in case of data: generated or collected.

of the system they control or engage with. Regarding the data perspective, this challenge in healthcare is described well by Kemppainen et al. (2019).[5][6]

As Sect. 4 describes in detail, threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights can originate in various of these tasks. In some cases, they are even rooted in more than one of them. This raises the question of which of these actors are sensible addressees for obligations in AI regulation.

3 Two Frameworks for Assigning Obligations to Actors in AI Systems

Most recent European policy papers and regulatory proposals targeting AI systems recognize the system's considerable formative power. They view the introduction of AI systems as an opportunity to "stimulate new kinds of innovations that seek to foster ethical values" and to "improve individual flourishing and collective wellbeing" (HLEG-AI, 2019, p. 9). Yet, they also acknowledge the risks that AI systems pose for the realization of ethical values, the consideration of ethical principles, and fundamental rights (European Commission, 2020). Here, recurring motives are risks for the respect for human decisions, self-determination, and agency; control over personal data; non-discrimination and fairness; accountability; technical robustness and safety; the rule of law; welfare systems; and democracy (see Datenethikkommission, 2019; European Commission, 2019, 2020; HLEG-AI, 2019). Therefore, they propagate the creation of a regulatory framework that allows harnessing the potential of AI systems while simultaneously mitigating the risks associated with them. Developing such a framework requires addressing regulatory challenges that are familiar from other contexts as well as regulatory challenges that are specific to AI. For instance, a challenge common to the regulation of AI systems as well as the regulation of many other computer systems is the involvement of many actors in development processes, which makes it difficult to hold individual actors accountable (Nissenbaum, 1994). On the contrary, a challenge specific to AI systems is the continuous learning "'in the wild,' that is, in uncontrolled real-world conditions" after deployment (Vallor & Bekey, 2017, p. 341).

The AI Act is a proposal for the regulation of artificial intelligence introduced by the EC seeking "to lay down harmonised rules for the development, placement on the market and use of AI systems" (Veale & Zuiderveen Borgesius, 2021, p. 2). The whitepaper *On Artificial Intelligence* is a document that sets out policy options on how to achieve "the twin objective of promoting the uptake of AI and of addressing the risks associated with certain uses of this new technology" (European Commission, 2020, p. 1). Both the AI Act and the whitepaper propose a risk-based approach to regulating AI, i.e., to apply different governance measures depending on a risk

[5] The well-publicized case of "Watson Oncology," for example,exhibits many of these characteristics (Ross & Swetlitz, 2018; Strickland, 2019).

[6] Pleasenote that the constellations of actors involved in AI systems are not uniform.Therefore, the features of the outlined case are not generalizable. It servesfor illustrative purposes only.

level assigned to the application based on its application area, features, and purpose. While AI systems that are considered to pose an unacceptable risk are outright prohibited, especially in the case of high-risk applications and limited risk applications,[7] a large proportion of suggested measures take the form of obligations for regulated actors (European Commission, 2020, 2021c; Veale & Zuiderveen Borgesius, 2021).[8] However, given the multitude of actors involved in the development, deployment, and operation of many AI systems, there are different approaches to assigning obligations to these actors.

The EC's whitepaper *On Artificial Intelligence* proposes a capability-based approach to assigning obligations. The whitepaper states that in the EC's view, "each obligation should be addressed to the actor(s) who is (are) best placed to address" the respective issue (European Commission, 2020, p. 22).[9] However, the whitepaper does not outline how to determine which actor involved in an AI system is best placed to address an issue in detail (Borutta et al., 2020, p. 6). It just illustrates the approach by suggesting that "[f]or example, while the developers of AI may be best placed to address risks arising from the development phase, their ability to control risks during the use phase may be more limited [in which case] the deployer should be subject to the relevant obligation" (European Commission, 2020, p. 22). Further elaborations regarding the addressees of obligations only concern the geographic scope of the proposed regulation. They do not further specify by which criteria regulators should determine which actor is best placed to address a specific risk (Borutta et al., 2020; European Commission, 2020, p. 22).[10] In the proposal for the AI Act, the EC departs from this view. It moves away from determining addressees of regulatory measures by evaluating their capability. Instead, it attempts to assign obligations to well-defined and clearly identifiable actors. Here, the focus is on "providers" and, to a lesser degree, "users" as the main addressees of obligations (Veale & Zuiderveen Borgesius, 2021). The EC defines "providers" as "a natural or legal person, public authority, agency or other body that develops an AI system or that has an AI system developed with a view to placing it on the market or putting it into service under its own name or trademark, whether for payment or free

[7] For further information on the risk-based approach and the classifications in the respective proposals, see European Commission (2021c, pp. 3, 6, 13) and European Commission (2020, p. 17).

[8] According to the AI Act, AI systems of two categories are considered high-risk. These are products "already covered by certain Union health and safety harmonisation legislation (such as toys, machinery, lifts, or medical devices)" (Veale & Zuiderveen Borgesius, 2021, p. 9), on the one hand, and AI systems for the use in further specified sensitive areas, on the other hand. Such sensitive areas are, for instance, biometric identification, law enforcement, and the administration of justice and democracy (European Commission, 2021a, 2021c).

[9] The whitepaper excludes questions of (civil) liability from this line of reasoning, arguing that it is not a premature judgement of question concerning "liability to end-users or other parties suffering harm and ensuring effective access to justice, which party should be liable for any damage caused" (European Commission, 2020, p. 22).

[10] Such a criterion could be, for example, bearing the (least) cost for addressing a given risk (least (or cheapest) cost avoider approach, cf. Calabresi, 2008).

of charge,"[11] and "users" as "any natural or legal person, public authority, agency or other body using an AI system under its authority, except where the AI system is used in the course of a personal non-professional activity" (European Commission, 2021c, pp. 39–40). With its approach to assign obligations to fixed addressees, the AI Act circumvents the necessity to engage with the possibly ambiguous setup of competencies and capabilities of actors involved in developing, deploying, and operating AI systems. However, Article 28 of the AI Act defines exceptions to this approach. According to Article 28, distributors, importers, users, and third parties are considered providers under the AI Act if "(a) they place on the market or put into service a high-risk AI system under their name or trademark; (b) they modify the intended purpose of a high-risk AI system already placed on the market or put into service; (c) they make a substantial modification to the high-risk AI system" (European Commission, 2021c, Art. 28).

4 On the Roots of Threats Posed by AI Systems

As stated in Sect. 1, AI systems can pose threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights. This section sets out several types of these threats that AI systems pose and traces their roots back to the various tasks in the process of developing, deploying, and operating AI systems and the actors involved in performing these tasks.[12] The list does not aspire to be exhaustive but merely demonstrates that these threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights can arise in the various tasks outlined in Sect. 2.

4.1 Purpose Setting

Some of the threats that AI systems pose are rooted in the setting of the system's purpose. The most salient issues discussed in the scholarly literature are use cases that are problematic from an ethical perspective, regardless of specific design decisions. First and foremost, these are AI systems for intentionally malign purposes such as "Prioritizing targets for cyber attacks using machine learning," "State use of automated surveillance platforms to suppress dissent" (Brundage et al., 2018), or attempting to mask intentional discrimination with ostensibly objective algorithms (Barocas & Selbst, 2016). Furthermore, while some use cases cannot be classified off-hand as based on malign intent, they nevertheless are posited at least at the edge of moral dubiousness because the prospect of deploying such a system raises severe

[11] The focus on providers can be explained in part by the fact that the AI Act draws heavily on existing European product safety regulation Veale and Zuiderveen Borgesius (2021).

[12] While the list is loosely oriented to the sequence of a development process, this is not meant to suggest that the development processes of most AI applications are linear. Often, prototypes are taken into an early deployment and subsequently evaluated and further refined during their use "in the wild.".

ethical concerns regardless of specific design decisions. Prime examples of such systems are, for instance, lethal autonomous weapons systems (Horowitz, 2016).[13]

The AI Act addresses some of these purposes already by assigning them to an "unacceptable risk" category. Within this category, the AI Act does not assign obligations to specific actors involved in the development, deployment, and use of the system but operates with "outright or qualified prohibitions" for respective applications. The current proposal for the AI Act "contains four prohibited categories, three prohibited in their entirety (two on manipulation, one on social scoring); and the last, 'real-time' and 'remote' biometric identification systems prohibited except for specific law" (Veale & Zuiderveen Borgesius, 2021, p. 3). In these cases, the framework discussed in this article does not apply.

4.2 Data Management and Data Preparation

Many of the most controversially discussed threats posed by AI systems based on machine learning have their roots in data management and data preparation. This is because the data used for training an AI system severely impacts how it determines its decisions later on. As Barocas and Selbst (2016) explain, "the data that function as examples […] train the model to behave in a certain way." Data preparation in this context "involves preparing, labeling, and cleaning the data to be used for models" (Dhinakaran, 2020). The main issues that can arise here can be divided into four categories: (1) the use of inaccurate data, (2) the use of nonrepresentative and insufficient data, (3) data containing pre-existing societal biases, and (4) the use of unsecured, protected data.

In many cases in which inaccurate data leads to AI systems posing threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights, inaccurate labeling is at the core of the problem. According to Barocas and Selbst (2016), "labeling examples is the process by which the training data is manually assigned class labels" and further that "the labels applied to the training data must serve as ground truth" for the system. Thus, inaccurate labels lead to a skewed ground truth. For instance, in the field of AI systems for medical image classification, "[i]mage labels are annotations performed by medical experts such as radiologists" (Willemink et al., 2020). Inaccurate training data (or, more specifically, inaccurate labels in the training data) here come about if medical experts categorize and annotate images incorrectly. Inaccurate data can also be deliberately injected into training data to manipulate an AI system's decisions. For instance, in recommender systems, inaccurate recommendations (e.g., fake product recommendations) made by users can shift the recommendations that a system provides. As Milano et al. (2020) note, "providing inaccurate or irrelevant recommendations directly harms a user by reducing the utility that they derive from the recommended option." Other forms of harm can occur if attackers manipulate the training data of AI systems that are applied in other domains.

[13] For a vivid illustration of the issue of morally problematic use cases and the importance of ethically sound purpose setting, see Keyes et al. (2019).

Besides inaccurate data, nonrepresentative data and data containing pre-existing societal biases (Friedman & Nissenbaum, 1996) play a crucial role in the threats that AI systems pose, and that can be attributed to the preparation and management of data. Many of the threats that are discussed under the heading "algorithmic bias" fall in this category. Nonrepresentative data here refers to data sets that omit to sufficiently take certain groups into account. For instance, image recognition software, in many cases, is trained with pictures of predominantly light-skinned persons. In an illustrative case, the error rates for identifying individuals of an image recognition tool developed by Amazon differed extensively between population groups, particularly lighter- and darker-skinned individuals. It mislabeled especially darker-skinned women disproportionately often (Arbel, 2019). From an ethical perspective, the problem is exacerbated by that "[e]rrors of this sort may befall historically disadvantaged groups at higher rates because they are less involved in the formal economy and its data-generating activities, have unequal access to and relatively less fluency in the technology necessary to engage online, or are less profitable customers or important constituents and therefore less interesting as targets of observation" (Barocas & Selbst, 2016). Furthermore, they conclude that this does not only affect the "quality of individual records of members of these groups be poorer as a consequence, but these groups as a whole will also be less well represented in datasets, skewing conclusions that may be drawn from an analysis of the data."

The same issues can arise even if the data is representative, i.e., reflects the overall population, but is insufficient in its extent regarding a certain group. For instance, from a technical perspective, the error of image recognition software would deliver equally problematic results if the training data was representative of the overall population, but the sample of persons with a given skin color would be minuscule in the overall population. It can be expected that the error rate in recognition of images of persons with that skin color would be higher than in other groups, not because it is underrepresented in the training data in that it "deviates from the actual population statistics" (Danks & London, 2017), but that that there is insufficient training data for a certain subset of the population to achieve high-quality outcomes, leading to differential treatment of different population groups.

Moreover, even if the training data is accurate, the data sample is representative of the overall population, and there are sufficient datasets concerning all relevant subgroups, it can still be skewed regarding a moral standard, leading to biased decisions if used as training data for AI systems. Such issues arise "if particular groups ([e.g.,] based on race, religion, ethnicity etc.) have historically suffered disadvantage" (Yeung, 2019, p. 32), and this fact is reflected in the data. For instance, several cases show that AI systems used in hiring processes often disfavored women and racial minorities even if the applicants from these groups had "credentials otherwise equal to other applicants" because the training data was based on historical hiring practices and mirrored existing discrimination (Barocas & Selbst, 2016; see also Dastin, 2018; Lowry & Macpherson, 1988; Yeung, 2019). In these cases, AI systems replicated and reinforced historical biases and perpetuated "injustice against disadvantaged groups and associated stereotypes and stigmatization" (Yeung, 2019, pp. 32–33), even though the data was neither incorrect nor misrepresenting the status quo. Instead, the underlying problem in these cases is that the "relevant moral

standard"—the equal or fair treatment of women and racial minorities—"is different from the current empirical facts" (Danks & London, 2017, p. 4693).

Additionally, making decisions on individual human beings requires working with some form of personal data. Here, especially personally identifiable information can lead to privacy issues. Moreover, the prioritization, classification, association, or filtering of individuals by AI systems can create sensitive insights, even if originally less sensitive data is used as an input. In such cases, "[p]ersonal harms emerge from the inappropriate inclusion and predictive analysis of an individual's personal data" (Crawford & Schultz, 2014, p. 94). In a prominently discussed case, the retail chain *Target* used customer data to make predictions on whether customers were pregnant. It then forwarded this information to marketers to target the respective customers with relevant products, even in cases where the customers did not announce their pregnancy publicly yet (Duhigg, 2012; Hill, 2012). In essence, Target's actions "resulted in the unauthorized disclosure of personal information" (Crawford & Schultz, 2014). Thus, AI systems bring about novel privacy-related issues, such as the "predictive privacy harms" (Crawford & Schultz, 2014) from the Target case.

4.3 Model Development

Also, activities during the development of an AI system's decision model can cause threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights. While many process steps are associated with model development (see Dhinakaran, 2020), most of the threats discussed in the pertinent literature concern the setting of target variables or feature selection.

The setting of target variables is partly determined by the purpose of the AI Application. Hence, the threats posed by AI systems outlined under "Purpose Setting" can manifest here if problematic purposes are operationalized as target variables in model development. However, target variables are not only determined by the purpose of an application. Target variables also contain further requirements to the output of a system that are of ethical relevance. Among these are, for instance, fairness metrics that determine how and to what degree fairness is considered in the differential treatment of various groups. As Binns (2017) points out, there are various metrics for, e.g., fairness, "including; 'accuracy equity', which considers the overall accuracy of a predictive model for each group; 'conditional accuracy equity', which considers the accuracy of a predictive model for each group, conditional on their predicted class; 'equality of opportunity', which considers whether each group is equally likely to be predicted a desirable outcome given the actual base rates for that group; and 'disparate mistreatment', a corollary which considers differences in false positive rates between groups." To complicate matters, the different measures are often "mathematically impossible to satisfy simultaneously except in rare and contrived circumstances, and therefore hard choices between fairness metrics must be made before the technical work of detecting and mitigating unfairness can proceed" (Binns, 2017; see also

Kleinberg et al., 2016). Failure to recognize and act on such concerns during the model development thus can cause severe issues concerning fairness, other ethical values, principles, or fundamental rights.

The process of "feature selection" refers to the process of making choices about what attributes in data sets to observe and subsequently fold into an analysis (Barocas & Selbst, 2016). It can pose threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights if features concern morally or legally sensitive attributes and make them a determining factor in a system's decision-making model. This can either manifest in the selection of a feature that directly represents sensitive attributes as race or gender or via proxy information that serve as a placeholder for such attributes (Danks & London, 2017, p. 4696). As Crawford and Schultz (2014, p. 100) note, by using proxies, AI systems can "circumvent anti-discrimination enforcement mechanisms by isolating correlative attributes that they can use as a proxy" for protected attributes.

Moreover, if data that causally relates to relevant variables is challenging to obtain, models can take correlating data into account as a proxy for unavailable data. This creates the risk of sensitive data being used where it "might be capable of serving as an informational proxy for a morally unproblematic, though hard to measure, variable or feature" (Danks & London, 2017, p. 4696). In some cases, these variables "have a very high degree of predictive value ([i.e.,] statistical relevance)" (Yeung, 2019, p. 27). For this reason, using sensitive attributes as features can be appealing to users even if they are morally problematic. However, individuals that AI systems make decisions on still have a "legitimate interest in not being evaluated and assessed based on considerations that are not causally relevant to the decision" (Yeung, 2019). As a result, the interests of operators and those affected by AI systems may differ extensively, leading to conflicts between different stakeholder groups.

Furthermore, as Barocas and Selbst (2016) point out, many cases discussed in pertinent literature suggest that model developers often settle for proxies which serve as a "highly imperfect basis upon which to predict" other features of an individual that are causally relevant for a decision. Prominently discussed cases are, for instance, the use of skin color as a proxy for the likelihood of an individual having a criminal record (Barocas & Selbst, 2016; Strahilevitz, 2008) and gender as a proxy for traits that correlate with job performance (Danks & London, 2017).

Other threats posed by feature selection arise because, in many contexts, not all groups are equally represented in the set of selected features. Barocas and Selbst (2016) elaborate: "Members of protected classes may find that they are subject to systematically less accurate classifications or predictions because the details necessary to achieve equally accurate determinations reside at a level of granularity and coverage that the selected features fail to achieve."

While the sources of the discussed issues are already rooted in the training data, feature selection can fail to take these issues of the training data into account. If the selection of problematic features is unavoidable, the model can integrate mechanisms that counter or offset adverse outcomes of an AI system's decisions. Here, fairness metrics can also play a role here.

4.4 Deployment, Use, and Refinement

Lastly, threats that AI systems pose can also be rooted in the way they are deployed, used, and refined. Regarding AI systems' deployment, especially its embedding in its socio-technical environment can be a cause of concern. In the technical domain, poor technology-environment design can bring about malfunctions, leading to erroneous decision-making with potentially severe consequences for affected individuals. Similarly, the deployment of an AI system into a social environment with characteristics differing from the ones assumed during its development can cause threats to individuals or society. For instance, Friedman and Nissenbaum (1996) observe that a mismatch between users and system design can occur if "the population using the system differs on some significant dimension such as expertise or cultural background from the population assumed as users in the design" and lead to biased system behavior. This holds true also for AI systems. Recalling the threats that can arise in data preparation and data management illustrates why: while a training data set might not be biased relative to a standard of statistical distribution in one context, it might be in another (Danks & London, 2017, pp. 4692–4693). Therefore, the threats described above, especially the ones stemming from regard to data preparation and data management, can unexpectedly occur if a system is deployed in a context that it was not designed for. Control mechanisms put in place for the intended context of use may fail in a different context.

Moreover, the continuous (re-)development, refinement, and feedback loops between users, developers, and the system give room to novel types of threats. A continuous expansion of training data and, building on that, a constantly evolving decision-making model require a continuous evaluation of the ethical soundness of AI systems over time. Even if an AI system is considered unobjectionable or harmless at one point in time, an evolvement of the system can lead to model instability and performance degradation (Cheatham et al., 2019).

Especially in security-related contexts, "asymmetric feedback" can be a source of performance degradation in systems that integrate continuous learning (O'Neil, 2016). Asymmetric feedback emerges if the setting that a system is placed in only allows for unilateral feedback. For instance, as Zweig et al. (2018, p. 193) note, "a criminal offender who is not released on bail on the recommendation of an ADM system has no way to prove that he would not have recidivated." In the case of a binary decision like this, the system thus only can get feedback on one type of decision and only learn from one type of mistake, leading to over-specialization in one direction while not recognizing and reacting to mistakes in the other direction (Zweig et al., 2018).

Another way cause for performance degradation of AI systems is learning from interactions with human actors that—intentionally or unintentionally—feed the system problematic input. An often-cited case illustrating this issue is the chatbot "Tay," released by Microsoft in 2016 to be shut down after only one day, because the model "quickly turned offensive and abusive after interacting with Twitter users" (Neff & Nagy, 2016, p. 4921). While the abusive behavior of the bot did not directly stem from the developer's actions—it "echoed the racism and harassment that was fed into it" (Neff & Nagy, 2016) by social media users—the developers were accused

of being accountable for not considering the possibility of such a performance degradation and putting "in place additional safeguards and testing procedures" (Wolf et al., 2017).

However, while the intervention of human actors can be necessary to deal with erroneous or biased AI systems, and overriding or intervening in decisions can be necessary, it can also be a source of further bias. As Cheatham et al. (2019) note, "human judgment can also prove faulty in overriding system results," leading to a biased or otherwise unethical decision and, potentially, feeding these decisions as new input data into the system. This, in turn, can result in future replication, i.e., similar—and similarly problematic—decisions made by the AI system.

Conversely, "automation bias" (cf. Skitka et al., 2000)—a tendency to trust or rely on technical artifacts to a higher degree than is warranted—can also be problematic as it leads to operators of automated systems paying "insufficient attention to monitoring the process and to verifying the outputs of the system" (Simon et al., 2020, pp. 12–13). This issue can be exacerbated if operators receive insufficient training to adequately assess an AI system's output as well as its reliability and do not "recognize when systems should be overruled" (Cheatham et al., 2019).

5 Discussion

Based on the analyses in the previous sections, this section discusses the merit of the shift from the capability-based framework outlined in the EC's whitepaper *On Artificial Intelligence* to the framework based on fixed addressees outlined in the AI Act. The evaluation of the frameworks rests on determining to what extent they are able to deal with three challenges that are derived from the findings in previous sections.

5.1 Ambiguity Regarding Which Actors Are Best Placed to Address Risks or Negative Consequences

It is inherently ambiguous which actor is best placed to address threats posed by AI systems. The EC's proposal that "each obligation should be addressed to the actor(s) who is (are) best placed to address any potential risks" (European Commission, 2020) therefore cannot be translated into practice straightforwardly. This is the case even if not only the risk or negative consequence itself but also its root of a threat posed by an AI system is in plain view. This is because identifying the root of a threat does not directly provide any information on who is best placed to address it. If, for instance, an AI system has proven to be biased and problematic features of the training data have been identified as the source of the issue, it can be resolved within the remit of different actors.

Firstly, the actors engaged in data preparation and data management can modify, replace, or delete data points in the training data that—in their aggregate—have shown to be biased. Secondly, the actors engaged in model development "can use a bias in the algorithmic processing to offset or correct for the data bias, thereby yielding an overall unbiased system" (Danks & London, 2017, p. 4695). Thirdly, the deployment

and use of an AI system can be adjusted by "restrict[ing] the scope of operation for the system in question so that there is no longer a mismatch in system performance and task demands" or, in case of decision support systems, by the user "deliberately employ[ing] a compensatory bias" instead of taking action "solely on the basis of the algorithm output" (Danks & London, 2017, p. 4695). Thus, in a capability-based framework, this ambiguity creates the need for regulatory bodies to further specify which actors should be addressed. Approaches such as the cheapest cost avoider principle could reduce this ambiguity by providing clear criteria. However, such a principle can only be applied to individual cases and does not provide generalizable rules for assigning obligations.

On the contrary, the AI Act's framework appears to be appropriate to address the challenge of assigning obligations to actors without regulatory gaps arising due to ambiguous addressees of obligations. By disentangling obligations from capability or judgments on who is "best placed," the proposal for the AI Act allows assigning obligations without engaging with the actor constellations in individual AI systems and the respective actor's capabilities. Yet, while the AI Act's approach results in clearer attribution of obligations, these are not strictly linked to the actual causes and solutions of the respective problems. Therefore, the actors addressed by the AI Act need to establish this link by identifying actors within the respective system that are capable of, e.g., providing documentation or securing ethically relevant technical features of the system and ensuring that these actors support them in fulfilling their obligations. Therefore, the question of which actor is capable of or best placed to address a threat posed by an AI system is not irrelevant in the framework underlying the AI Act. Addressing this issue is merely delegated from the regulatory authority to providers of AI systems.

However, the AI Act recognizes that providers cannot fulfill their obligations under some circumstances. As noted in Sect. 3 of this article, Article 28 determines that if distributors, importers, users, or third parties modify the intended purpose of a high-risk AI system or make substantial modifications to it, they take the role of the provider of that AI system henceforth. All obligations of the original provider are transferred to them (European Commission, 2021c, Art. 28). Here, the AI Act does not strictly follow the framework of fixed addressees but engages in redefining roles and reassigning obligations based on specific actions by involved actors. This deviation from assigning obligations based on a framework based on fixed addressees poses similar challenges as the challenges for the capability-based framework mentioned above: ambiguities arise, which are hard to address due to the complexity and lack of uniformity of AI systems. Smuha et al. (2021, p. 28) raise the question if there are "cases in which a user may legitimately 'misuse' a particular AI system to protect fundamental rights (could the user then change the intended purpose of an AI system without incurring the obligations of a provider under Article 28) [and, if so] who decides these thresholds?" Furthermore, as Ebers et al. (2021, p. 597) note, in the case of AI systems that are "used for many different purposes (general-use AI systems), there may be circumstances where such an AI technology gets integrated into a high-risk system, without the provider having any or only limited influence over the compliance obligations of high-risk AI systems." Here, the question arises if such an integration is considered misuse or in accordance with the purpose of the

system.[14] These ambiguities can be illustrated revisiting the example of applying a multi-purpose NLP system in a healthcare setting introduced in Sect. 2. Which obligations the AI Act assigns to the provider of the system and which it assigns to the user of the system depends on various factors. First, while this circumstance has been criticized (see, e.g., Ebers et al., 2021, p. 594), applying AI systems in sensitive contexts such as healthcare does not automatically qualify this system as a high-risk system. Therefore, the obligations to providers, users, and other actors defined in chapter 2 of the AI Act do not apply by default. However, if the purpose of the system meets the criteria defined in Annex II and Annex III of the AI Act (see Sect. 3), the AI Act assigns additional obligations for high-risk AI systems to the respective actors. For instance, if the AI system is a medical device, it would meet the criteria defined in Annex II. If this is the case, by default, the majority of obligations falls on the provider, and only a few requirements are assigned to the user (e.g., ensuring "that input data is relevant in view of the intended purpose of the high-risk AI system" and monitoring "the operation of the high-risk AI system on the basis of the instructions of use") (European Commission, 2021c, Art. 29). However, two further factors determine this attribution of obligations. On the one hand, in the case of healthcare data, providers will often not have access to the relevant data (e.g., patient records) to fulfill their obligations (Kemppainen et al., 2019). Based on the distribution of control over the data between user and provider, the obligation to "ensure that input data is relevant in view of the intended purpose of the high-risk AI system" is assigned to one or the other (European Commission, 2021c, Art. 29). On the other hand, the question of whether integrating a multi-purpose AI system in a high-risk application is to be considered a modification of the purpose of the system or a misuse of the system remains relevant. As mentioned above, if either is the case, the entirety of obligations originally assigned to the provider would be instead assigned to the user (European Commission, 2021c, Art. 28).

5.2 The Insufficient Informational Basis for Addressing Threats Posed by AI Systems

To address threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights posed by AI systems, involved actors need an informational basis to do so. For instance, if model developers are supposed to decide on whether to integrate a compensatory bias in an AI system's decision model to account for an initial bias in the training data, they need a solid informational basis provided by the actors responsible for data collection, preparation, and management on relevant features of the respective datasets. Thus, the feasibility of meeting obligations is often dependent on receiving information from actors who are better placed to assess features of technical components, use cases, and application

[14] In its reply to the AI Act, Google makes a similar argument claiming that the rules for shifting obligations from providers to other involved actors lack clarity and that "companies will be forced to take a conservative position, imposing a significant chilling effect on the release of general-use APIs and OSS until the issue is resolved in the courts" (Google, 2021, p. 4).

contexts (Digital Europe, 2021, p. 5). However, the more independent actors are involved in developing, deploying, and operating an AI system, the less likely it is that the necessary information transfer will occur.

Moreover, recent trends in the development of AI systems stand in the way of the necessary disclosure and consideration of information flowing in both directions: from the collection and management of training data, over model development, to deployment and use as well as the other way around. This flow of information is often limited due to the involved actors' business interests. While algorithms and data are non-rivalrous goods, openly sharing them can still lead to a "loss of advantage over competitors" (Keller et al., 2018, p. 11). If little information is shared about the training data or algorithms used to train an AI system's decision model, it becomes increasingly difficult for actors who integrate the model into end-user applications or actors who deploy and operate them to react adequately to ethically problematic properties of the system.

Conversely, actors who are involved in managing training data or developing decision models in many cases lack or cannot fully account for information on societal features of the context of use of an end-user application in order to adjust the AI system to this context and, e.g., avoid bias (see Friedman & Nissenbaum, 1996). This is because the context of use is often deliberately not fully defined to allow components of a system to be used in more than one application context. Pre-trained models are a prime example of this. Empirical research has already linked existing pre-trained models to bias (Webster et al., 2020) as well as security-related issues (Gu et al., 2019). It needs to be emphasized, however, that the problem is not limited to the use of pre-trained models but is more general.

While the proposal for the AI Act requires providers of AI systems of "high-risk" category to "establish a sound quality management system, ensure the accomplishment of the required conformity assessment procedure, draw up the relevant documentation and establish a robust post-market monitoring system" (European Commission, 2021c, p. 31), it refrains from defining the scope of information obligations among the other actors involved in the system. The information and documentation obligations are directed primarily at providers. Nevertheless, establishing a structure for sharing information among the involved stakeholders is, in practice, a prerequisite for providers to fulfill the AI Act's obligations. For instance, if providers of the system are not themselves directly in charge of data management, data preparation, or model development, they can hardly establish a quality management system (European Commission, 2021c, Art. 17), provide technical documentation (European Commission, 2021c, Art. 11), or ensure adequate data governance (European Commission, 2021c, Art. 10) without entering into an intensive exchange of information with the actors in charge of the respective tasks. This is because, as Digital Europe (2021, p. 5) notes, questions such as "what is relevant and representative at a given time when developing the AI system will vary based on the use case," and in many cases, providers need to rely on users to assess use cases. Thus, to ensure that it is feasible to fulfill their obligations, providers need to ensure that development practices (e.g., the use of pre-trained models) or business practices (e.g., restricting access to resources such as data or algorithms) do not obstruct necessary information sharing between the involved actors.

The capability-based framework—as presented in the whitepaper *On Artificial Intelligence*—does not define any requirements for information sharing among involved actors or between the involved actors and regulatory authorities. Nevertheless, the approach is affected by insufficient data-sharing practices because whether or not a given actor possesses access to information determines if it is well placed to address a threat posed by an AI system. For instance, actors in model development could be well posited to address a given bias that results from skewed training data by integrating a compensatory bias. However, one could only meaningfully describe these actors as well-positioned to address the respective issue if information about the relevant features of the training data is shared with them. Actors controlling data could refrain from sharing such information, for instance, based on business interests (Keller et al., 2018) or, as described in the healthcare case, due to data protection regulation. Therefore, a capability-based approach applied in practice needs to spell out either an information-sharing ruleset or it needs to evaluate actual information-sharing practices on a case-by-case basis to determine which actor is best placed to address an issue. Both approaches would involve a major regulatory burden if applied by regulatory authorities to prevent regulatory gaps due to ambiguous addressees of obligations. The elaborations in the whitepaper *On Artificial Intelligence* do not engage with this issue and therefore leave this central issue unaddressed.

Thus, the framework introduced in the proposal for the AI Act based on fixed addressees avoids further central problems that occur in the capability-based framework proposed in the whitepaper *On Artificial Intelligence*. While it does not provide a clear path to how information sharing should be structured among the involved actors, it establishes a well-defined, clearly identifiable actor—the provider—who is the main addressee of most obligations. Thereby, the proposal for the AI Act makes it possible to delegate the micromanagement of information sharing without allowing regulatory gaps due to ambiguous addressees of obligations to arise.

5.3 Systemic, Cumulative Effects of AI Systems

Finally, some threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights posed by AI systems are not rooted in one specific application, let alone individual actions during the process of developing, deploying, and operating it. Most importantly, these are issues concerning the "cumulative effect from widespread and systematic reliance on algorithmic decision-making [which] could erode and destabilize the core constitutional, moral, political, and social fabric upon which liberal democratic societies rest and upon which our shared values are rooted" (Yeung, 2019, pp. 41–42). For instance, the use of AI systems by social networks has been criticized for increasing political polarization (see, e.g., Hao, 2021), whereas the use of AI systems by health insurances is being critically examined regarding whether individually justified differentiations can lead to a loss of solidarity in society (see, e.g., Datenethikkommission, 2019). Furthermore, the capacity of AI systems to make inferences about individual's intimate aspects of life and decisions that determine their future based on data that individuals produce

by going on with their everyday life "may have a corrosive chilling effect on our capacity to exercise our human rights and fundamental freedoms" (Yeung, 2019, p. 36).[15][16]

The question of who bears responsibility for the systemic, cumulative effects of the widespread use of AI systems for society is not addressed by either framework. In the case of the AI Act, this might be caused by that it is modeled after EU product law, especially regulation concerning product safety (Veale & Zuiderveen Borgesius, 2021, p. 3). Thus, irrespective of which framework is applied, further policy considerations are required to address threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights that fall outside the scope of the respective framework.

6 Conclusion

Currently, a broad range of academic literature and numerous (European) policy papers, as well as regulatory proposals, discuss how AI systems can and should be regulated. Within this discourse, one key challenge that is discussed is how to determine which of the actors involved in developing, deploying, and operating AI systems should be obliged to address threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights such systems can pose. The present article contributes to this discourse by discussing the appropriateness of the frameworks to assign such obligations to involved actors proposed by the EC in the whitepaper *On Artificial Intelligence* and the proposal for the AI Act, respectively.

To do so, this article first provides an overview of the different tasks that exist in the process of developing, deploying, and operating AI systems and the actors involved in performing these tasks (Sect. 2). Then, it introduces the frameworks to assign obligations outlined in the EC's whitepaper *On Artificial Intelligence* and the AI Act, respectively (Sect. 3). Subsequently, the article links the threats posed by AI systems to the various tasks in the process of developing, deploying, and operating AI systems (Sect. 4). Finally, it discusses challenges for the two frameworks and the merit of the shift from the capability-based framework outlined in the EC's whitepaper *On Artificial Intelligence* to the framework based on fixed addressees outlined in the AI Act (Sect. 5).

The capability-based framework—targeting actors who are "best placed" to address threats—suffers from the fact that one threat, for instance, bias against a protected group, can have various roots and paths to resolve. Therefore, which involved actor is most capable of or best placed to address a threat posed by an AI system remains highly subjective. Furthermore, threats posed by AI systems often do not emerge as a result of one actor's activity but due to an insufficient flow of

[15] For a more exhaustive discussion, see Yeung (2019).

[16] Systemic, cumulative effects of AI systems are not listed in Sect. 4, as Sect. 4 attempts to trace the roots of threats back to the various tasks in the process of developing, deploying, and operating an individual AI system.

information among several involved actors. An actor who, in theory, is capable of addressing a threat posed by an AI system does, in practice, often not have access to information that allows it to recognize this circumstance and react adequately to it. Both concerns can lead to a diffusion of responsibility as they give involved actors leeway to reject obligations to address a specific threat assigned to them. Therefore, if the capability-based framework for regulating AI systems is applied, it would require extensive micromanagement of regulatory authorities in assigning obligations. As there are no more tangible proposals dealing with the challenges identified in Sect. 5, the capability-based framework, as outlined in the whitepaper *On Artificial Intelligence*, does not provide an appropriate path to assigning obligations to actors involved in developing, deploying, and operating AI systems. While providing additional criteria for what "best placed" means in practice could reduce ambiguity, it would not reduce the need for extensive micromanagement by regulatory authorities.

The framework based on fixed addressees outlined in the proposal for the AI Act is less affected by both challenges: the ambiguity regarding which actor is best placed to address a given threat posed by an AI system, and the insufficient informational basis for actors to address such threats. By obliging actors who place a product on the market to ensure that it meets a given set of criteria, this framework avoids the necessity for regulators to engage in-depth with the actor constellations within a given AI system, as the obligations are simply assigned to a well-defined and clearly identifiable actor. The responsibility to gather the necessary information and ensure certain system properties, even in ambiguous setups, is delegated to these actors. By relying on this framework, the proposal for the AI Act resolves some of the core problems of earlier policy papers and regulatory proposals and is thus more appropriate in this crucial respect.

However, in cases in which providers are not in charge of the whole process of developing, deploying, and operating AI systems, they might need to rely on additional actors involved in an AI system to cooperate to fulfill obligations such as establishing a quality management system (European Commission, 2021c, Art. 17), providing technical documentation (European Commission, 2021c, Art. 11), or ensuring adequate data governance (European Commission, 2021c, Art. 10). To do so, providers need to identify actors capable of providing information and carrying out modification and oblige them to do so. In practice, this could result in that it is unfeasible for providers to cooperate with actors who rely on engineering and business practices that are common in the development, deployment, and operation of AI systems but are incompatible with such requirements. Affected could be, for instance, the use of pre-trained models or business practices built around disclosing little information about the training data.

Yet, since the proposal for the AI Act categorizes the AI systems in question as posing a high risk or bringing about negative consequences for individuals or the society, the higher weighting of ensuring that no regulatory gaps arise over specific engineering and business practices is only consequential. Accordingly, the EC's shift from a framework focused on capability to a framework focused on fixed addressees is appropriate in that it ensures that there is a well-defined and identifiable actor that obligations can be assigned to.

Nonetheless, just like capability-based approaches for assigning obligations, the AI Act's framework based on fixed addressees aims at addressing threats posed by individual AI systems. It does not consider cumulative effects of AI systems resulting from a "widespread and systematic reliance on algorithmic decision-making" (Yeung, 2019). Thus, irrespective of the selected framework, this issue must be addressed through alternative policy considerations. Moreover, it is important to note that how a regulatory proposal deals with the challenge of determining sensible addressees for the respective obligations is by no means the only factor that determines its appropriateness. Both the proposal for the AI Act as well as the whitepaper *On Artificial Intelligence* have been criticized for other reasons, as, e.g., the appropriateness of the risk categorization, the appropriateness of risk-based approach in general, a wide scope for interpretation, an extensive bureaucratic burden, relying heavily on (self-) conformity assessments and proportionality assessments, and (further) regulatory blind spots (see, e.g., Borutta et al., 2020; Hoffmann, 2021; Smuha et al., 2021; Veale & Zuiderveen Borgesius, 2021). Therefore, this article should be understood as a contribution to the discourse on the respective proposal's appropriateness regarding the framework for assigning obligations to actors involved in developing, deploying, and operating AI systems and not as an exhaustive assessment of the respective proposals.

Future research and regulation can address some of the outlined issues for assigning obligations in AI regulation. Regarding the capability-based framework, developing criteria for what "best placed" means in practice is crucial. This could help to explore whether a more elaborate form of the framework would be a feasible alternative approach to assigning obligations. The AI Act, however, has to be complemented by further regulation. The EC is already planning to address (civil) liability issues "related to new technologies, including AI systems" that it did not address in the AI Act such as "revisions of the sectoral safety legislation and changes to the liability rules" (European Commission, 2021b, p. 1). Future research should, therefore, investigate to what extent the frameworks and concepts discussed in this article can be transferred to this context. Furthermore, the EC needs to address cumulative effects of AI systems on society with additional regulation.

Author Contribution Mattis Jacobs and Judith Simon contributed to the article's conception. The literature analysis was performed by Mattis Jacobs. The first draft of the manuscript was written by Mattis Jacobs. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. Mattis Jacobs and Judith Simon acknowledge financial support from the German Federal Ministry of Education and Research in the project GOAL–About algorithmic behaviour control and artificial intelligence under the reference 01IS19020.

Data Availability Not applicable, no data, code, or other materials (e.g., images, maps, archival documents, photographs, audio or film recordings, field notes, spreadsheets, interview notes) were generated or analyzed.

Declarations

Ethics Approval Not applicable; conducting the study did not involve research with human participants and/or animals.

Consent to Participate Not applicable; conducting the study did not involve research with human participants.

Competing Interests The authors declare no competing interests.

References

Arbel, T. (2019). *Researchers say Amazon face-detection technology shows bias*. https://abcnews.go.com/Technology/wireStory/researchers-amazon-face-detection-technology-shows-bias-60630589?cid=social_twitter_abcn

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. l. Rev., 104*, 671.

Binns, R. (2017). Fairness in machine learning: Lessons from political philosophy. ArXiv Preprint ArXiv:1712.03586.

Borutta, Y., Haag, M., Hoffmann, H., Kevekordes, J., & Vogt, V. (2020). *Fundamentalkritik des White Papers und des Datenstrategiepapiers der EU-Kommission vom 19. Februar 2020*. https://goal-projekt.de/wp-content/uploads/2020/03/Fundamentalkritik-1.pdf

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., & . . . Amodei, D. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. https://arxiv.org/pdf/1802.07228

Calabresi, G. (2008). *The Cost of Accidents: A Legal and Economic Analysis*. Yale University Press.

Cheatham, B., Javanmardian, K., & Samandari, H. (2019). *Confronting the risks of artificial intelligence*. McKinsey Quarterly. https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/confronting-the-risks-of-artificial-intelligence

Crawford, K., & Schultz, J. (2014). Big data and due process: Toward a framework to redress predictive privacy harms. *Boston College Law Review*, *55*(1), 93–128. https://heinonline.org/HOL/P?h=hein.journals/bclr55&i=93

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. In F. Bacchus & C. Sierra (Eds.), *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (pp. 4691–4697). International joint conferences on artificial intelligence organization. https://doi.org/10.24963/ijcai.2017/654

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Datenethikkommission. (2019). *Gutachten der Datenethikkommission*. https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf

Dhinakaran, A. (2020). *The AI Ecosystem is a MESS: Why is it impossible to understand what AI companies really do?* Towards Data Science. https://towardsdatascience.com/the-ai-ecosystem-is-a-mess-c46bdfbf43e4

Digital Europe. (2021). *DIGITALEUROPE's initial findings on the proposed AI Act.* Digital Europe. https://www.digitaleurope.org/wp/wp-content/uploads/2021/08/DIGITALEUROPEs-initial-findings-on-the-proposed-AI-Act.pdf

Duhigg, C. (2012). *How Companies Learn Your Secrets.* The New York Times. https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Ebers, M., Hoch, V. R. S., Rosenkranz, F., Ruschemeier, H., & Steinrötter, B. (2021). The European Commission's Proposal for an Artificial Intelligence Act—A Critical Assessment by Members of the Robotics and AI Law Society (RAILS). *J*, *4*(4), 589–603. https://doi.org/10.3390/j4040043

European Commission. (2019). *Building trust in human-centric artificial intelligence*. https://ec.europa.eu/digital-single-market/en/news/communication-building-trust-human-centric-artificial-intelligence

European Commission. (2020). *On artificial intelligence - A European approach to excellence and trust: Whitepaper*. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

European Commission. (2021a). *Annexes to the proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative ACTS*. https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_2&format=PDF

European Commission. (2021b). *Commission staff working document impact assessment accompanying the proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts*. https://eur-lex.europa.eu/resource.html?uri=cellar:0694be88-a373-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

European Commission. (2021c). *Proposal for a Regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts*. https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems, 14*(3), 330–347. https://doi.org/10.1145/230538.230561

Google. (2021). *Consultation on the EU AI Act Proposal: Google's submission*. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2662492_en

Gu, T., Liu, K., Dolan-Gavitt, B., & Garg, S. (2019). BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. *IEEE Access, 7*, 47230–47244.https://doi.org/10.1109/ACCESS.2019.2909068

Hao, K. (2021). *The Facebook whistleblower says its algorithms are dangerous. Here's why*. MIT Technology Review.https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/

Hill, K. (2012). *How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did.* Forbes. https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-herfather-did/?sh=20f35caa6668

HLEG-AI. (2019). *Ethics guidelines for trustworthy AI.*

Hoffmann, H. (2021). Regulierung der Künstlichen Intelligenz: Fundamentalkritik am Verordnungsentwurf zur Regulierung der Künstlichen Intelligenz der EU-Kommission vom 21. 4. 2021. *Kommunikation & Recht*, 369–374.

Horowitz, M. C. (2016). The Ethics & Morality of Robotic Warfare: Assessing the Debate over Autonomous Weapons. *Daedalus, 145*(4), 25–36. https://doi.org/10.1162/DAED_a_00409

Keller, J. R., Chauvet, L., Fawcett, J., & Thereaux, O. (2018). *The role of data in AI business models.* Open Data Institute. https://theodi.org/wp-content/uploads/2018/04/376886336-Therole-of-data-in-AI-business-models.pdf

Kemppainen, L., Pikkarainen, M., Hurmelinna-Laukkanen, P., & Reponen, J. (2019). Data Access in Connected Health Innovation: Managerial Orchestration Challenges and Solutions. *Technology Innovation Management Review*, *9*(12), 43–55. https://doi.org/10.22215/timreview/1291

Keyes, O., Hutson, J., & Durbin, M. (2019). A Mulching Proposal. In S. Brewster, G. Fitzpatrick, A. Cox, & V. Kostakos (Eds.), *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–11). ACM. https://doi.org/10.1145/3290607.3310433

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent Trade-Offs in the Fair Determination of Risk Scores*. https://arxiv.org/pdf/1609.05807

Krafft, T. D., Zweig, K. A., & König, P. D. (2020). How to regulate algorithmic decision-making: A framework of regulatory requirements for different applications. *Regulation & Governance, 104*, 671. https://doi.org/10.1111/rego.12369

Lowry, S., & Macpherson, G. (1988). A blot on the profession. *British Medical Journal (clinical Research Ed.), 296*(6623), 657–658. https://doi.org/10.1136/bmj.296.6623.657

Microsoft. (2018). *Pre-trained machine learning models for sentiment analysis and image detection.* Microsoft. https://docs.microsoft.com/en-us/machine-learning-server/install/microsoftml-install-pretrained-models

Milano S, Taddeo M, Floridi L (2020) Ethical aspects of multi-stakeholder recommendation systems The Information Society 1–11. https://doi.org/10.1080/01972243.2020.1832636

Neff, G., & Nagy, P. (2016). Talking to bots: Symbiotic agency and the case of Tay. *International Journal of Communication*.

Nissenbaum, H. (1994). Computing and accountability. *Communications of the ACM, 37*(1), 72–80. https://doi.org/10.1145/175222.175228

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy* (First paperback edition). Broadway Books.

Quan, X. I., & Sanderson, J. (2018). Understanding the Artificial Intelligence Business Ecosystem. *IEEE Engineering Management Review, 46*(4), 22–25. https://doi.org/10.1109/EMR.2018.2882430

Ross, C., & Swetlitz, I. (2018). *IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show*. https://www.statnews.com/wp-content/uploads/2018/09/IBMs-Watson-recommended-unsafe-and-incorrect-cancer-treatments-STAT.pdf

Russell, S. J., & Norvig, P. (1995). *Artificial intelligence: A modern approach*. *Prentice Hall series in artificial intelligence*. Prentice Hall; London : Prentice-Hall International.

Simon, J., Wong, P.-H., & Rieder, G. (2020). Algorithmic bias and the Value Sensitive Design approach. *Internet Policy Review*, *9*(4). https://doi.org/10.14763/2020.4.1534

Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies, 52*(4), 701–717. https://doi.org/10.1006/ijhc.1999.0349

Smuha, N. A., Ahmed-Rengers, E., Harkens, A., Li, W., MacLaren, J., Piselli, R., & Yeung, K. (2021). *How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act*. Advance online publication. https://doi.org/10.2139/ssrn.3899991

Strahilevitz, L. J. (2008). Privacy versus antidiscrimination. *The University of Chicago Law Review, 75*(1), 363–381.

Strickland, E. (2019). IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum, 56*(4), 24–31. https://doi.org/10.1109/MSPEC.2019.8678513

Vallor, S., & Bekey, G. A. (2017). Artificial Intelligence and the Ethics of Self-learning Robots. In P. Lin, R. Jenkins, & K. Abney (Eds.), *Robot ethics 2.0: New challenges in philosophy, law, and society* (pp. 338–353). Oxford University Press.

Veale, M., & Zuiderveen Borgesius, F. (2021). *Demystifying the Draft EU Artificial Intelligence Act*. Preprint, July 2021. Version 1.1. https://doi.org/10.31235/osf.io/38p5f

Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., Pavlick, E., Chen, J., Chi, E., & Petrov, S. (2020). *Measuring and Reducing Gendered Correlations in Pre-trained Models*. http://arxiv.org/pdf/2010.06032v2

Willemink, M. J., Koszek, W. A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L. R., Summers, R. M., Rubin, D. L., & Lungren, M. P. (2020). Preparing Medical Imaging Data for Machine Learning. *Radiology, 295*(1), 4–15. https://doi.org/10.1148/radiol.2020192224

Wolf, M. J., Miller, K. W., & Grodzinsky, F. S. (2017). Why we should have seen that coming: comments on microsoft's tay "Experiment," and Wider Implications. *The ORBIT Journal*, *1*(2), 1–12. https://doi.org/10.29297/orbit.v1i2.49

Yeung, K. (2019). Why Worry about Decision-Making by Machine? In K. Yeung & M. Lodge (Eds.), *Algorithmic regulation* (pp. 21–48).

Zweig, K. A., Wenzelburger, G., & Krafft, T. D. (2018). On Chances and Risks of Security Related Algorithmic Decision Making Systems. *European Journal for Security Research, 3*(2), 181–203. https://doi.org/10.1007/s41125-018-0031-2w