Coding for Respect

A Philosophical Framework for Evaluating the Ethics of Autonomous Machine Action

Michael Skirpan
University of Colorado
Boulder, CO
michael.skirpan@colorado.edu

Willie Costello
Stanford University
Palo Alto, CA
williec@stanford.edu

ABSTRACT

The influence of autonomous machine actions on human life is growing rapidly. This paper presents an interpretative framework toward future policymaking and ethical engineering that protects and promotes human interests in the face of intelligent machine systems. Drawing on insights from contemporary moral philosophy, we offer a model for thinking about respectful action between humans and then translate this to the machine case. Applying a novel rubric of analysis, we develop moral intuitions around what it means for machine-intelligent systems to act with "respect" toward humans. Working from a set of cases, we delimit the space of interactions between human autonomy and autonomous machine action. We end by discussing the possibilities of employing our framework for future policy and technical work in the area.

KEYWORDS

Ethics, Policy, Machine Learning, AI, Fair, Respect, Autonomy, Responsible Machine Action

1 INTRODUCTION

In a world where machines are increasingly being used to act autonomously in the place of humans, we all have an interest in preserving an ethical character to autonomous machine action (hereafter "machine action"). Indeed, this is often part of the very reason why we want machines to act for us in the first place: to remove the human flaws, biases, and oversights which lead such actions to be unethical or otherwise harmful when performed normally by humans. However, focusing solely on the ethically negative action characteristics (such as bias) that machines *might* [27] be able to avoid overlooks the ethically positive action characteristics that machines must still act in accordance with. Otherwise, in replacing human action with machine action, we will simply be trading one kind of unethical action for another – and machine actions might end up being unethical in even more problematic ways than the human actions they are replacing.

Some work in this direction is already happening. The current literature of fair machine learning works to overcome the problem of machines inheriting bias from their training data [7]. Aiming to discover [3] and constrain [31, 48] algorithmic bias, however, limits our focus to solving a single ethical issue of preventing illegal discrimination in machine action. And while we know we like models that are interpretable [37, 44], it is still unclear what boundaries we should be putting on the behavior of intelligent machines. As we move into the future, it is going to be critical, both socially and legally, that we build intuitions about what constitutes an ethical machine action, and formulate guidelines that engineers, technologists, policymakers, and lawyers can keep in mind when designing, evaluating, and regulating machine actions.

Our recommendation in this paper is that we should look to *philosophy* to help build these intuitions and formulate these guidelines. Philosophy, moral philosophy in particular, has the resources we need to expand and hone our ethical vocabulary, and to develop interpretive metrics for regulating algorithmic systems.

In this vein, our specific contribution here is to bring attention to the importance of the concept of *respect* to any adequate account of ethical action – that is, respect *for persons*, or treating others never simply as means but always as *ends in themselves*. Respect is widely recognized in contemporary moral philosophy as a crucial component of ethical action, and, as we argue, it should be seen as a crucial component of ethical machine action as well. As we demonstrate, respect stands behind many of the ethical aspects of machine action that we already recognize and care about, such as privacy, fairness, accountability, and transparency. However, work still needs to be done to know how to systematically apply the concept of respect to the case of machine action. In addition, machine action includes some further, hitherto unrecognized dimensions along which respect also needs to be preserved and promoted.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

FAT, 2018*

© 2018 Copyright held by the owner/author(s). Publication rights licensed to Association for Computing Machinery.

The structure of our paper is as follows: In §2, we present a brief philosophical introduction to the concept of respect, and explain why respect is so important to the ethics of human action. In §3, in order to know better how to apply the concept of respect to the case of machine action, we present an analysis of machine actions into their six most ethically salient components. In §4, pairing together the insights from the preceding two sections, we show how the concept of respect can be applied systematically to the case of machine action, and detail the specific ethical considerations that emerge from this application. In §5, we briefly indicate the upshots of this framework for technologists, policymakers, and other practitioners.

Through this discussion, we have the following broad goals: First, to provide a philosophical framework for thinking through and evaluating the ethics of machine action. Second, to clarify the some of the specific ethical questions we should be asking about machine actions. Third, to highlight some ethical dimensions of machine action that are not being discussed much at present, but which must be in order to ensure a fully ethical character to machine action.

2 THE CONCEPT OF RESPECT

Our approach to the ethics of machine action takes as its starting-point the concept of *respect*. We choose to focus on respect for three reasons. First, respect is generally thought of by contemporary philosophers to be one of, if not *the*, fundamental concept of morality, from which all our other, more specific moral obligations follow. Thus respect cannot be ignored in *any* discussion of ethical action. Second, as we detail below, respect has been fruitfully applied to a variety of other real-world ethical issues. Thus there is reason to look to respect for guidance in the case of the real-world issue of machine action as well. Third, as we show in §4, respect captures many of the more ethical moral norms of machine action that we already know and care about. Thus a better understanding of respect will clarify the importance of these norms, as well as possibly recommend further, hitherto unrecognized norms for our consideration.

What, then, *is* respect? Though there are many different kinds of respect [19, 35], one kind in particular has been of primary concern to moral philosophers: namely, respect *for persons*, that is, the respect that each person is due simply in virtue of being a person, the respect that is due to all persons *as such*. The basic idea is that, although in certain contexts some people might deserve more respect than others (due to differences in position, status, talents, or the like), there is nonetheless a baseline kind of respect that *all* deserve, due to the distinctive moral status as persons that we all share. "Persons, it is said, have a fundamental moral right to respect simply because they are persons" [20, §2].

This concept of respect should not seem foreign; indeed, it is natural for us to think that we have special categorical obligations to all other persons [14]. Yet respect would not have the place it has today were it not for the work of eighteenth-century German philosopher Immanuel Kant. For Kant, all our specific moral obligations and duties flowed from one ultimate moral principle, the "Categorical Imperative", which in its second formulation (the "Formula of Humanity") he expressed as the command that one act so as to treat persons "never simply as a means but always at the same time as an end" [28, 4:429]. Whatever else this dictum means, it is a command to respect persons *qua* persons [19, p. 36], and this basic idea continues to animate moral philosophers today [18, 29, 47]. Some contemporary philosophers, following Kant, have taken respect to be the basis of all morality [21, 22]; and while others deny this [25], all agree that respect is an important ethical concept. For the purposes of this paper, we need not take a stand on this issue; it is sufficient merely that respect is accepted as an important ethical concept, in light of which moral principles and judgments can and should be formulated.

Why, then, is respect important? Take the following simple but powerful example (adapted from [24, 42]). Imagine a doctor who kills one of her patients, who is perfectly healthy, in order to harvest that patient's organs and give those organs to five of her other patients, who are each about to die from a different kind of organ failure. In one respect, the doctor has promoted "the greatest good for the greatest number" (the well-worn slogan of utilitarian moral theories), saving five people while losing only one, whereas were she to do nothing five would have died and only one would have lived. Nonetheless, the doctor's action clearly strikes us as morally wrong, as the doctor has treated her healthy patient merely as a means to her other patients' health. If the doctor had respected her healthy patient and recognized his fundamental moral rights (in this instance, his right to life and to his own body), she would not have engaged in such morally wrong behavior.

So what, more specifically, does respecting persons actually involve? What does it mean, in Kant's words, to treat others never simply as a means but always as an end? Primarily, respect involves *refraining* from certain actions and behaviors, such as the exploitation, manipulation, and debasement of others, violations of their rights, and interference with their decision-making or self-governance. In other words, respect sets a "boundary condition" for moral action, demarcating a subset of actions that are absolutely morally wrong. Respect does not always dictate exactly what we *should* do, but it does always indicate what we should *not* do. This is why Kant refers to the Formula of Humanity as "the supreme limiting condition of the freedom of action of every human being" [28, 4:431].

In addition to this behavioral dimension, respecting persons also involves a *deliberative* dimension. Respect requires not only that one act so as to respect persons as such, but also that one give consideration to persons as such in one's deliberations about how to act. Kant, and many other moral philosophers since, have emphasized the deliberative dimension of respect over the behavioral. One reason for this is because respect itself is generally thought to be, fundamentally, an attitude or state of mind, and thus is more properly revealed in deliberation than behavior. But a deeper reason is because the deliberative dimension is thought to explain the behavioral: when one gives due consideration to persons as such in one's deliberations, one will, as a result, generally act respectfully towards them [12, p. 210]. Thus the best way to ensure respectful behavior is by encouraging respectful deliberation. This will be an important point to keep in mind as we move forward.

Lastly, it should be noted that respect for persons is not just some general principle of abstract moral theory; it also has a number of real-world application contexts. In the past half-century, moral and political philosophers have drawn on the concept of respect to explain and justify the nature and importance of moral rights [23], equality [10, 26, 46], social justice [34], privacy [9], political liberalism [12], and multiculturalism and the politics of recognition [40]. Of particular note is how attention to respect transformed the field of biomedical ethics, as well as actual health care practice, by emphasizing the importance of patient autonomy, which is now widely recognized as a basic principle of bioethics, and which has provided an essential counterpoint to the traditional norm of physician paternalism [8].

3 AN ANALYSIS OF MACHINE ACTION

In the previous section, we introduced the concept of respect and explained its importance to the ethics of (human) action. Thus, if we wish to preserve an ethical character to machine action, we must ensure that machine action lives up to this normative ideal, too. Yet it is not immediately obvious how the concept of respect should be applied to the case of machine action, as machine action is in many ways different from normal human action. However, this dissimilarity should not discourage us. As we will see, with a rigorous enough understanding of machine action, applying the concept of respect to machine action becomes relatively straightforward. Furthermore, clarifying the differences between machine action and human action will reveal dimensions of respect that are unique to the case of machine action and thus all the more important to explicitly attend to, as we will argue in §4.

To know better how to apply the concept of respect to the case of machine action, we first need a more fine-grained analysis of machine action. Here we present an original analysis, which we believe captures the most ethically salient components of machine action. However, we should note at the outset that this analysis is not intended as definitive or unassailable; others can and should modify or expand this analysis as they see fit. The present analysis is intended, rather, as a template or framework for the sort of analysis of machine action that we believe is needed to think more rigorously and systematically about the ethics of machine action.

We analyze machine action into six components, divided into two groups (Table 1). The first group encompasses what we call the "kinds" of machine action, or the types of action or sub-action that a machine intelligent system might perform. Here we identify three basic kinds: observations, classifications, and interventions. The second group encompasses what we call the "strata" of machine action, or the levels at which a machine action might be carried out. Here we identify, similarly, three basic strata: individual, collective, and iterative.

Group	**Component**		
Kinds	Observation	Classification	Intervention
Strata	Individual	Collective	Iterative

Table 1: The components of machine action

Some basic examples will help clarify what we mean by each of these different components. Let us begin with the three kinds of machine action.

"Observations" are actions that correspond to some mode of data capture [4]. These actions are often defined specifically by an engineer who designs a data mining or behavioral tracking system; however, autonomous experimentation [13] is beginning to replace the need for human definitions. Examples of observations are taking behavioral analytics, server logging, storing revision histories, and sensor time series. Observations are, of course, a component of human action as well; but notably, machine observation elevates the ability for tracking and experimenting beyond what a human could do. Furthermore, whether constructed by a human or an autonomous algorithm, machine observation establishes the space of possibilities, or ground truth, that will determine all further actions that the machine or any associated machine intelligent systems will take.

"Classifications" are actions that assign users classifiers and change how they are "seen" or treated by a machine intelligent system (typically, on the basis of data that has been collected from a prior observation action). Classifications are the process of using statistical and computational methods to cluster, organize, or label users and their data [39]. Examples of classification include training a classifier that predicts a user's religion [30] or mental health [32, 36], the results of

which will then change what ads or content the user sees. Trained models may be construed as a species of classifications, as such models form the basis from which future decisions are made. The act of creating classifications allows for the further embedding of those classifiers or models into any relevant autonomous machine system.

"Interventions" happen when a machine actually does something, changes something, or interacts with a user directly (typically, on the basis of previous observations and/or classifications). Interventions are the positive product of a machine acting in the world. Examples of interventions are establishing the visual order of an interface or suggesting auto-complete text. Interventions are the most material and manifest action that a machine may take. The results of an intervention are felt by, and may even harm [43], the user, whether or not they comprehend its occurrence or consequence.

By combining these three kinds of machine action together, we can form a pipeline where each kind of action feeds into the next. Thus an observation can create data, which can then be categorized or structured via some analysis or training, which can result in a model or set of categories which allows for a specific intervention on a user. A simple example of this would be the filtering of social media feeds: In this machine intelligent system, data is first captured on the basis of user clicks or likes of content (the observation actions); next, data that has been collected from many users goes into the training of a system that ranks content given a specific history (the classification action); and finally, an autonomous system filters and orders live content, using those learned ranks, at the moment a page request is made (the intervention action).

Here it should be noted that, although interventions may seem like the purest and most important form of machine action (since they are the finite moments when a human consequence is actually felt), each kind of action in this group uniquely impacts the conception of and behavior towards the user in a machine intelligent system. This point is especially relevant when applying the concept of respect to machine intelligent systems, since, as was noted above, respect involves not only a behavioral but also a deliberative dimension. In this regard, the steps that feed into and inform any machine intervention (i.e., the machine's "deliberations") are just as important to respect as the intervention itself (i.e., the machine's "behavior"). We return to this point below, in §4.

Let us now turn to our second group of components, the three strata of machine action. Beyond the *kinds* of actions that machines may take, there are also separable *loci* at which machine actions take place. The "individual" stratum considers a machine action from the perspective of its impact on a single user. Examples of individual-stratum actions include observing a single user's history of responses to a marketing campaign, assigning a single user a category relevant for targeting, and rendering a specific article or ad at the top of a user's feed.

The "collective" stratum views machine actions from the perspective of how they affect a population of users. Examples of collective-stratum actions include collecting data from all employees in a workplace; a data classification that (directly or indirectly) serves as a proxy for a protected class, such as race or political orientation; or an image captioning system that exclusively mistags black faces [1]. To be sure, collective actions emerge out of the combination of many individual actions; nonetheless, they warrant separate ethical consideration, as the combination of individual actions is itself a matter of ethical concern. This is especially true when the combination of individual actions results in impacts linked to a shared attribute that targets a specific community (e.g., disparate impact [7]).

The "iterative" stratum is the most distinctive component in our analysis of machine action. Iterative actions are actions viewed from the perspective of the dynamics and impacts that occur due to same basic action being repeated and reiterated numerous times (as, indeed, many machine actions are). We separately identify this stratum due to the fact some actions, though harmless when looked at in isolation, become disconcerting once the action is performed repeatedly and continuously. Take, for example, the persistent shaping of a user's news feed. If we were to analyze the personalization of a news feed on the individual stratum, we may find nothing wrong with it. But when we consider the same personalization being iterated over time, to the extent that a user sees nothing but what is personalized for them, the psychological and social consequences of the action become tangible.

The preceding discussion should suffice to clarify our six-component analysis of machine action. If it is not already clear, any machine action can be described according to both its kind and stratum. Thus, by treating our two groups of components as axes and crossing them, we can create a 3x3 matrix, in which we can locate any particular machine action according to its kind and its stratum (Table 2).

This analysis of machine action may seem illuminating in its own right, but recall that our ultimate reason for introducing it was so as to know better how to apply the concept of respect to the case of machine action. And indeed, as we will argue presently, for each kind and stratum of machine action that we have identified, there is a distinct and specific moral obligation, arising out of the fundamental moral obligation to respect persons, which must not be violated if the action is to be ethical. (Note that this is a necessary, but not necessarily sufficient, condition for ethical machine action.) Understanding these different obligations is, we believe, foundational to the structuring of any future regulation that looks to protect humans in the face of actions taken by autonomous machine

systems. In the next section, we outline what each of these obligations are, how they derive from the more basic obligation to respect persons, and explain why they are important and relevant to the case of machine action.

4 RESPECT IN MACHINE ACTION

Recall that, when applying the concept of respect to the evaluation of any action, the basic question we must ask is: Does the action respect the persons involved in and/or affected by the action? (Does the action treat them simply as means, or as ends in themselves?) Accordingly, when applying the concept of respect to the evaluation of machine actions, the basic question we must ask is: Does the machine action respect the users involved in and/or affected by the action? (Does the machine action treat the users purely as instrumental to some goal, or as autonomous individuals with the basic moral standing of persons?)

As stated, this basic question is still too general and abstract to offer any specific guidance in the evaluation of machine action. However, more specific and focused questions can be posed, in light of the analysis of machine action introduced in §3. This is because, as we argue in this section, respect is manifested in a distinctive way for each kind and stratum of machine action defined above. These results are summarized in Tables 3 and 4 below. But to more effectively introduce and convey these ideas, we first present a number of concrete cases that illustrate the different forms that respect takes for each of our kinds and strata of machine action. These cases will also clarify how prior work in machine learning and computing ethics fits into our framework.

A preliminary note: One of the distinctive aspects of our framework is our assumption that respect must be preserved and promoted for *all* kinds and strata of machine action. This assumption may strike some as surprising, as it may seem that, when it comes to the ethics of machine action, it is only machine *interventions* that we should be worrying about, since this is where the discrete harms of machine actions actually materialize. However, here again we point to the fact that respect for persons involves more than the here performance of respectful *behavior*. In addition, it requires manifesting respect in one's *deliberations*, or the considerations and subactions that lead up to and result in one's behavior. Thus, it is crucial to recognize that we are here outlining a matrix for discovering the *provenance* of a machine's disrespect for its users, as opposed to identifying specific machine interactions that actually cause harm. Comprehending this broader spectrum of ethical action is, we believe, the only way to tease apart the increasingly complex interactions that machines are having, and will continue to have, with humans.

That being said, let us turn now to our cases. We present one case for each of the nine regions in our matrix (Table 2). We begin, in §4.1, with some familiar cases of machine action (Cases A–D), and then turn, in §4.2, to some emerging and less familiar cases (Cases E–I).

Strata \ *Kinds*	Observation	Classification	Intervention
Individual	A	C	F
Collective	B	D	H
Iterative	E	I	G

Table 2: Cases in our matrix of machine action

4.1 Familiar cases

Some of the kinds and strata of machine action in our analysis already have a rich literature concerning the ethical problems they raise. Here we briefly detail four of the nine regions in our matrix, whose ethical issues are widely recognized in machine learning and computing ethics. Yet as we demonstrate, these well-known ethical concerns can all be seen to derive from the basic moral obligation to respect the persons (or users) involved. This, in turn, provides some initial justification for our respect-based framework.

Case A: Private data collection

[Individual Observation]

Example: Devices and applications that process personal information about users often violate reasonable expectations of privacy. The use of technology to collect or show private information at unexpected, unknown, or unwarranted junctures has already occurred in many cases. The growth of the internet of things, and ubiquitous computing more generally, gives rise to machines able to gather information beyond what an individual can comprehend or manage. This threatens our individual ability to manage what is known about us and to whom. A 2015 study of mobile applications [49] showed that personal information such as location and email address are being shared with third-parties with no notification to the user. The FTC recently settled a case with Vizio for not disclosing how it was collecting user information [2]. Uber's God View [11] further proved that without reasonable checks, the data we expect to be private could be reappropriated outside of our control.

Analysis: There are two clear ethical concerns in this case. First, such data collection violates the user's rights, namely, their right to privacy. Second, such data collection is (often) opaque to the user. Our framework explains why these are legitimate ethical concerns, for both can be seen as failures to respect the user. Moreover, both failures can be seen to correlate to the kind and stratum of such actions, that is,

the fact that they are individual observations. First, for any individual action to respect the user it affects, it must refrain from violating that user's fundamental human or moral rights, that is, the rights the user has simply in virtue of being a person [23], and these include the right to privacy [9]. Second, for any observation action to respect the user it observes, it must ensure that the observation is transparent, such that the user can be aware of it. Observations that fail in this regard end up treating their users purely instrumentally, as mere data points to be collected, and not as the persons that they are.

Case B: Uneven data collection

[Collective Observation]

Example: Deploying a data capture system that does not collect data evenly across populations creates the problem of over- or under-sampling certain populations. We expect the data we collect to represent the true state of affairs, eliminating the biases that humans carry. However, we are finding that technology-enabled data capture carries its own potential to exacerbate inequality between different communities [6]. The escalation of predictive policing practices has raised serious questions about the fairness of the impacts. Attention has already been paid to the fact that facial recognition databases disproportionately represent black faces [38]. Conversely, data collected to solve problems might bias solutions to aid only those with the most access. Such was the case with the StreetBump app which helped get potholes get fixed, but due to faulty data collection methods, did so primarily for richer neighborhoods [16].

Analysis: This case introduces a further ethical concern, namely, that such data collection seems unfair, in that it disproportionately and unequally affects the users in its population. Again, our framework explains why this is a legitimate ethical concern, as this, too, can be seen as a failure to respect the users involved. Furthermore, this concern arises from the fact that, in this case, we are considering data collection at the collective stratum, that is, from the perspective of how such actions affect a population of users. For any collective action to respect its users, it must ensure that they are all treated fairly and given equal consideration [10, 46]. This is because all users share the same fundamental moral status as persons. Flagrantly unequal and unfair treatment of user populations fails to respect this shared moral status.

Case C: Triangulating sensitive information

[Individual Classification]

Example: When personal data is handed over to companies, users can be unaware of what they are truly disclosing. The ability to predict undisclosed attributes about individuals using data they provided has already been proven [30]. There have even been cases of companies selling highly-sensitive, anonymized medical data to companies that can de-anonymize the data set using complimentary data [41].

Analysis: The ethical concerns in this case are slightly more subtle. Like Case A, part of the issue is that such actions seems to violate the user's right to privacy [17]. This is not surprising given our framework, since this case, like Case A, is considering action at the individual stratum; and as we saw above, respect at the individual stratum requires not violating the user's rights. In addition, the actions of this case seem objectionable because the user has not been given a voice, or say, in their classification; rather, the classification was made on their behalf, without their consultation or consent, and without even the possibility for the user to object or intervene. Again, our framework explains why this is a legitimate ethical concern, as such "voiceless" classification of a user can be seen as a failure of respect. Indeed, for any classification action to respect the user it is classifying, it is essential that the user be given some degree of voice in their classification. Otherwise, the classification ends up treating the user paternalistically, as a mere means to some desired data-set, and not a free and self-determining individual.

Case D: Sentencing algorithms

[Collective Classification]

Example: The highly-publicized case of the COMPAS recidivism scoring algorithm highlighted the threat of machines classifying people along racial lines [5, 15]. Once machines codify relationships in high-dimensional feature spaces required for complex models, they have the potential of constructing unfair classifiers that discriminate on the basis of identity. Further, the use of a machine system to solve a contentious human problem minimizes the affected person's ability to understand and potentially redress any harm.

Analysis: This case raises two ethical concerns, which we have already seen above. As a collective action, we must ensure that the action is fair; and as a classification action, we must ensure that the action gives users a voice. On both counts, this case seems problematic, treating the users in its population unequally, and not giving its users any say in their classification.

Summary of familiar cases

These familiar cases highlight the relevance of respect to the ethics of machine action, and detail how respect can be more thickly conceptualized within our framework. To ensure respect, individual actions must not violate the user's *rights*, collective actions must treat all users *fairly* and *equally*, observations must be *transparent*, and classifications must give users a *voice* (Table 3).

These specific norms and considerations should already seem familiar to those in the machine ethics community. Yet our discussion highlights that these acknowledged ethical concerns derive from the basic moral obligation to respect persons. In other words, in emphasizing privacy, fairness, accountability, and transparency, practitioners in the machine ethics community have already tacitly been working to ensure respect in machine action.

This is significant, because respect manifests itself in the other dimensions of machine action, as well, in ways that are *not* at present being discussed. A complete ethics of machine action must also pay attention to these forms of respect.

4.2 Emerging cases

Thus far, we have looked at cases of individual, collective, observation, and classification machine actions. What remains to be discussed are cases of iterative and intervention actions. Such cases are especially relevant within our framework, since respect in these cases is not as well understood as it is in the others. As we argue below, respect in these cases involves preserving and promoting the user's *autonomy*: the user's fundamental right to act freely and of her own accord, rather than being coerced or having decisions made for her.

To further comprehend the threat of these less-understood issues, we present novel case studies meant to aid in our collective understanding of the dimensions of respect these areas of our matrix put into question.

Case E: Low-level behavioral tracking

[Iterative Observation]

Example: Elaine uses her mobile phone to read news. Even though she chooses not to click on certain articles, she sometimes pauses, shocked by the headlines. To try to minimize the amount of shock content she receives, she often copies and pastes information to a separate application to fact check. While she believes to be avoiding giving information about what shocks her, the fact observation continues across all facets of an interaction, she cannot help but inform an intelligent machine of the truth via her behavioral metrics. However, without some other feedback mechanism for Elaine, it is possible, if not likely, she will be fed more shocking content.

Analysis: This case is an example of an iterative observation: it is an observation, because data is being collected; and it is iterative, because the data collection is done expansively and continuously. Like other observation actions, part of the problem here is that the data collection is not transparent to Elaine. But this case also has a distinctive problem due to its iterative dimension. Despite Elaine's best efforts to prevent the app from knowing too much about her habits, the app's iterative data collection outstrips her ability to do so. In this way, the app undermines her ability to act freely and of her own volition, and in this regard compromises her autonomy.

Case F: Targeted advertising

[Individual Intervention]

Example: Frederick wants to go to Harvard for college. Leading up to receiving his admission decision, he displays anxious behaviors by scouring the internet for information about who has received admission and posting messages to his friends showcasing how much the uncertainty is bothering him. Adapting to his social media data and behavioral metrics, he begins receiving information related to anxiety and depression medication. Despite never having thought of himself as depressed or feeling unhealthy anxiety before, upon being rejected from Harvard, in a moment of vulnerability, he is shown an ad framed as "Is anxiety harming your performance?" and chooses to click and purchase.

Analysis: This case is an example of an individual intervention. Its distinctive ethical concern comes from its intervention dimension. The machine is interacting with Frederick in an attempt to influence him and make him respond. In itself there is nothing wrong with a machine (or a human, for that matter) doing so. However, because the machine is using its trove of data to show Frederick the ad precisely when it knows he is at his most vulnerable, the intervention can easily seem coercive or exploitative. If a human were to intervene in this manner, we would consider it predatory. In its specificity, the intervention compromises Frederick's autonomy and disrespects his status as a free, autonomous individual.

Case G: Influential advertising

[Iterative Intervention]

Example: Just before bed, Gina tends to browse the internet for new clothing. Recently, she has noticed more photos and ads with fashion models showing up across all of her platforms, especially at night. Gina is now beginning to feel insecure when she's tired and can't stop being bombarded with expensive clothing and thin models. After a few nights where her mind began to wander toward very negative thoughts, she decided she should stop looking at clothes before bed and read something else instead. However, now that her interests and browsing times have been learned, even on unrelated platforms, she is continuing to be shown models in every ad bar. Her insecurity worsens and she is now considering counseling for body image issues which she never used to have.

Analysis: This case is an example of an iterative intervention, combining elements from Cases E and F. As we have already seen, both iterative and intervention actions run the risk of compromising a user's autonomy. In this case, the iterative dimension is what is particularly ethically problematic. As

a one-off intervention, showing an ad to a user does not disrespect their autonomy; we can reasonably expect the user to display some resilience. Yet when performed iteratively, this intervention becomes more of a concern, as it can shape, influence, and determine a user's self-conception, and thereby compromises a user's ability to form that conception freely and for themselves. These tactics have already been discussed as an emerging possibility in targeted marketing [33].

Case H: Filter bubbles

[Collective Intervention]

Example: Heather and Henry are friends in real life and social media. Though they have very separate interests, their friendship is built on a mutual respect for one another. One day Heather sees information in her news feed about a new federal policy proposal that she finds concerning for her family. Knowing that her friends may care to know how this could impact her, she posts a heartfelt plea to stand against this policy, outlining the impact it would have on her particular family situation. She asks that if you disagree not to bring it up in front of her brothers or sisters who are having issues. Meanwhile Henry receives news that explains a polar opposite view of the same policy and his feed filters out Heather's plea due to distinct political interests. Doug then raises the issue in front of Heather and her sister while at a party, which hurts Heather given her public plea. Neither knows what the other has seen and thus both assume the other is completely unreasonable.

Analysis: This case is an example of a collective intervention. The intervention itself (like Cases F and G) is concerning due to the fact that it is opaquely filtering information in a way that restrains individuals' – such as Henry's and Heather's – ability to freely communicate. Moreover, this case is a collective intervention, due to the fact that individuals are being clustered along interest lines and the effect of the intervention is to segment the user population. Thus, while the personalization of information feeds is not always in itself problematic, in this case it both limits the communicative capacity of the individuals and compromises the assumed equality of public dialogue by restraining information flows across groupings. Critically, the fact Henry and Heather *want* to communicate with each other in a public setting, yet are being *opaquely* undermined in doing so, disrespects their autonomy.

Case I: Collateral classification

[Iterative Classification]

Example: A large online search engine begins classifying words and semantics that are likely to signal highly politicized or fake news. Leading up to a major federal election, Isaac is following a surging fringe candidate building conversation around universal basic income. Though the topic is getting traction, it does not get taken up seriously by major media networks. Beyond covering this surging candidate, a number of smaller publications have begun publishing articles using unsound claims around environment, refugees, and medicine. Seeing results being overcrowded by fake news, the search company decides to apply its classifier to its search engine to improve people getting news with sound sources. With only a few days until the election, all articles around universal basic income along with those discussing unsound science disappear from the top of search results. By the time the company realizes the issue, the election day has come and went and the surging candidate lost traction and exposure.

Analysis: This case is an example of iterative classification. It is a classification because we are dealing with a trained model classifying news content that is likely to be fake or biased; and it is iterative because the same internal logic of the model is being applied over and over again to all articles. In this case, the model was very good at identifying fake news and lowering its rank in search results. However, it came with a hidden cost of misclassifying articles representing real, factual political discourse that carried features intertwined with those signaling fake news. Yet the ethical concern in this case goes beyond the model's failure to accurately classify real news. More fundamentally, the model impeded voters' ability to find news related to their issue of choice [45]. It decided what was relevant for them, rather than allowing users to freely make this decision for themselves.

4.3 Summary

In this section we have argued, on the basis of the above cases, that the general moral obligation to respect persons can be seen to take on a more specific form according to the kind and stratum of machine action at issue. Thus, each kind and stratum can be associated with a specific norm of respect (Table 3), and each of these norms can be understood in light of a corresponding ethical question (Table 4).

Group	Component	Norm
Kinds	Observation	Transparency
	Classification	Voice
	Intervention	Autonomy
Strata	Individual	Rights
	Collective	Fairness
	Iterative	Autonomy

Table 3: Norms of respect

Norm	Question
Transparency	Is the user (able to be) aware of the action?
Voice	Is the user able to influence the action?
Rights	Does the action violate the user's rights?
Fairness	Does the action treat all users fairly and equally?
Autonomy	Is the user still able to act freely and of her own accord?

Table 4: Questions of respect

Here, then, is our basic recommendation: If you are considering the ethics of any particular machine action, first identify which cell(s) in our matrix the action falls under, and second pose to yourself the corresponding questions of respect.

Admittedly, this recommendation leaves many of the hard questions unanswered. We do not assume that it is obvious when, for example, an observation is sufficiently transparent, or a user is able to act freely and of her own accord. Yet this is not the purpose of our framework. Its purpose, rather, is to clarify the kinds of questions that we should be asking when evaluating the ethics of machine action. By having these questions in view, it is our hope that the task of designing ethical machine actions will seem more tractable and a little less daunting. The hard work of answering these questions, however, lies with all of us.

5 CONCLUDING REMARKS

This paper has addressed the ethics of machine action using a philosophical framework based on the concept of respect. Our fundamental assumption has been that a linchpin of all ethical action is that persons – the users in intelligent machine systems – must be treated as persons, that is, as autonomous agents and ends in themselves. Based on the multifarious abilities of intelligent machines, we applied our framework of respect to clarify the various ways in which machine actions can undermine this fundamental moral obligation and disrespect their users, by treating them as purely instrumental to their objective. We believe this framework encompasses much of what has already been looked at through the lenses of privacy, fairness, accountability, and transparency, while at the same time expanding our intuitions about ethical machine action in emerging cases.

Our framework of respect is a pathway towards what could ultimately guide organizational policy, engineering best practices, and government regulation. When building systems that interact with humans and their data, a first evaluation should be to ask, "What is the conception of the user in this system?" Namely, "Does our system treat the persons involved as ends in themselves, or purely as means to the system's objectives?" If we cannot meaningfully answer these questions, then we may be designing a system that unethically instrumentalizes and dehumanizes its users.

Researchers are already undertaking the challenge of quantifying and systematizing definitions of "fairness", "accountability", and "transparency". We believe our rubric may aid in helping practitioners know when they should be applying these tools. For example, our framework clearly distinguishes between systems whose objectives are normatively contentious, insofar as their actions impinge on their users' rights and autonomy (e.g., passive data capture mechanisms) and systems whose objectives are normatively unproblematic, insofar as their actions do not violate their users' rights or autonomy (e.g., screening medical imagery for cancer).

We encourage researchers to further clarify and expand the kinds and strata of machine action we have proposed. Many of these cases, particularly the iterative stratum, involve complex social dynamics that will need more robust definitions as new cases emerge. There is also work to be done in designing better protocols for transparency and systems analysis in terms of the interpersonal dynamics promoted for affiliated users. As we have seen with fairness, once we can build robust conceptions of what ethical treatment looks like, we can move on to quantifying and implementing approaches. One major recommendation of our framework is that future systems should also be designed to give users a voice, by, for example, receiving live feedback from users and incorporating this feedback into future actions.

Admittedly, this paper does not solve the ethical problems we've highlighted; but this was not our aim. Our goal was to clarify our intuitions around how machines may curtail the rights and freedoms that we all see as ethically fundamental.

We hope readers walk away with a better understanding of the kinds of questions they should be asking of intelligent machine systems and an improved comprehension of the ethical space that machine actions occupy in human life. Given our current discourses around rights and regulatory mechanisms to protect those rights, we believe a framework like ours can take us further towards a systematic approach to the ethics and regulation of machine action.

REFERENCES

[1] 2015. Google apologises for Photos app's racist blunder. *BBC News* (July 2015). http://www.bbc.com/news/technology-33347866

[2] 2017. VIZIO to Pay $2.2 Million to FTC, State of New Jersey to Settle Charges It Collected Viewing Histories on 11 Million Smart Televisions without Users's Consent. (Feb. 2017). https://www.ftc.gov/news-events/press-releases/2017/02/vizio-pay-22-million-ftc-state-new-jersey-settle-charges-it

[3] Julius Adebayo and Lalana Kagal. 2016. Iterative Orthogonal Feature Projection for Diagnosing Bias in Black-Box Models. *arXiv:1611.04967* (2016).

[4] Philip E Agre. 1994. Surveillance and Capture: Two Models of Privacy. *The Information Society* 10.2 (1994), 101–127.

[5] Julia Angwin, Surya Mattu, Jeff Larson, and Lauren Kirchner. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. (May 2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[6] Solon Barocas. 2014. Data Mining and the Discourse on Discrimination. In *Data Ethics Workshop, Conference on Knowledge Discovery and Data Mining*.

[7] Solon Barocas and Andrew D. Selbst. 2014. *Big Data's Disparate Impact*. SSRN Scholarly Paper ID 2477899. Social Science Research Network, Rochester, NY. http://papers.ssrn.com/abstract=2477899

[8] Tom L. Beauchamp and James F. Childress. 1979/2012. *Principles of Biomedical Ethics*. Oxford University Press, Oxford.

[9] Stanley I. Benn. 1984. Privacy, freedom, and respect for persons. In *Philosophical Dimensions of Privacy: An Anthology*, Ferdinand David Schoeman (Ed.). Cambridge University Press, Cambridge, 223–244.

[10] Stanley I. Benn. 1988. *A Theory of Freedom*. Cambridge University Press, Cambridge.

[11] Johana Bhuiyan and Charlie Warzel. 2014. "God View": Uber Investigates Its Top New York Executive For Privacy Violations. (Nov. 2014). https://www.buzzfeed.com/johanabhuiyan/uber-is-investigating-its-top-new-york-executive-for-privacy

[12] Colin Bird. 2004. Status, Identity, and Respect. *Political THeory* 32, 2 (April 2004), 207–232.

[13] Sarah Bird, Solon Barocas, Kate Crawford, Fernando Diaz, and Hanna Wallach. 2016. *Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI*. SSRN Scholarly Paper ID 2846909. Social Science Research Network, Rochester, NY. https://papers.ssrn.com/abstract=2846909

[14] Sarah Buss. 1999. Respect for Persons. *Canadian Journal of Philosophy* 29, 4 (December 1999), 517–550.

[15] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv:1703.00056* (2017).

[16] Kate Crawford. 2013. The Hidden Biases in Big Data. (April 2013). https://hbr.org/2013/04/the-hidden-biases-in-big-data

[17] Kate Crawford and Jason Schultz. 2013. *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*. SSRN Scholarly Paper ID 2325784. Social Science Research Network, Rochester, NY. http://papers.ssrn.com/abstract=2325784

[18] Stephen Darwall. 2006. *The Second-Person Standpoint*. Harvard University Press, Cambridge, MA.

[19] Stephen L. Darwall. 1977. Two Kinds of Respect. *Ethics* 88, 1 (October 1977), 36–49.

[20] Robin S. Dillon. 2016. Respect. In *The Stanford Encyclopedia of Philosophy* (winter 2016 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[21] Alan Donagan. 1977. *The Theory of Morality*. University of Chicago Press, Chicago.

[22] R. S. Downie and Elizabeth Telfer. 1969. *Respect for persons*. Schocken Books, New York.

[23] Joel Feinberg. 1970. The nature and value of rights. *The Journal of Value Inquiry* 4, 4 (December 1970), 243–260.

[24] Philippa Foot. 1967. The Problem of Abortion and the Doctrine of the Double Effect. *Oxford Review* 5 (1967), 5–15.

[25] William K. Frankena. 1986. The Ethics of Respect for Persons. *Philosophical Topics* 14, 2 (1986), 149–167.

[26] Harry Frankfurt. 1997. Equality and Respect. *Social Research* 64, 1 (1997), 3–15.

[27] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. *arXiv:1609.07236* (Sept. 2016). http://arxiv.org/abs/1609.07236 arXiv: 1609.07236.

[28] Immanuel Kant. 1785/1998. *Groundwork of the Metaphysics of Morals*. Cambridge University Press, Cambridge.

[29] Christine M. Korsgaard. 1996. *Creating the Kingdom of Ends*. Cambridge University Press, Cambridge.

[30] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110, 15 (April 2013), 5802–5805. https://doi.org/10.1073/pnas.1218772110

[31] Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. 2016. *Accountable Algorithms*. SSRN Scholarly Paper ID 2765268. Social Science Research Network, Rochester, NY. http://papers.ssrn.com/abstract=2765268

[32] Lydia Manikonda and Munmun De Choudhury. 2017. Modeling and Understanding Visual Attributes of Mental Health Disclosures in Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 170–181.

[33] Lucia Moses. 2013. Marketers Should Take Note of When Women Feel Least Attractive. (Oct. 2013). http://www.adweek.com/brand-marketing/marketers-should-take-note-when-women-feel-least-attractive-152753/

[34] Martha C. Nussbaum. 1999. *Sex and Social Justice*. Oxford University Press, Oxford.

[35] The Nature of Respect. 1980. Stephen D. Hudson. *Social Theory and Practice* 6, 1 (1980), 69–90.

[36] Andrew G Reece and Christopher M Danforth. 2017. Instagram photos reveal predictive markers of depression. *EPJ Data Science* 6, 1 (Dec. 2017). https://doi.org/10.1140/epjds/s13688-017-0110-z

[37] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1135–1144.

[38] Olivia Solon. 2017. Facial recognition database used by FBI is out of control, House committee hears. *The Guardian* (March 2017). http://www.theguardian.com/technology/2017/mar/27/us-facial-recognition-database-fbi-drivers-licenses-passports

[39] Felix Stalder and Christine Mayer. 2009. The Second Index. Search Engines, Personalization and Surveillance. In *Deep Search: The Politics of Search Beyond Google*, Konrad Becker and Felix Stalder (Eds.). Transaction Publishers, London, 98–115.

[40] Charles Taylor. 1992. The Politics of Recognition. In *Multiculturalism: Examining the Politics of Recognition*. Princeton University Press, Princeton, 25–73.

[41] Sam Thielman. 2017. Your private medical data is for sale - and it's driving a business worth billions. *The Guardian* (Jan. 2017). https://www.theguardian.com/technology/2017/jan/10/medical-data-multibillion-dollar-business-report-warns

[42] Judith Jarvis Thomson. 1976. Killing, Letting Die, and The Trolley Problem. *The Monist* 59, 2 (April 1976), 204–217.

[43] Zeynep Tufekci. 2015. Algorithmic Harms beyond Facebook and Google: Emergent Challenges of Computational Agency. *Colorado Technology Law Journal* 13.2 (2015), 203–218.

[44] Berk Ustun and Cynthia Rudin. 2016. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning* 102, 3 (2016), 349–391.

[45] Daisuke Wakabayashi. 2017. As Google Fights Fake News, Voices on the Margins Raise Alarm. *The New York Times* (Sept. 2017). https://www.nytimes.com/2017/09/26/technology/google-search-bias-claims.html

[46] Bernard Williams. 1973. The idea of equality. In *Problems of the Self: Philosophical Papers 1956–1972*. Cambridge University Press, Cambridge, 230–249.

[47] Allen W. Wood. 1999. *Kant's Ethical Thought*. Cambridge University Press, Cambridge.

[48] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2016. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. *arXiv:1610.08452* (Oct. 2016). http://arxiv.org/abs/1610.08452 arXiv: 1610.08452.

[49] Jinyan Zang, Krysta Dummit, James Graves, Paul Lisker, and Latanya Sweeney. 2015. Who knows what about me? A survey of behind the scenes personal data sharing to third parties by mobile apps. *Proceeding of Technology Science* (2015).