Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

doi:10.3233/FAIA230686

Respect for Autonomy in the Machine Learning Pipeline

Paula SUBÍAS-BELTRÁN [a,b,1], Oriol PUJOL [c], Itziar DE LECUONA [b,d], and Vicent RIBAS RIPOLL [a]

[a] *Eurecat, Centre Tecnològic de Catalunya, Barcelona, Spain*
[b] *Bioethics and Law Observatory - UNESCO Chair in Bioethics, Universitat de Barcelona, Barcelona, Spain*
[c] *Dept. de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Spain*
[d] *Dept. of Medicine, Universitat de Barcelona, Barcelona, Spain*
ORCiD ID: Paula Subías-Beltrán https://orcid.org/0000-0003-1167-1259, Oriol Pujol https://orcid.org/0000-0001-7573-009X, Itziar de Lecuona https://orcid.org/0000-0002-5081-5756, Vicent Ribas Ripoll https://orcid.org/0000-0002-7266-6106

Abstract. The machine learning (ML) community recognizes the potential impact of ML systems on human rights, especially regarding privacy and discrimination. To address these concerns, the community has conducted various studies on fairness, accountability, and transparency in developing and deploying ML systems. Despite these efforts, the importance of autonomy, a fundamental principle underlying many human rights, has often been overlooked. This oversight is concerning as it could jeopardize individuals' decision-making and exercise effective control over their lives, resulting in a violation of their rights. This article examines the principle of autonomy and its significance in the ML pipeline from a transdisciplinary perspective. The authors argue that autonomy remains a theoretical concept and does not translate well into the real use of ML, leading to contradictory outcomes. The absence of an effective approach to integrating autonomy leads to the persistence of disparities.

Keywords. human autonomy, machine learning, ethics of AI, bioethics and human rights

1. Introduction

We face several demands to constantly consume and produce as a result of the current economic and social structure. As people and organizations are concerned about failing to keep up with fast-moving events [2], these pressures create a climate in which notions such as "if we do not do it, someone else will" grow more prominent. Thus, these demands for haste frequently prevent us from effectively considering how our activities affect society and the environment.

[1] Corresponding Author: Paula Subías-Beltrán, paula.subias@eurecat.org.

By streamlining systematic operations, automation plays a critical role in advancing development by freeing up time and resources to be applied to more complicated tasks. By being able to process enormous volumes of data and find insights that might not be obvious through manual analysis alone, Machine Learning (ML) enables us to address more complicated problems. In light of this, ML-based technologies are spreading ubiquitously. These technologies are intrinsically vehicles for the companies that develop them to achieve their goals, such as expanding their reach, boosting user engagement, and so on. Because of their nature, the main objective of these technologies is not the respect of societal values, but it is a requirement that we must demand.

Autonomy is the capacity of individuals to make free and self-determined judgments. The premise is that people require autonomy in order to conduct their life according to their own views, values, and needs, as long as they do not damage others. Autonomy is a core value in the development and promotion of the principles and human rights we embrace as European society. Making decisions regarding privacy and intimacy requires autonomy, and maintaining privacy is crucial to sustaining intimate relationships. These are interconnected and interrelated principles that form the cornerstone of the society in which we live.

The importance of autonomy in algorithmic decision-making is often underappreciated in the Artificial Intelligence (AI)[2] community. Instead, emphasis is being placed on other crucial values, such as fairness and privacy, but autonomy-related issues are not being sufficiently addressed. The topic of autonomy in AI calls for a more philosophical approach, requiring more critical and in-depth views on how to include the value of autonomy in technology. There is still work to be done to fully achieve effective respect for autonomy, despite the fact that advancements have been achieved in this area [5,8,14,22].

To address these concerns, the AI community needs an effective and systematic strategy, one that is accessible to individuals with technical backgrounds and may be less conversant with philosophical frameworks. The innovation of this work is that it addresses the risks posed in each phase of the ML pipeline in a methodical manner, allowing it to be easily transferred to real-world development carried out by AI practitioners. This paper begins by describing the concept of autonomy and stating its importance as a core principle in European society in §2. The full examination of the hazards to autonomy at the various stages of the ML pipeline will then follow in §3. Finally, §4 concludes by summarising the research's findings.

2. Autonomy as a foundational value

Autonomy is a very complex concept that has been debated by different disciplines. The notion that unites them all, though, is self-government and self-determination, which elevates autonomy to the status of a distinguishing quality of free moral actors. People are therefore at their most autonomous when they behave in accordance with their own interests, beliefs, and desires [9]. The disputes around autonomy are frequently brought on by the term autonomy being confused with one specific conception of autonomy rather than a more comprehensive perspective.

[2]In folk language usage, AI and ML are often interchanged, but in this context, we use "AI" to describe the general approach discussed in the public discourse, reserving "ML" for addressing specific real-world implementations.

In bioethics, one of the four basic principles that govern the field is autonomy. This principle is addressed within the decision-making process and is connected to individual responsibility. Beauchamp and Childress, two of the main authors of the reference work that examines the principles of Bioethics [3], claimed that to demonstrate respect for autonomous agents implies "to acknowledge their right to hold views, to make choices, and to take actions based on their personal values and beliefs".

From a liberal philosophical approach, autonomy is seen as an individual right necessary to make one's own decisions or as a psychological ideal of independent thinking and rational self-control [12]. According to this view, individuals should be free to make their own choices, as long as they do not harm others. On a different note, there is the concept of "relational autonomy", which is an umbrella term indicating a variety of linked ideas rather than referring to a single understanding of autonomy. Relational autonomy is grounded in the belief that people are socially constituted and that social relationships and a variety of intersecting social variables, including race, class, and gender, impact people's identities. Thus the focus of this approach is to analyse the effects of the intersubjective and social aspects of selfhood and identity on concepts of individual autonomy as well as moral and political agency [15]. This view is rooted in feminist philosophy and challenges traditional views of autonomy. Autonomy is caricatured in feminist critiques of autonomy as a self-sufficient individualist male who adheres to libertarian doctrine. According to this view, autonomy is created not just by individual decisions but also by cultural and social structures that favor certain groups over others. In line with feminist views, full autonomy can be attained only when society is more egalitarian and everyone has equal access to important resources and opportunities.

These many philosophical viewpoints on autonomy have significant ramifications for how autonomy is interpreted and valued in different scenarios, including AI technologies. It is crucial for AI practitioners to be aware of the importance of autonomy and to engage in critical reflection on how autonomy is addressed during each phase of the ML pipeline. This was backed up by the European Commission's High-Level Expert Group on AI in its well-known Ethics Guidelines for Trustworthy AI [1]. They listed respect for human autonomy as one ethical principle to take into account in the context of AI, despite the fact that they did not even describe autonomy and did not offer any practical advice for making this respect effective.

In this work, we understand autonomy as a principle that enables people to behave independently of others or in connection to their society, in accordance with their own values, beliefs, and desires. In democratic environments, autonomy is considered a vital element of human freedom, and self-determination since it enables people to take part in the political process as educated and engaged citizens [20]. And any widely used technology, such as ML, can have an impact on it. Excessive reliance on ML systems, for example, may reduce the overall level of human skill required, impairing our ability to make educated and independent decisions and, as a result, reducing our autonomy in the medium or long term. ML-based solutions have the potential to improve social justice or exacerbate existing disparities. Therefore, it is the duty of ML practitioners to think about the wider ramifications of their work in order to comprehend the potential effects and lessen any harm. This calls for a dedication to moral standards and a readiness to consider the implications of their job critically. By doing this, ML professionals may strive toward a society in which their solutions promote everyone's rights and principles, such as respect for autonomy. Our study targets ML applications where decisions are

made based on the output produced, which includes both intended designs and situations where there is an excessive reliance on the result. The focus of our research is thus restricted to ML models that are utilized for decision-making. Next, we take a thorough look at how autonomy is affected by each of the major stages of the ML pipeline.

3. Autonomy in the ML pipeline

Our strategy to promote autonomy has been to examine the impact of autonomy within the ML pipeline, with the goal of engaging more technical profiles in this dialogue. We seek to raise awareness of the ethical aspects of AI and encourage AI practitioners to prioritize respect for autonomy in their work by identifying and considering the potential dangers to autonomy at each phase of the ML pipeline. In order to structure and facilitate the discussion around the impact of autonomy within the ML pipeline, we have simplified the pipeline into the following phases, which are connected in the pipeline shown in Figure 1:

- **Design** involves setting expectations for the performance of the model, as well as considering the ethical and social implications of the model's use in practice.
- **Data** entails its collection or creation, and its transformation so that it will be used to train and validate the model. It includes data collection, data cleaning, imputation, preprocessing, data transformation, and supervised annotation, among others.
- **ML model** is primarily concerned with the internal workings of the model, rather than the input and output of the system. It includes the definition of the functional hypothesis space and model architecture, loss function, optimization algorithm, regularization definition, and other kinds of inductive biases as well as dependencies of third-party components and libraries.
- **Utilization** involves not only the technical aspects of deploying the model but also the practical considerations of how the system will be used and monitored over time.

Next, we delve into the identified phases and expand on how ML may threaten the exercise of effective autonomy.

3.1. The world as it is

"The world as it is" serves as the foundational factual starting point for the design and development of ML-based systems. This is because the world as it is provides a source of data that represents current dynamics. However, it is up to us to choose whether to step in to alter present dynamics or to remain silent and let everything play out as it has in the past. We as a species do not have common values and even the parts we do share change over time. The world as it ought to be provides an opportunity to adopt a vision of the world based on our ideals rather than how things are now. It inspires us to think of a time when technology is employed to advance human flourishing rather than to uphold injustice or cause damage. It serves as a reminder that we have the ability to mold our environment and technological advancements to match our goals and ideals. An important design choice depends on two different perspectives, one where we presume

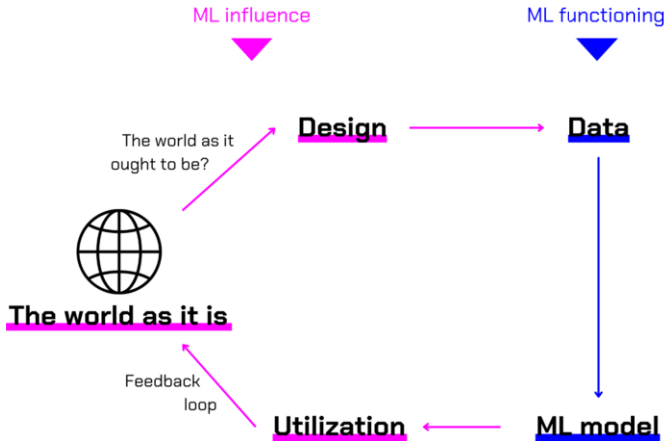

Figure 1. Phases that form the ML pipeline and their impact on the world. Pink highlights correspond to the phases that are external to the technicalities of ML in itself, while blue identifies the phases that explicitly affect the functioning of ML.

that the data is accurate and that the observed disparities are due to differences that have a causal motivation, and the other where we assume that structural artifacts and historical factors are to blame for the observed differences. This is assuming that all relevant factors are considered in the problem. If we were to do a fairness analysis, the conclusion from the latter viewpoint might be that we are all equal and that any found inequities should be corrected while considering independence criteria.

The discrepancy between the world as it is and as it ought to be is one form of misalignment. But the alignment problem [7] can also manifest when one uses proxies, non-representative data, or mix up ground truth with the information captured by data. Misalignment impedes the correct assessment of ML, rendering the ability to exercise autonomy more challenging. Our attempts to be precise may be hampered by the limitations of our own understanding and the fact that the map we construct based on that understanding does not exactly match the territory we are modeling.

These practices, which rely on the excessive use of ML systems, may jeopardize the value of autonomy by eroding individual autonomy and decision-making. The three actors of interest are the ML-based solution, the organization that owns it, and the end user. The solution unites the two key stakeholders, and its execution reveals the importance organizations place on end-user autonomy. In social media, we see personalized feeds that automatically limit the information we receive; in automated processes, we find a lack of adaptive automation based on risk; in diagnostic tools, we find absolute judgments decontextualized from their uncertainty; etc. These examples show how organizations advocate full automation or information limitation in the name of efficiency, objectivity, and personalization while minimizing end-user autonomy. Respect for autonomy needs to happen at all levels, between individuals and from institutions. Devaluing autonomy can have major ethical consequences and weaken confidence in the system's legitimacy.

3.2. Design

As stated in the AI Act, ML systems cannot be used for manipulating or deceiving people [10]. These unethical behaviors may have an impact on how the decision space is

formed and the ability to weigh the consequences of actions. Therefore, it is essential to carefully evaluate the effects of the design decisions in order to create the necessary safeguards to stop misuse or abuse.

One tool that may foster respect for people's autonomy is meaningful human control over the system. This technique may increase the system's accountability and transparency, enabling people to contest or challenge the decisions made [17]. At the same time, it will avoid over-reliance on ML and ensure that humans remain in charge of the decision-making process. In this line of gaining system control, the adequate allocation of responsibilities is key so that users can evaluate the risks that the given solution entails. This helps prevent agency laundering [19], in which responsibility is transferred away from people who built or deployed the system, and responsibility gaps [16], in which no one is held accountable for the actions or choices of the system.

Another technique for defending the respect of autonomy is related to the automation capabilities of ML-based solutions. Automation entails the cession of autonomy by definition. In the decision-making process, the cession of autonomy means delegating decision-making to the model. A successful system should be able to give up control of the appropriate items at the appropriate times to the end-user. This concept is known as adaptive automation and it is founded on the premise that systems should be modular and always supported by an explainability layer that allows useful information to be exchanged with their users [21]. This strategy will give users control at the right times and adapt to the level of risk they are willing to accept.

3.3. Data

Data is the cornerstone of any ML-based system. Starting with data collection, ML systems that rely on large amounts of interconnected data may violate individuals' privacy rights because of their ability to expand users' profiles. In addition, this may endanger users' autonomy by allowing sensitive information to be utilized without their explicit consent and by allowing decisions affecting their lives to be made without their involvement or comprehension. This might result in judgments being made based on unreliable assumptions and correlations, which would restrict people's autonomy by preventing them from taking charge of their own life and making their own decisions. This might reinforce structural injustices and go against the idea of respecting autonomy, which also entails respecting consent. Sensitive data should not be utilized, even tangentially, if individuals have not given their authorization to do so since it may lead to unjust and discriminatory effects. Furthermore, because some user types are a minority, pre-processing procedures like data cleaning and normalization might change the distribution of the data and ignore them, which reduces their representation in the entire sample. As a result, the representation may be insufficiently representative, contributing to the persistence of systemic disparities and, consequently, the persistent deterioration of their autonomy.

Data can be collected, but it can also be manufactured. This is the case with taxonomies. We will concentrate on taxonomies that characterize people's non-visual qualities without their agreement in order to focus on examples that worsen autonomy. This was the situation with ImageNet previous to 2020, when people were classified based on their presumed occupation, sexual orientation, and gender [24]. In this instance, a group of people categorized others without their consent, infringing on their capacity for self-determination. It is always necessary to include those who are affected by the taxonomy in the decision-making process to ensure that their autonomy is respected.

In any case, data must be accurate in order to solve the problem at hand. But we find sometimes that what the model actually predicts is not always what we think it is predicting. Selecting a "ground truth" that really is the ground truth is not always straightforward. There are cases where complexity is unavoidable, and the only quantifiable data may come from human abstractions, such as the "number of prior offenses", which does not correspond with the desired measure, such as the probability of reoffense. This misalignment between what the data gather and the interpretation we assign to the data is a risk factor for the system's correct behavior.

Data is also utilized to predefine the decision space of the solution. The decision space selected may influence how many alternatives users perceive themselves to have, prompting them to exclude particular possibilities. A limited decision space risks excluding alternative possibilities and assuming that the indicated decision space is complete. However, the data may not contain all possibilities and hence may not truly represent the variety of the whole set. In this scenario thinking beyond the box can be tough, and it usually does not happen at all.

3.4. ML model

Models are the outcome of a series of decisions, such as which metrics to optimize, which models to train, which hyper-parameters to tweak, etc. But also their level of understandability, the level of control they provide to end users, etc. Control over these decisions is lost when ML is outsourced. This is true for foundation models as well as other ready-to-use solutions. The loss of control may be avoided if ML-based systems were sufficiently understandable so that users could get all of the information required to create their own judgment about ML's behavior and limitations. The general lack of transparency in these systems makes it harder to scrutinize them, limiting users' agency which might ultimately lead to a misconception of the assurance on the system.

Models are intended to do a certain task while operating within established boundaries. When models are compelled to answer questions outside of their intended scope, they may deliver erroneous responses. The risk of models making mistakes because they are forced by design to give an answer may jeopardize autonomy insofar as the user may not know the underlying uncertainty the model faced and, thus, erroneously rely on the system. Rejection algorithms [6] are a recognized but less used solution for these situations since they provide users the choice to withhold their response when the uncertainty is greater than the specified threshold.

When ML is used to make judgments or interact with the environment, such as when using reinforcement learning or prediction systems for decision-making, the balance between the algorithm's exploratory and exploitative capabilities must be evaluated. While the former broadens the variety of options, the latter maximizes user adaption. It is common to devote greater weight to exploitation than to exploration in response to commercial strategies. However, this comes at the price of people's autonomy. Giving exploratory capacity less credit involves restricting users' access to information, which already compromises their autonomy, but can be exacerbated by making them more likely to overlook opportunities by threatening the creation of their decision space.

The opt-out mechanism is one technique that allows for the exercise of autonomy with respect to voluntary exclusion from a system. As stated in the Article 21 of the GDPR [11], consent can be withdrawn by the user at any time, and ML-based systems

must be prepared to respond accordingly. Furthermore, systems must also account for individuals exercising their right to rectification (Article 16 of the GDPR) so that no judgments are made based on wrong data, which might further erode people's autonomy.

Last but not least, we consider it crucial to examine how the context of the users is integrated into the model's decision-making. This is addressed by decision theory, which is concerned with how decisions are formed depending on the information available and the context in which it is perceived. The most prevalent technique in ML is to maximize expected utility [23]. This translates to maximizing the probability of the correct outcome or, alternatively, minimizing the empirical risk. However, it is not always the best strategy to choose for combining decision and prediction theories. For example, prospect theory [13] examines how people view their loss and gain perspectives asymmetrically. The degree to which the model is applicable to the users' environment depends significantly on the decision theory used, which has a negative impact on adaptation and, consequently, on the users' degree of effective autonomy.

3.5. Utilization

Both human flourishing and the greater entrenchment of systemic disparities are possible in the post-ML world. To avoid creating a dystopian society, it is crucial to evaluate how ML solutions affect various subgroups of the population in order to spot any possible problems early on. The world is changing, and so is the data. Data is the backbone of ML-based solutions. And for that reason, it is critical that it remains as accurate as possible. If not, the solutions would become outdated since they would be lacking relevant knowledge. This may occur as a result of improvements in the quality that technology can deliver, individuals' change in behavior, populations becoming more diverse, etc. In any case, the actual utilization of ML-based solutions should directly affect the first phases of the ML pipeline to ensure that the adequacy of the solution is not affected.

It is essential to have been transparent about the system's intended usage for users to successfully utilize it. Therefore, it is crucial that ML practitioners convey all decisions made throughout the system's development so that the user can benefit from this knowledge. Lack of knowledge of this information may cause assurance to be interpreted incorrectly, overconfidence in the system, and opportunities to be missed, among others. But not every communication is effective. It is crucial to consider the potential misunderstanding between model creators and end users. Sharing assumptions and making the method as self-explanatory as possible can help close this gap. By doing this, we can encourage a better knowledge of the model and its implications, enabling users to act independently and make more educated decisions, fostering thus the exercise of their autonomy.

Another challenge is deskilling, which in this context occurs when the excessive usage of ML systems leads to a reduction in the required human expertise. Although ML automation might be advantageous to increase efficiency, it can also have negative consequences, such as lowering the overall level of human ability and knowledge required in a specific profession. As a result, decision-making may become more reliant on ML models, which are not always precise and can contribute to wider social and economic inequalities. As a result, our autonomy has been put at risk.

One of the greatest risks of ML models is that they become reality. "The world is its own best model" [4], hence, rather than modeling the world, appropriate perception-to-

action systems can be employed to engage with it directly. Performative predictions [18] are forecasts that have an effect on the target. Thus, performativity might take the shape of a distribution shift as the decision-maker applies a prediction model. The ambiguity of performativity's effects extends beyond ML practitioners to end-users, who might be concerned that more interaction with the system could lead to future deterioration of their interaction with the system.

4. Conclusions

As the number of accessible ML systems expands significantly, it becomes more worrying that ethical issues are not sufficiently integrated throughout their development. The rights of individuals, notably their autonomy, are directly impacted by this absence. Instead of fostering autonomy, ML systems actually act against it by weakening people's ability to make free and informed decisions. The difficulty of incorporating ethical concerns from the start is being made more difficult by the speed at which ML systems are being developed and used in daily life. This helps to explain why there have not been as many efforts to operationalize autonomy effectively, which would have allowed for the translation of abstract ideas into logical languages that allow for its practical application. In order to ensure that autonomy is properly understood, operationalized, and deployed, we propose an ethical analysis from the inception of ML systems.

In this study, we analyze the potential hazards that could emerge in each of the indicated ML pipeline phases and pose a threat to the successful exercise of our autonomy. We emphasize the need of recognizing the constraints of modeling the world as it as well as examine all aspects of the alignment problem from the start. We highlight the need of anticipating mechanisms to ensure human control in order to avoid excessive reliance and responsibility gaps in the design phase, the impact of the data chosen on the perpetuation of current practices, and the effect of data selection on the formation of the perceived decision space. We also emphasize the need of putting the end-users at the heart of the ML model in order to avoid neglecting their context. Finally, we analyze the risks associated with the utilization of ML systems, which may result in deskilling or even changing the world as it is through feedback loops.

We propose a transdisciplinary approach in a society that takes for granted that the decisions we are going to make with the support of ML systems prevail over our own. We question that this basis does not allow us to make autonomous decisions and we conduct a critical diagnosis to identify violations of the respect for autonomy at the different phases of the ML pipeline. Our work aims to provide useful advice that guides AI practitioners in identifying potential risks when developing ML solutions that endanger the respect for autonomy and, ideally, advise effective actions to take to reduce the hazards that are identified.

5. Acknowledgements

This work is partially supported by MCIN/AEI/10.13039/501100011033 under project PID2019-105093GB-I00 and PID2022-136436NB-I00.

References

[1] AI HLEG. Ethics Guidelines for Trustworthy AI. Technical report, High-Level Expert Group on Artificial Intelligence, Brussels, 4 2019.

[2] Zygmunt Bauman. *Liquid life*. Polity, 2005.

[3] Tom L Beauchamp, James F Childress, and Others. *Principles of Biomedical Ethics*. Oxford University Press, USA, 2001.

[4] Rodney A Brooks. Intelligence without reason. In *The artificial life route to artificial intelligence*, pages 25–81. Routledge, 2018.

[5] Rafael A. Calvo, Dorian Peters, Karina Vold, and Richard M. Ryan. Supporting Human Autonomy in AI Systems: A Framework for Ethical Enquiry. *Philosophical Studies Series*, 140:31–54, 2020.

[6] C Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46, 1970.

[7] Brian Christian. *The Alignment Problem: machine learning and human values*. W.W. Norton & Company, New York, 1 edition, 2020.

[8] Itziar de Lecuona, María Jesús Bertrán, and et al. Guidelines for reviewing health research and innovation projects that use emergent technologies and personal data. *Review of biomedical research and innovation that uses emergent technologies and personal data: the challenges facing research ethics committees in a data-driven society*, 2020.

[9] Edward L. Deci and Richard M. Ryan. Intrinsic Motivation and Self-Determination in Human Behavior. *Intrinsic Motivation and Self-Determination in Human Behavior*, 1985.

[10] European Commission. 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. Technical report, European Commission, Brussels, 4 2021.

[11] General Data Protection Regulation. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC. Technical report, Official Journal of the European Union, 4 2016.

[12] Thomas E Hill. Kantian autonomy and contemporary ideas of autonomy. *Kant on moral autonomy*, pages 15–31, 2013.

[13] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292, 1979.

[14] Arto Laitinen and Otto Sahlgren. AI Systems and Respect for Human Autonomy. *Frontiers in Artificial Intelligence*, 4:151, 10 2021.

[15] Catriona Mackenzie and Natalie Stoljar, editors. *Relational autonomy: feminist perspectives on autonomy, agency, and the social self*. Oxford University Press, New York, 2000.

[16] Andreas Matthias. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology 2004 6:3*, 6(3):175–183, 2004.

[17] Frank Pasquale. *The black box society: The secret algorithms that control money and information*. Harvard University Press, 2015.

[18] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative Prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 11 2020.

[19] Alan Rubel, Clinton Castro, and Adam Pham. Agency Laundering and Information Technologies. *Ethical Theory and Moral Practice 2019 22:4*, 22(4):1017–1041, 10 2019.

[20] Alan Rubel, Clinton Castro, and Adam Pham. *Algorithms and Autonomy: The Ethics of Automated Decision Systems*. Cambridge University Press, 2021.

[21] Mark W Scerbo. Theoretical perspectives on adaptive automation. In *Automation and human performance: Theory and applications*, pages 37–63. CRC Press, 2018.

[22] Paula Subías-Beltrán, Oriol Pujol, and Itziar De Lecuona. The forgotten human autonomy in Machine Learning. In Desara Dushi, Francesca Naretto, Cecilia Panigutti, and Francesca Pratesi, editors, *Proceedings of the Workshop on Imagining the AI Landscape after the AI Act (IAIL 2022)*, Amsterdam, 9 2022.

[23] John von Neumann and Oskar Morgenstern. *Theory of games and economic behavior*. Princeton University Press, 3 2007.

[24] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 547–558, 2020.