Original Research Article

Assessing biases, relaxing moralism: On ground-truthing practices in machine learning design and application

Florian Jaton

Abstract
This theoretical paper considers the morality of machine learning algorithms and systems in the light of the biases that ground their correctness. It begins by presenting biases not as a priori negative entities but as contingent external referents—often gathered in benchmarked repositories called ground-truth datasets—that define what needs to be learned and allow for performance measures. I then argue that ground-truth datasets and their concomitant practices—that fundamentally involve establishing biases to enable learning procedures—can be described by their respective morality, here defined as the more or less accounted experience of hesitation when faced with what pragmatist philosopher William James called "genuine options"—that is, choices to be made in the heat of the moment that engage different possible futures. I then stress three constitutive dimensions of this pragmatist morality, as far as ground-truthing practices are concerned: (I) the definition of the problem to be solved (problematization), (II) the identification of the data to be collected and set up (databasing), and (III) the qualification of the targets to be learned (labeling). I finally suggest that this three-dimensional conceptual space can be used to map machine learning algorithmic projects in terms of the morality of their respective and constitutive ground-truthing practices. Such techno-moral graphs may, in turn, serve as equipment for greater governance of machine learning algorithms and systems.

I cannot understand regret without the admission of real, genuine possibilities in the world. Only *then* is it other than a mockery to feel, after we have failed to do our best, that an irreparable opportunity is gone from the universe, the loss of which it must forever after mourn. (James, 1912: 176)

Fragility is not the opposite of solidity, duration or solemnity of things, it is not on our margins, it is neither a defect to be repaired nor a temporary state, it is our common fate. (Hennion and Monnin, 2020: 1. My translation)

Introduction

Machine learning (ML) algorithms—computerized methods of calculation that infer rules of computation from sets of data to make predictions and support decision-making tasks—are now powering many commonly used devices such as Web search engines (Richardson et al., 2006), social media applications (Hazelwood et al., 2018), online purchasing platforms (Portugal et al., 2018), and surveillance systems (Chokshi, 2019). In reaction to the growing ubiquity of these statistical methods of computation—that have greatly participated in the resurrection of artificial intelligence (AI)—scholars in *Science and Technology Studies* (STS)[1] have accounted for some of their constitutive relationships (Bechmann and Bowker, 2019; Crawford, 2021; Grosman and Reigeluth, 2019; Jaton, 2017, 2019, 2021; Neyland, 2019). By providing fine-grained depictions of ML algorithmic systems, these works have effectively acted as provisional

STS Lab, University of Lausanne, Lausanne, Switzerland

Corresponding author:
Florian Jaton, STS Lab, University of Lausanne, Lausanne, Switzerland.
Email: florian@florian-jaton.com

countermeasures to the promotional rhetoric of AI over-enthusiasts and provided seminal means for greater governance of algorithmic systems (Radfar, 2019; Shellenbarger, 2019).

Among the issues these studies helped to bring to light, the problem of *biases*—unquestioned and contingent sociocultural habits that orient output calculations—has certainly received the most attention. From the European Commission (AI HLEG, 2019) to IBM (McDade and Testman, 2019), and McKinsey (Silberg and Manyika, 2019), the so-called *AI bias problem*—generally associated with ethics and morality (Mittelstadt et al., 2016)—is now one of the most frequently discussed topics. Although salutary in many respects, this rush toward the issue of AI bias has led to some confusion, prompting several authors to take steps toward clarifying the situation. What are biases? How can one spot them? Should they be stamped out? Under the threat of an AI ethics-washing (Wagner, 2018) that allows powerful industrials to take refuge in intellectual vagueness, it seems more important than ever to analyze the elements at stake and specify the objects of debate. In the wake of recent efforts made by Barocas et al. (2017) and Mittelstadt (2019), this paper contributes to providing conceptual tools capable of further refining the notion of bias and making it somewhat more operational.

To do so, this paper begins by introducing a positive view on biases. Instead of considering them as intrinsically deleterious, it appreciates biases as necessary, yet contingent, external referents. Often gathered in repositories called *ground-truth datasets* (Grosman and Reigeluth, 2019; Henriksen and Bechmann, 2020; Jaton, 2017), these constructed external referents operate as supervisors of learning processes: They define what needs to be learned and allow for performance measures. The paper then shows that these supervising biases concern a wide range of ML algorithms: As recent studies indicate, computer scientists have to confront—and be biased by—ground-truth datasets while shaping and implementing supervised *and* unsupervised ML algorithms. I then argue that these *ground-truthing practices*—that fundamentally involve establishing biases to enable learning procedures—can be described by their respective morality, here defined, in the wake of pragmatist philosopher William James, as the more or less accounted experience of hesitation when faced with "genuine options" (James, 1912)—that is, choices to be made in the heat of the moment that engage different possible futures. I then stress three constitutive dimensions of this pragmatist morality, as far as ground-truthing practices are concerned: (I) the definition of the problem to be solved (problematization), (II) the identification of the data to be collected and set up (databasing), and (III) the qualification of the

targets to be learned (labeling). I finally suggest that this three-dimensional (3D) conceptual space can be used to read ML algorithmic projects in terms of the morality of their respective and constitutive ground-truthing practices. Such *techno-moral graphs* may, in turn, serve as equipment for greater governance of AI systems. In the conclusion, I briefly expand on the outlined propositions.

Learning to be biased: On the centrality of ground truths

Examples of what Noble (2018) coined *algorithmic oppression* abound: search engines that marginalize women (Carpenter, 2015); health prediction algorithms that consider Black patients riskier than White patients (Obermeyer et al., 2019); recommendation algorithms favoring violent, racist, and misogynist content (Hao, 2019); and crime prediction systems whose scores are more influenced by skin color than by criminal record entries (Angwin et al., 2016). These harmful *biases* must be fiercely criticized and combated. In order to do so, they must be identified through insightful statistical inquiries (Courtland, 2018) but also, which tends to be taken for granted, through the affirmation of universal moral precepts such as fairness and equality. While such principles are not easy to describe rigorously (Verma and Rubin, 2018), it is still possible to roughly outline them and agree on the worldviews they suggest.

This moralism—in the sense of a confident attitude toward the moral values to be defended—is crucial today, not least in order to infuse political affections into the thematic universe of ML algorithms that, not so long ago, was still confined to mere technical considerations (Jaton and Vinck, submitted). However, framing the still-to-be-fought algorithmic oppression mainly in terms of *bias* (in singular) runs the risk of not being taken seriously, or at least, as suggested by Manders-Huits and Zimmer (2009), of not being invested with meaning by some of the people who may be very much concerned with the issue: computer scientists who work every day, in academia and industry, to shape new ML algorithmic tools.

To understand this risk of expert disbelief in ML bias as this notion is described in the critical literature on algorithms, it is needful to turn to another, older discourse that has progressively become quite inaudible outside the spheres of computer science. This classical and authoritative view comes from Tom Mitchell's pioneering work on statistical learning. As early as 1980, he showed that biases—understood as external and arbitrary sources of information[2]—are necessary for the inductive leap underlying any learning process.

This specific instance of what Bowker and Star (2000: 111–123) call the *bootstrapping problem* led Mitchell to argue, in turn, that an unbiased learning algorithm would be senseless:

> Although removing all biases from a generalization system may seem to be a desirable goal, in fact the result is nearly useless. An unbiased learning system's ability to classify new instances is no better than if it simply stored all the training instances and performed a lookup when asked to classify a subsequent instance. (Mitchell, 1980: 2)

In short, no biases, no learning: As Domingos recently summarized: "in ordinary life, bias is a pejorative word: preconceived notions are bad. However, in machine learning, preconceived notions are indispensable; you can't learn without them" (Domingos, 2015: 64).

Put crudely, then, biases are crucial for ML: Any classification task needs a referent that lies outside of the task in order to ground its classificatory principle. The centrality of *ground-truth datasets* (Figure 1) for the training and evaluation of new ML algorithms is a striking illustration of this necessity: Without benchmarked databases that provide the referents of what ML algorithms must find (Henriksen and Bechmann, 2020), no learning operation is conceivable since there is no a priori indication of what is appropriate to learn. Removing all biases from ML algorithms—as it is sometimes suggested (Gibney, 2020)—would therefore be tantamount to removing central parts of what allowed them to come into existence, namely the contingent, yet necessary, external referents that operate as their initial impetus. For the specific case of ML algorithms and systems, morality seems then obliged to deal with this state of affairs visible as soon as one walks through the door of a computer science laboratory (Jaton, 2021): A bias-free ML algorithm is an oxymoron.

The supervision of the "unsupervised": From ground truths to ground-truth*ing*

Before moving forward and further examining the issue of morality with regard to the ground truths that bias—and enable—learning operations, it is important to look more precisely at another technical discussion related to ML algorithms. Papers and manuals on statistical learning methods are extremely numerous and varied. However, in this innovative and constantly changing nebula, one notion remains stable: that of *supervision* (and its opposite, *unsupervision*). With the possible exception of the category of "reinforcement learning"—which I will not discuss in this paper—computerized methods of calculation inferring classification or regression rules from aggregated data—what I refer to here as *ML algorithms*—are indeed divided, in the specialized literature, in two main families: supervised and unsupervised. This is a widely shared, standard statement: Supervised
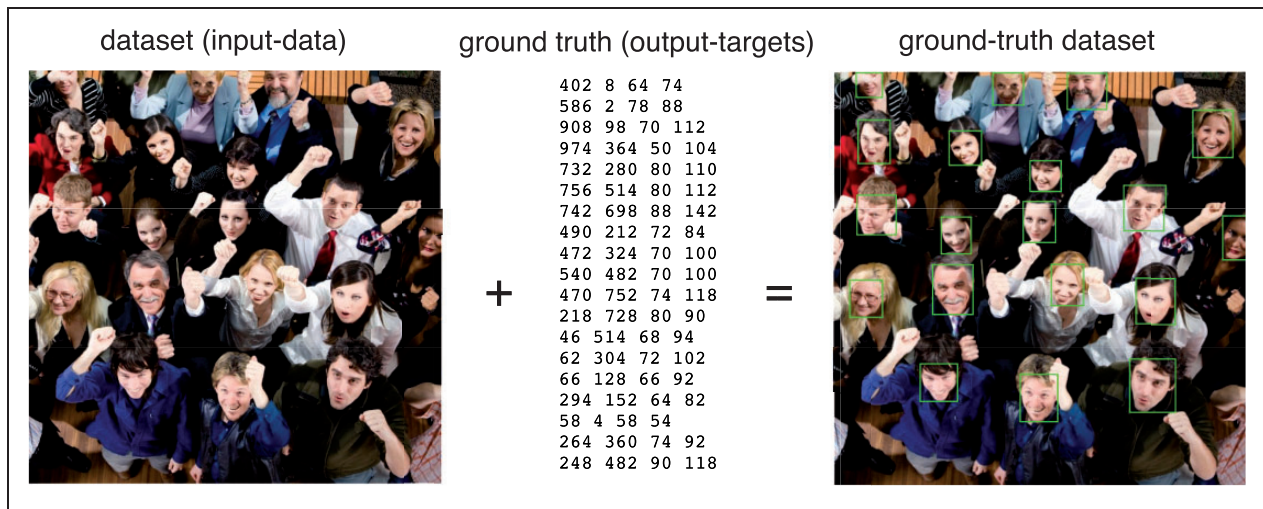

Figure 1. Taken from Yang et al. (2016), sample from WIDER FACE ground-truth dataset. On the left, one among the 32,203 images of the publicly available dataset for face-detection research. In the middle, the face annotations for this specific image. Since each annotation belongs to the coordinate space of the digital image, it can be expressed by a set of four numerical values, the first two expressing the start position of the label along the *x* and *y* axes, the third one expressing the number of pixel wide, the fourth one expressing the number of pixels high. This numerical information, that correspond to the bounding boxes of the image on the right, were produced manually by one human annotator and cross-checked by two others (Yang et al., 2016: 5527). As such, they constitute the ground truth of the image with regard to face detection; they bias the input data in order to provide something to learn and formulate. The images and their labels can then be used to train supervised ML algorithms to recognize faces in photos, the ground-truth dataset operating as the list of the very best answers to such task.

algorithms for which an external supervisor "provides the correct values, and the parameters of a model are updated so that its output gets as close as possible to these desired outputs" are fundamentally different from unsupervised algorithms that do not require any supervisor and whose purpose is "to find the regularities in the input, to see what normally happens" (Alpaydin, 2016: 112). To put it in equivalent terms, while supervised algorithms need a ground-truth dataset gathering input data and manually constructed output targets in order to learn the predictive rules of computation, unsupervised algorithms rely *only* on input data to detect patterns and regularities.
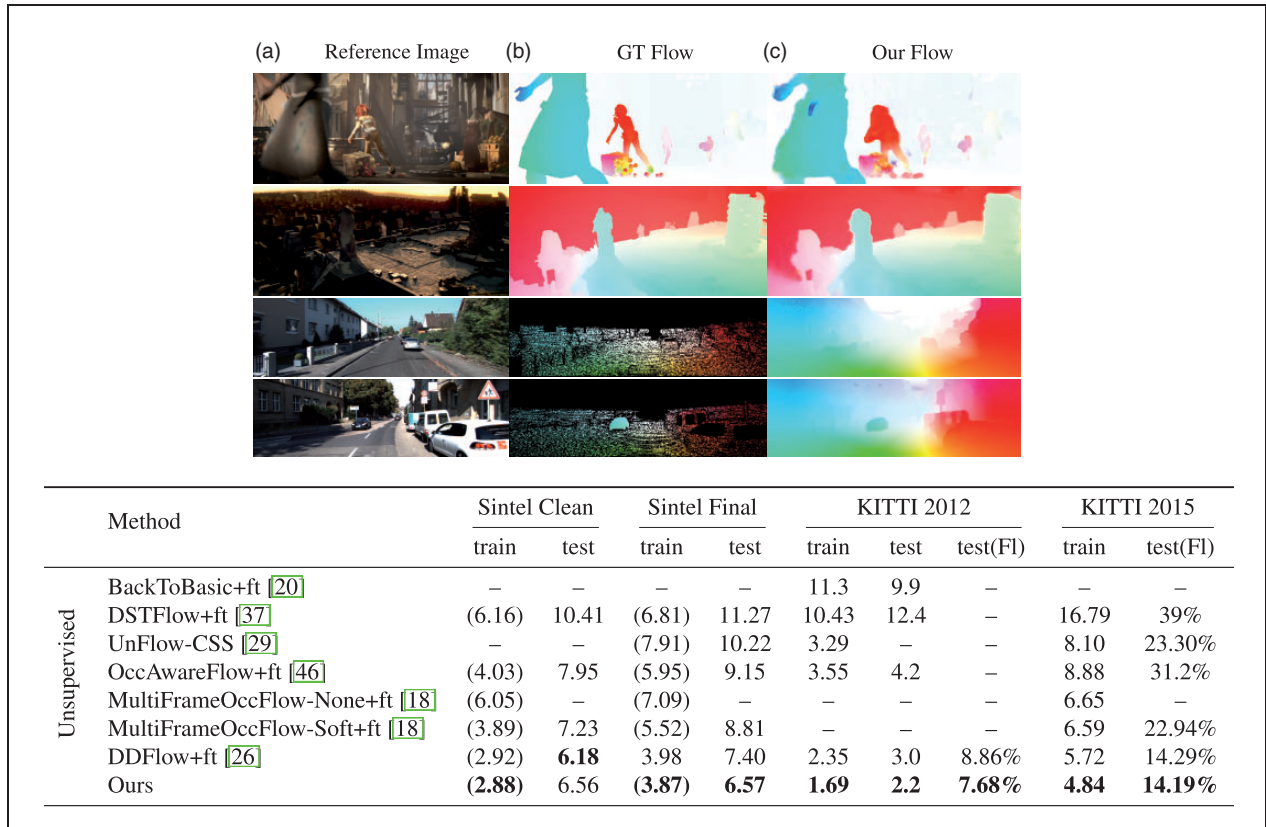
For academic and industrial researchers who recognize this fundamental distinction (i.e., the vast majority), further development of unsupervised ML algorithms carries with it great hope since these algorithms do not depend, theoretically, on any external supervision of the input data. Substantial gains in time, resources, and purity are explicitly envisaged: Because unsupervised ML algorithms use only the information contained in their "raw" learning data, they would not be concerned with the formation of costly ground-truth datasets whose labels are tedious to produce and potentially influenced by the sociocultural habits of their human generators and curators. The sheer enthusiasm for unsupervised ML algorithms is evidenced in the now famous "cake analogy" proposed by Yann LeCun, one of the main initiators of convolutional neural networks and the winner of the prestigious Turing Award in 2018 (with Yoshua Bengio and Geoffrey Hinton), where "the bulk of the cake is unsupervised learning, the icing on the cake is supervised learning, and the cherry on the cake is reinforcement learning" (LeCun, 2016).[3]

It would be incorrect to assert that this distinction between supervised and unsupervised ML algorithms is erroneous: When considered from a confined, theoretical perspective, unsupervised ML algorithms are not bounded to ground-truth datasets gathering input data and output targets, whereas supervised ML algorithms are. However, when considered "in the wild" (Hutchins, 1995), which is from a down-to-earth perspective, one realizes that the story is more intricate: It is indeed attested that ground truths and their concomitant referential practices do in fact impact, albeit in a less visible way, the practical shaping and use of ML algorithms presented as "unsupervised."

The first way to consider the subtle attachment of unsupervised ML algorithms to ground truths is simply to read recent award-winning papers presenting new unsupervised ML algorithms, such as Wan et al. (2019), Lorenz et al. (2019), Fu et al. (2019), and Liu et al. (2019).[4] Although these (quite) arbitrarily selected papers are all related to computer vision and image

processing, each deals with a different problem: 3D hand pose estimation for Wan et al., part-based disentangling of photographed objects for Lorenz et al., image translation from one domain (e.g., natural photograph) to another (e.g., painting) for Fu et al., and optical flow estimation for Liu et al. Moreover, each paper attempts to convince readers of the relevance and efficiency of its unsupervised ML algorithm. Also, in the specific evidence-production regime of applied computer science, the acceptable way to do so is to rely on a ground-truth dataset—also called a benchmarked dataset—containing human-produced labels and operating as a measurement reference capable of generating statistical results (see Figure 2). For each paper, the very topic of the computational operation is framed by, and dependent on, the availability of a ground truth previously constructed and used by other groups of researchers to develop and compare supervised ML algorithms.[5] It thus appears that even though these unsupervised ML algorithms do not rely upon any labeled ground truth for their learning tasks, they need available ground-truth datasets to attest to the significance of their results. Rather than a technical necessity, this is a practical imperative: Without referring to a ground truth operating as a yardstick between competing algorithms, the aforementioned researchers—but also, I believe, many others—cannot quantitatively measure the performances of their algorithms according to the standard statistical measures and cannot, therefore, make their algorithm exist within a searchable and quotable paper. This practical imperative is linked to the fact that unsupervised ML algorithms are not intended to remain theoretical: They are designed to be ultimately used and worked upon, which implies comparing them to benchmarked ground-truth datasets in order to show their relevance and efficiency. And if ground truths—together with their *labels* (and their biases)—are not necessary for the definition of the algorithms' learning functions, they remain essential to make them exist as devices producing valuable results.

Another reason why unsupervised ML algorithms remain attached to, and biased by, supervised "truths" can be understood by exploring the backstage of computer science work. Using the genre of autoethnography in an interdisciplinary data analysis laboratory, Bechmann and Bowker (2019) documented the inconspicuous supervisory operations involved while applying an unsupervised algorithm (the Latent Dirichlet Allocation model Text2vec) to Facebook user data. As they made clear, an irremediable succession of arbitrary (but justifiable) choices were necessary to produce results that made sense by virtue of the research question, itself reworked as the multilateral relationships between the collected survey data, the provisional results produced by the algorithm, and the interpretive

Figure 2. Two evaluations of Liu et al.'s (2019) unsupervised ML algorithm (ours). On top, qualitative evaluations of Liu et al.'s algorithm with respect to *Sintel* (Butler et al., 2012) and *KITTI* (Geiger et al., 2012; Menze and Geiger, 2015) ground-truth datasets for flow estimation. On bottom, quantitative comparisons between the performances of Liu et al.'s algorithm (ours) and previously-published algorithms with respect to *Sintel* and *KITTI* ground-truth datasets. The main performance metrics for these datasets is the average endpoint error: the overall comparison between the estimated optical flow vectors provided by the algorithm and those provided by the ground truth. *KITTI 2012* and its augmented version *KITTI 2015* also include the percentage of erroneous pixels of the algorithms' estimations. Except on the *Sintel Clean* test set, Liu et al.'s unsupervised algorithm outperforms all the others. The outperforming results are highlighted in bold. Source: Taken from Liu et al. (2019: 4–5), reproduced with the permission of the IEEE, 31 March 2021, 5039350523845, Pengpeng Liu, June 2019.

work unfold. Bechmann and Bowker convincingly showed, in turn, that in order to make the unsupervised algorithm operational, it was crucial to *supervise* its end-to-end deployment—that is, to refer to elements a priori external to the algorithmic process per se (e.g., their own values, doubts, disappointments, ambitions) in order to ground its correctness. As they sum it up: "a seemingly unsupervised model becomes extremely supervised due to classification work such as setting number of topics, cleaning data in a particular way with an a priori understanding of 'meaningful' clusters and interpreting clusters with parent classes manually" (Bechmann and Bowker, 2019: 7).

This second occurrence of biased supervision within the deployment of unsupervised ML algorithms suggests the need to somewhat extend the notion of ground truth. Indeed, in the case study of Bechmann and Bowker, a ground-truth dataset had not really been involved because the developers did not directly

refer to assumedly correct, labeled responses listed in a benchmarked dataset as the authors of the image-processing papers mentioned earlier did. Yet, in effect, "truths" have been grounded—and biases established—because the researchers did refer to external sources of information (e.g., research questions, interpretations of results) that, in the end, attested to the correctness of the applied algorithm: Without their supervision work, the researchers could not make the "unsupervised" ML algorithm produce results useful to their research. Here, turning the notion of ground truth into a gerund seems somewhat necessary: Since the design and application of an unsupervised learning algorithm must, apparently, refer to elements that lie outside of its own functioning to effectively support and prove its efficiency and correctness, one should rather talk about *ground-truthing* rather than ground truths. If one sticks to the noun form, great are the risks to invisibilize the subtle grounding practices

taking place during the design and implementation of seemingly unsupervised algorithms.

At the risk of proposing a redundant expression—but, sometimes, a little redundancy does not hurt—I propose calling *ground-truthing practices* the heterogeneous courses of actions aimed at attesting to the correctness of a computerized method of calculation.[6] Such practical biasing operations may range from the early problematization of an algorithmic project to the arbitrary selection of the relevant data and the actual construction, publication, and use of benchmarked ground-truth datasets (more on this later). Although these practices do not, by far, cover the whole process of algorithmic design—moments such as the mathematical characterization of the relationships between input data and output targets or the actual writing of computer code refer, for example, to qualitatively different processes (Jaton, 2021)—they nonetheless represent a non-negligible part of it. More than just setting up referential repositories, ground-truthing practices also support and make possible the very correctness of computerized methods of calculation.

Morality as collective hesitation

Both supervised *and* unsupervised ML algorithms are attached to ground-truthing practices in order to be shaped, published, and applied in real-world situations. Without practical efforts to attest to ML algorithms' correctness and, therefore, temporarily move away from their technical functioning to establish biases and ground their efficiency, most ML algorithms could not exist and not a single one could be effectively used. Ground-truthing practices and the many biases they allow to be established are part and parcel with the constitution of ML algorithms: They contribute to making them designable, commensurable, and even, sometimes, efficient and useful.

By including this small realistic modality—ML algorithms need constructed and more or less contingent referential biases to show their correctness and come into existence—another landscape soon unfolds. Instead of irresistible ventures confidently inserting the depths of human cognition into digital devices, computer science industry, and, more particularly, its applied subdomains such as computer vision and image processing, start to appear radically fragile and uncertain. Indeed, a single change in the ground-truthing practices underlying the shaping and use of an ML algorithm could be enough to significantly modify its tenor. For the Facebook data analysis project considered by Bechmann and Bowker, a somewhat different problematization of the research effort would have led to quite different algorithmic results. The same is true for the work of the computer scientists mentioned above: A single change in the ground-truth datasets used to prove their results (e.g., a different question asked to the human annotators who labeled the collected data, different choices regarding the extraction and curation of the labels) would have led to the development of different unsupervised ML algorithms since they would have had to confront different ground truths. By stressing the ground-truthing practices underlying the shaping and use of ML algorithms, one highlights a trivial but often forgotten feature of computer science research and industry: It could be otherwise.

Hence, occasionally, a certain surprise on reading accounts of the shaping or application of ML algorithms. Even though what underlies the correctness of these models—which triumphantly permeate our daily lives—is often the result of contingent and arbitrary processes, few authors publicly admit to this constituent fragility and document the ins and outs of the alternatives that have been, at some point, available to them (Geiger et al., 2020). To put it in another, more philosophical way, while the ground-truthing work subtending ML algorithms is punctuated by irremediable choices, relatively few accounts explicitly attest to what pragmatist philosopher William James (1912) classically called "genuine options"—that is, choices to be made in the heat of the moment that engage different possible futures or, in our case, different possible learning-enabling biases. What happens in these ground-truthing moments is indeed decisive, for as the biases implemented by computer experts will strongly frame what will later become, perhaps, a consulted ML model easily enrollable in broader corporate systems.[7] Could the morality of ML lie, at least in part, in these genuine options, moments when expert hands may shiver at the idea of grounding a "truth" that will then be reproduced, and thus promoted, computationally?

It is one of the many merits of James' philosophical work—and his contemporary interpretations (Hache, 2011; Hennion and Monnin, 2020)—to have detected the moral tonality of these moments of hesitation where actors are attentive to the fragility of what they are accomplishing. Contrary to the Kantian tradition which considers morality as the result of a judgment subtended by a universal law,[8] the pragmatist tradition considers morality as the temporary experience of, and inquiry into, concerns and scruples. The turning around is thus complete: Instead of considering morality as compliance with a transcendent norm, morality is considered as the follow-up on the spark of a genuine option, triggering in turn an investigation of the ins and outs of a situation of uncertainty. During pragmatist moral experiences, what was initially considered a simple *means* (e.g., a crowdsourcing

company, a ground-truth dataset, a research direction, an evaluation metrics) is transformed, temporarily, into an *end* whose trajectory depends on many other intertwined entities. And it is the uncertain exploration, long or short, of the connections between these entities that constitutes the specific signature of morality.

Considering morality as what is happening when there is an investigation on the fragilities and uncertainties of a genuine option that sparks may sound odd at first, but it effectively addresses a mundane, increasingly common experience: *that car* I drive, *that meat* I grill, or *that search engine* I use; these a priori unproblematic entities become frantically animated as my scruple sets in and my investigation unfolds. *That car* soon connects needs for mobility with workers in Northern France and birds stuck in oil spills; *that meat* soon connects cliché summer partying with traceability networks and fertilizers for fodder fields; and *that search engine* soon connects urgent desires for access to certified references with new forms of alienations subtending digital capitalism: As soon as a scrupulous doubt as to a means turns it into an end, a relational experience is set in motion that brings about many intertwined human and nonhuman entities. As Hache pointed out, this problematization "engages a conception of relational morality in which one cannot be moral on one's own" (Hache, 2011: 52. My translation); the intermingling operated during the more or less long moments of hesitant investigation turns moral experiences into collective enterprises.

From this, it follows that morality—in its pragmatist understanding—can blossom in some settings and wither away in some others. Certain arrangements can favor great moral development by valuing the expression of doubt and the hesitant exploration of scruples, to the point of even instituting them—sometimes—as a working habit.[9] Conversely, other arrangements can, voluntarily or not, repress moral development, thus making what Latour calls the "emission of morality" (Latour, 2012: 454) inaudible. In these constricted settings, scruples are stifled; hesitations as to the distinction between means and ends are ignored; genuine options are pale glimmers, practically invisible and incapable of suggesting anxious inquiries.

Pragmatist-inspired morality is thus what is happening when there is a collective exploration of the fragilities and uncertainties of a genuine option that sparks. During moral moments—whose development can be sustained or repressed—means are temporarily turned into problematic ends through concerned and hesitant collective inquiries. From there, if we return to ML algorithms and take the point of view of those who work every day to build new robust and innovative ones (who may be different from those who talk about ML and AI during keynotes and distinguished lectures), two opposite ways of experiencing ground-truthing practices emerge. Either ignore (or keep quiet) the genuine options that dot ground-truthing practices and consider their constituent elements as unproblematic *means* for the completion of the algorithmic project; or, at the other end of the spectrum, be systematically sensitive to the scruples and uncertainties. Variable intensities are, of course, conceivable, but it seems a priori fair to posit that ML designers go through differentiated *moral experiences* that could be summarized as follows: being more or less eager (or encouraged) to confront and unfold the fragility of ground-truthing practices; being more or less eager (or encouraged) to respond to what they are bound to while building referential bases for the proper shaping and deployment of ML algorithms.

By virtue of the perspective adopted in this paper, ground-truthing practices (and the biases they establish) contributing to ML algorithmic projects are then not always morally equivalent: Some are more sensitive to the irruption of genuine options and the exploration of their underlying uncertainties than others (see Figures 3 and 4). Also, without suspecting anyone of negligence, one may consider that many current ML projects are not especially moral in the sense defined here. There are numerous reasons for this, and the race for valuable innovation and publication is certainly part of the phenomenon (Mirowski, 2011). However, the sheer act of attributing the adjective "unsupervised" to highly supervised—and thus biased—algorithms may illustrate a lack of sensitivity to, and emphasis on, the hesitation between the means and ends of many contemporary ML developments (Geiger et al., 2020).

However, more and more organizations are beginning to show, through their actions, that the moral issue of algorithms is an integral part of their concerns. By opening their doors to sociologists, philosophers, journalists, anthropologists, or ethnographers and, in particular, encouraging them to document the practices by which "truths" are instituted in order to establish the correctness of algorithms being shaped and used, the data analysis laboratory mentioned by Bechmann and Bowker (2019), the image-processing laboratory followed by Jaton (2017, 2021), the European automatic surveillance projects studied by Grosman and Reigeluth (2019) and Neyland (2019), or the Scandinavian AI firm investigated by Henriksen and Bechmann (2020) show, for example, a genuine desire for morality, understood as a propensity to make more explicit, and therefore real, the exploratory hesitations and doubts contributing to algorithmic projects. The current situation seems then to be quite mixed: Whereas many algorithm-related organizations,
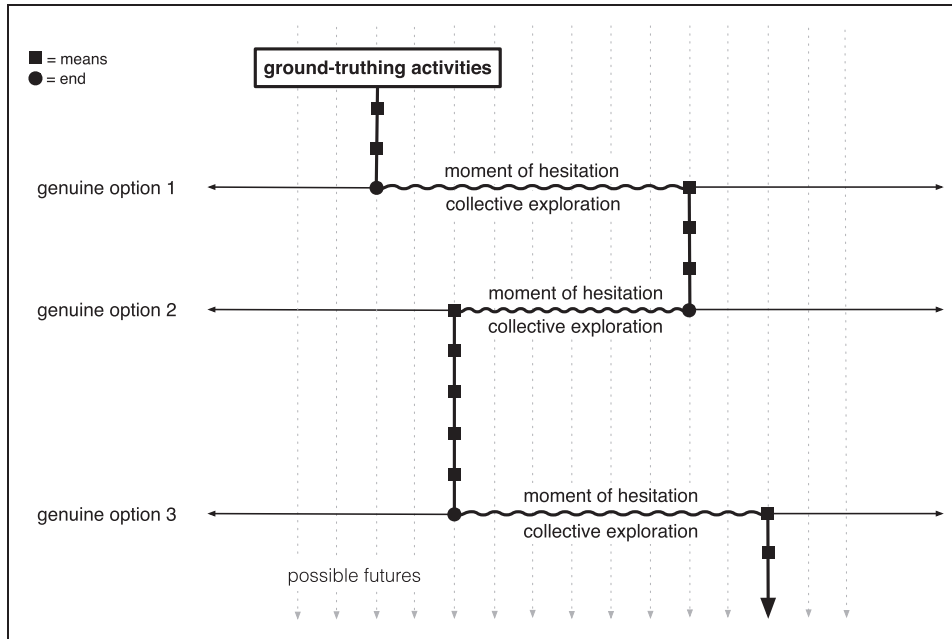
Figure 3. Schematics of ground-truthing practices sensitive to encounters with genuine options. Let us imagine ground-truthing courses of action using means (e.g., benchmarked ground-truth datasets, evaluation metrics) as part of an ML algorithmic project. When doubts arise regarding a means (genuine option 1), this scrupulous hesitation temporarily shifts the project to exploring the issues underlying the option, considering and, eventually, rejecting (or embracing) sets of possible futures (vertical dotted arrows). Once the investigation has been provisionally completed—making the end become, temporarily, an acceptable means—the design decision as to the genuine option moves the project forward to, potentially, another genuine option (genuine option 2) whose fragility triggers, in turn, another collective exploration of the whys and wherefores of the uncertain situation.
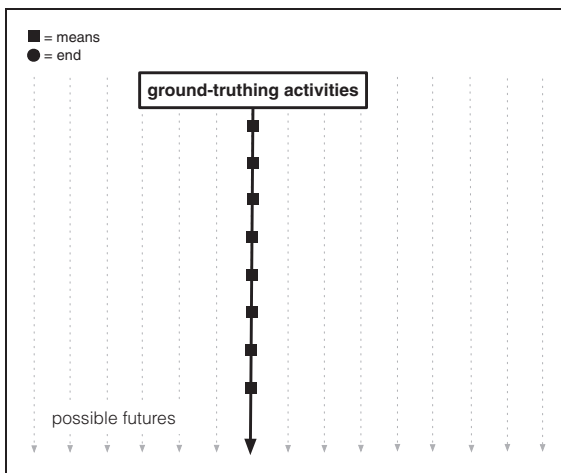

Figure 4. Schematics of ground-truthing practices not sensitive to encounters with genuine options. In this case, the ground-truthing courses of action have not encountered/made appear any genuine options. No form of hesitation has turned a means into an end. The progress is straight-line and does not expressively take into account alternative futures.

especially the most powerful ones, are frankly reluctant to make hesitations and uncertainties visible and thus favor the modernist path of inevitable mastery (Pasquale, 2016), others, fewer in number but

nevertheless increasingly present, are ready to make the morality of the devices they build positive. By inviting lay actors to co-investigate with computer science professionals on algorithmic work *cases*, these ecological institutions—in the sense of institutions sensitive to networks of interdependencies (Latour, 2017)—agree to be accountable for some of their actions and, thus, accept to become more response-able. This is something important to point out: In computer science and industry, morality is nowadays actively opposed but also supported (albeit still tentatively).

Toward techno-moral graphs: Problems, data, and labels

If we accept to consider morality as the act of responding to what Latour (2004: 216) calls "the generalized revolt of means" by embarking on collective investigations of genuine options—and thus distinguish it from moralism understood as the important, yet sometimes limited, injunction to observe universal moral precepts—it still remains, in our case, to somewhat specify the environment in which this morality can be deployed. At this point, the term "ground-truthing practices" remains too vague to hope to detect moral

differentials with regard to the contingent, yet crucial, biases that make ML algorithms possible and usable. How can one see more clearly in this imbroglio that I call "ground-truthing practices"?

One may start by pointing out something that is trivial in hindsight: The problem an ML algorithm solves is not a priori given; it is the result of *problematization practices* aimed at establishing the terms of a problem that can be solved computationally. The institutional, and contingent, definition of algorithms as "computerized problem-solving methods" may have contributed to putting aside the basic fact that problems that aim to be solved by algorithms are temporary results of processes engaging habits, desires, skills, and values (Jaton and Vinck, submitted). As Lehr and Ohm nicely summarized, for ML algorithms whose tasks are to predict and estimate something, "the first step of any analysis is to define what that *something* should be and how it should be measured" (Lehr and Ohm, 2017: 672–673. Emphasis in the original).

A first dimension of ground-truthing practices may then refer to the courses of action that participate in problematizing a state of affairs. Here, the external referents, or biases, to be defined are the terms of the problem the algorithm will have to solve. The image-processing project followed by Jaton (2017) is illustrative in this respect. In order to launch their project of a new image-processing algorithm for saliency detection, the computer scientists had to start by critically examining the state-of-the-art literature, notably by equipping themselves with authoritative—and highly indexed—papers in philosophy and cognitive science, and by imagining new industrial applications based on their criticism. I shall group under the term "problematization" this type of practical work that can take many different forms but still consists, ultimately, in making a situation problematic and presenting the would-be algorithm as a solution potentially generating positive differences.

A second dimension specific to ground-truthing practices, and the biases they contribute to establishing, can be detected with another trivial observation: Algorithms, especially ML ones, need sets of data. This aspect of the preparatory work required for the shaping and use of ML algorithms was made particularly visible in recent years by the (temporary) advent of the term "Big Data" and, simultaneously, by more or less successful attempts to enforce data privacy. Unsurprisingly, then, it is in the fields of law and ethics that one finds the most refined explorations of the issues related to the massive online collection and organization of digital data to be transformed into "raw" materials for the shaping of ML algorithms and systems (Barocas and Selbst, 2016; Custers et al., 2012; Lehr and Ohm, 2017). Building upon these

important works to further specify the heterogeneity of ground-truthing practices, I shall call "databasing" the actual work of collecting, compiling, organizing, and cleaning the data to be used for the shaping and/or training of new ML algorithms. As its name suggests, this particular dimension of ground-truthing practices—which often takes place in parallel with problematization activities and impacts on them—cannot be reduced to the collection of data alone. It also includes choices and actions contributing to the aggregation, probing, organization, and cleaning—in short, the *setting up*—of what may be called at some point the "raw" data (Gitelman, 2013; Jaton and Vinck, 2016).

Finally, a third dimension specific to ground-truthing practices is to be found in the categories or labels (or output targets) that are superimposed on the collected "raw" input data by means of more or less standardized devices and procedures. As summarized by Grosman and Reigeluth, this refers to the

> "ground truth" part of the term "ground-truth dataset": the categories or labels which humans—for example, domain experts, computer scientists or Amazon turkers—have attributed to each sequence. It supplies the system with answers to the problem: the algorithm now has an external check for assessing the correctness of its classification. (Grosman and Reigeluth, 2019: 3)

The increasing availability of crowdsourcing services proposed by companies such as Amazon (via Amazon Mechanical Turk) or ClickWorker since 2010 had a considerable impact on the production capacity of labels and, as a result, on the proliferation of algorithms capable of retrieving and thus reproducing (and promoting) these labels.[10] Consequently, more than being strictly about computer science research, the production of labels also had, and has, a wider socio-economic impact—the often worrying ramifications of which have been well studied by Gray and Suri (2019) and Casilli (2019). Here, I shall use "labeling" as an umbrella term to designate the heterogeneous, but assignable, practices involved in defining, organizing, remunerating, and, sometimes, refining the labels that provide ML algorithms—supervised and unsupervised—with answers to the problems they try to solve. While this axis of ground-truthing practices might be, for the time being, the one with the most obvious ramifications to global socioeconomic issues, it does not unfold independently from the other two axes. Indeed, as suggested in Jaton (2021: 31–86), some algorithmic projects may start to organize the production of labels once their problem has been defined and their data collected and set up, other projects may start by considering the facilities available for the production of labels in order to define the terms of a problem

in a different way, and other projects may use available data to define a problem and the labels that can be used to solve it. Although referring to different practices, the three axes—problematization, databasing, and labeling—echo each other in ways that are specific to each individual ML endeavour.

If we combine the elements presented in this section, we find ourselves with, at least, three dimensions—or axes—on which, schematically, ground-truthing practices contributing to ML projects can be represented: a first axis for the problematization practices (e.g., criticizing previous research results, capitalizing on past achievements), a second axis for the databasing practices (e.g., web scrapping, distribution analyses), and a third axis for the labeling practices (e.g., designing a crowdsourcing task, implementing an available benchmarked ground truth). And the whole of this space, now delineated, constitutes the bulk of ground-truthing practices that are crucial for the shaping and use of ML algorithms, whether supervised or unsupervised (see Figure 5). If we now include the somewhat speculative (yet informed) elements of moral philosophy presented in the previous section, each axis of this 3D ground-truthing space becomes staggered by potential genuine options that, themselves, refer to potential explorative and collective hesitations between means and ends (Figure 6).

From there, the theoretical space of ground-truthing practices becomes a *visual and conceptual space* on
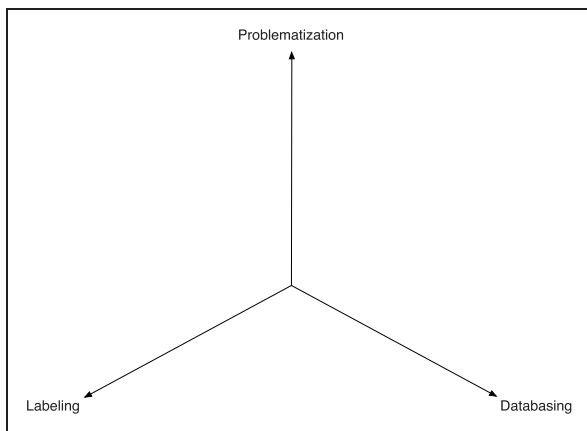
which specific courses of action can eventually be reported. This scriptural technology, rudimentary but refinable, could then support *maps* differentiating ML algorithm projects with regard to their ground-truthing practices (see Figure 7). These *techno-moral graphs*— that still need to be put to the test—would be a way to visualize the moral narratives of the ground-truthing practices contributing to ML projects—that is,
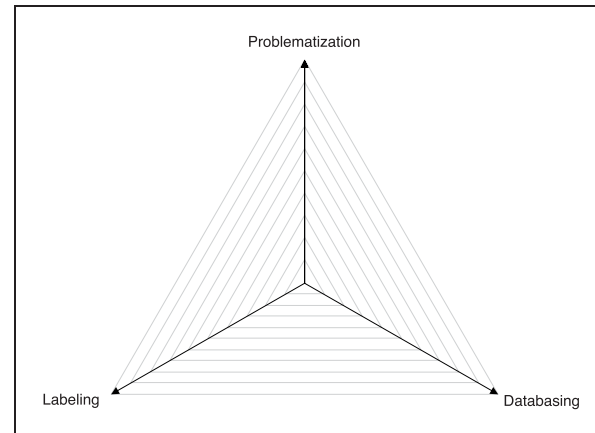
Figure 6. Schematics of the three dimensions of ground-truthing practices when staggered by the genuine options that, potentially, spot their deployment. Each intersection between lines and axes corresponds to a potential genuine option partaking the deployment of problematization, databasing, or labeling actions.
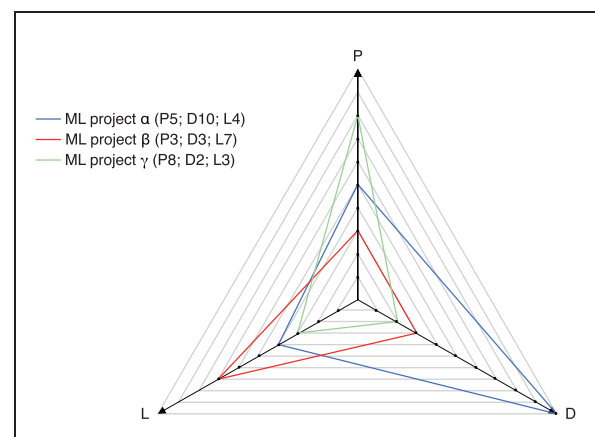
Figure 5. Schematics of the three dimensions of ground-truthing practices. On the top, the "problematization" dimension refers to the practices partaking the definition of the terms of a solvable problem. On the right, the "databasing" dimension refers to the practices partaking the aggregation, probing, organization, and cleaning of the data to be used for the shaping and/or training of a new algorithm. On the left, the "labeling" dimension refers to practices partaking the definition, organization, collection, and access to the labels that provide the new supervised or unsupervised ML algorithm with answers to the problem it should solve.

Figure 7. Techno-moral graph of three hypothetical ML projects. If, by convention, each encounter with a genuine option counts as 1, the addition of these attested meetings/explorations allows to report a value on one (or more) of the three axes *P* (problematization), *D* (databasing), and *L* (labeling). The pragmatist morality of each ML project—as far as its ground-truthing practices are concerned—could then be summarized by its more or less extended map on the coordinate system. The more accounted (and available) explorative hesitations on each of the axes (small black dots on the axes), the more morality.

narratives that are all the more voluminous as the morality maps of their related projects are large. Similar to the sociotechnical graphs introduced by Latour et al. (1992), techno-moral graphs would have no value by themselves: They could only make sense if they point to already existing documents and reports, notably, and above all, those accounting for the collective investigations underlying genuine options. This scriptural device, for the moment quite speculative, may operate as a reflexive instrument whose main aim

> like that of any other instrument, is to get rid of most of the initial information, while outlining the features that are deemed relevant to our inquiry, [and] offer a quick and easy comparative basis for many narratives coming from many sources. (Latour et al., 1992: 37)

In short, techno-moral graphs would be a way, among other possible ones, to visualize, make explicit, and compare some of the constitutive biases of ML algorithms.

Despite the apparent simplicity of these techno-moral graphs—they are, after all, only a superficial way of visualizing the quantities of assignable hesitations and scruples that have dotted the ground-truthing practices underlying specific ML projects—they do point to deep issues. First, techno-moral graphs suggest that, for the specific case of ML algorithms—entities that easily pervade our lives—morality must be supported by accessible inscriptions and writings: Without a tangible record of the emergence of a genuine option and its correlated exploration, there is no way to further attest to its existence. This denotes, in turn, the power of inscriptions and their centrality in the composition of the collective world (Jaton, 2021: 12–17). Second, techno-moral graphs suggest that morality can also be thought of as a quantifiable continuum: when there are more accounted scruples about means, there is more morality (i.e., the surface of the map is larger). While the consultation of the accounts reporting on the collective explorations would be crucial to assess the seriousness of the enterprise (the only way to ensure that the graph is the expression of thoughtful moral concerns), techno-moral graphs would also give a rough indication about the *number* of genuine options explored. If this somewhat personal operationalization of James' pragmatist morality may seem fussy, even bureaucratic, it has the merit of giving the graphs a potential binding effect: An ML algorithm could be considered all the more moral as it has gone through many explored, and accounted, genuine options. The graphs may then become (costly) moral backings capable of serving, eventually, as a basis for subsequent evaluations. In that sense, if, in conjunction with the development of ML-related projects, techno-moral graphs were also required (and this would entail dedicated *moral secretaries*), the biases that are constitutive of ML algorithms would finally start to be assessed instead of being repressed.

Discussion and conclusion: ML governance equipped with assessed biases?

In a recent paper, Jobin et al. (2019) identified 84 public–private initiatives describing sets of principles to guide the moral/ethical development of ML, increasingly affiliated with AI. While this principled approach is certainly important in its ability to, sometimes, evoke political affects among non-expert publics, it is not sufficient to make effective differences. With regard to ML, moralism may establish a horizon to be reached but fails to enforce clear procedures. Worse, the active participation of AI companies in these high-level ethical and moral issues contributes to the ambient vagueness while also encouraging "policy-makers with a reason not to pursue new regulation" (Mittelstadt, 2019: 501). It is, in turn, difficult to be satisfied with the current situation: If one adheres to the ideals contained in the principled approach to ML and AI ethics, one has to admit that this moralism, on its own, does not have the means to achieve its ambitions. It still needs to be supported by new devices, procedures, and habits that have yet to be invented.

In this theoretical paper, I proposed a moral device based on the notion of bias and, more broadly, on what I called *ground-truthing practices*. I started by showing that biases should not be seen as a priori negative: According to Tom Mitchell's pioneering work on statistical learning, biases are necessary to define learning functions. Asking an ML algorithm to be bias-free would be tantamount to asking a tree to have no roots: a genuine contradictory injunction that has no chance of engaging binding mechanisms. I then considered ground-truth datasets to be repositories of biases that are central to the development of ML algorithms. I first proposed that, contrary to what is sometimes asserted in the specialized literature, these referential databases—and their underlying design practices that I call *ground-truthing practices*—contribute to the development of supervised *and* unsupervised ML algorithms. In order to come into existence and be used in real-world situations, ML algorithms depend indeed on biases that are the results of assignable practices. However, these practices can be more or less moral depending on their ability to make themselves sensitive to what William James (1912) called *genuine options*: moments of doubt about the consequence of an action that is necessary, irremediable, and imminent. The morality of ML processes may then be coextensive

with the exploration of their radical fragility: When there are more scrupulous doubts about the practices for defining the problem to be solved, about the data to be set up, and about the targets to be learned, there is more morality. I then suggested that these three practical axes that specify and delimit ground-truthing practices—problematization, databasing, and labeling—could be used as a scriptural space capable of hosting what I call *techno-moral graphs*: instruments for reporting and comparing the quantity of moral operations that took place during the grounding of ML-related projects. These graphs, like any other graph, would not be useful by themselves but by the many elements they compile and format. The techno-moral graph of an ML project would permit the visualization of the moral space of its ground-truthing part and—together with the qualitative consideration of the documents that each point incrementing the graph refer to—ensure that alternatives have indeed been the object of collective investigations. In that sense, it is a whole ecology of texts and documentation that would be summarized in techno-moral graphs that could not, obviously, exist on their own.

To imagine that each ML-related project proposes, in addition to its inner workings and performances, a techno-moral graph referring to the accessible accounts of the doubts and hesitations that occurred in its ground-truthing practices may seem utopian, at best. Devoting care and time to exposing the essential fragilities of ML algorithms and systems is not, by far, the priority of the actors involved in this industry whose products increasingly, and triumphantly, contribute to irrigating our daily lives. However, if biases are indeed constitutive of ML (which remains to be further demonstrated by in situ ethnographic studies), the current invisibility of biases must also be the result of a work of purification (Latour, 1987: 45–62) aimed at distinguishing them from the algorithms and systems they underlie. From there, integrating more morality into ML-shaping practices might be a matter of mere substitution: In place of the many efforts to make the biases underlying ML algorithms and systems invisible, one may think about valuing other efforts, such as those that would make biases visible and consultable.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ORCID iD

Florian Jaton 🄳 https://orcid.org/0000-0002-5001-9098

Notes

1. STS are a subfield of social sciences that aims to document the co-construction of science, technology, and the collective world. What connects the practitioners of this heterogeneous research community is the conviction that science is not only the expression of a logical empiricism, that knowledge of the world does not preexist, and that scientific and technological truths are dependent on collective arrangements, instrumentations, and dynamics (Dear and Jasanoff, 2010). I consider myself fully part of this research community.

2. More precisely, Mitchell describes bias as "any basis for choosing one generalization over another, other than strict consistency with the observed training instances" (Mitchell, 1980: 1).

3. At the 2019 International Solid-State Circuits Conference (ISSCC videos, 2019), LeCun refined his "cake analogy." The bulk of the cake is now self-supervision, a subcategory of unsupervised learning where the data provides its own supervision. This method makes convolutional neural networks independent from labels during their learning operations.

4. I selected these articles because they were all finalists for the "best paper award" of the 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). CVPR is one of the most prestigious and selective conferences in computer vision and image processing. For the 2019 edition, among the 5160 papers submitted, 1294 were accepted (25%) and 45 (0.87%) ended up being part of the final best paper award list.

5. More precisely, to test their unsupervised 3D hand pose estimator, Wan et al. (2019) relied upon the ground truth *NYU Test Set* initially proposed by Tompson et al. (2014). To evaluate their algorithm for part-based disentangling of object shape, Lorenz et al. (2019) used *CelebA* by Liu et al. (2015), *Cat Head* by Zhang et al. (2008), *CUB-200-2011* by Wah et al. (2011), *BBC Pose* by Charles et al. (2013), *Human3.6M* by Ionescu et al. (2014), *Penn Action* by Zhang et al. (2013), *Dogs Run* (self-made), and *Deep Fashion* by Liu et al. (2016). To evaluate their unsupervised domain-mapping algorithm, Fu et al. (2019) used *Cityscape* by Cordts et al. (2016), *MNIST* initially developed by LeCun, Cortes, and Burges and further publicized by Deng (2012), and

SVHN by Netzer et al. (2011). Finally, to evaluate and compare their self-supervised algorithm for optimal flow, Liu et al. (2019) used *MPI Sintel* developed by Butler et al. (2012), *KITTI 2012* by Geiger et al. (2012), and *KITTI 2015* by Menze and Geiger (2015).

6. In a recent paper, Henriksen and Bechmann (2020) proposed to call such grounding practices "truth practices." To me, this term is completely equivalent to "ground-truthing." However, there are two reasons why I prefer to use this somewhat complicated (and not very phonic) terminology: (1) It is a vernacular expression: "ground-truthing" is sometimes used in the field of applied computing to designate the action of defining external referents to support an algorithmic process. (2) Its tends to limit metaphysical speculation: the expression "ground-truthing" somewhat hides the philosophically very loaded term "truth."

7. In a recent TechCrunch article (Constine, 2019), Jordan Fisher—chief executive officer of Standard Cognition, a start-up that specializes in image recognition for autonomous checkout—talked about the enrolment of ML publications for industrial purposes: "It's the wild west—applying cutting-edge, state-of-the-art machine learning research that's hot off the press. We read papers then implement it weeks after it's published, putting the ideas out into the wild and making them production-worthy."

8. Moral experiences, according to Kant, only make sense in so far as they fall under the jurisdiction of a universal and necessary law. And the categorical imperative—formulated in Kant (1998 [1785])—is the finite version of the universal law that is applicable by us, finite humans, to our ordinary actions.

9. On the possibility to build habits around the exploration of genuine options, see the ethnographic work of Haeringer and Pecqueux (2020) on the space for dialogue "Parlons-en" in Grenoble, France.

10. With its millions of annotated images, the ground-truth dataset ImageNet is certainly the most illustrative example of the correlation between advances in computer science research and the construction and dissemination of ground-truth datasets. For a (quick) history of the formation of ImageNet, see Gray and Suri (2019: 6–8) as well as Gershgorn (2017) and Markoff (2012).

References

AI High-Level Expert Group – AIHLEG (2019) *Ethics guidelines for trustworthy AI*. Text, 8 April. Brussels: European Commission. Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed 14 May 2019).

Alpaydin E (2016) *Machine Learning: The New AI*. Cambridge: MIT Press.

Angwin J, Larson J, Mattu S, et al. (2016) Machine bias. *ProPublica*, 23 May. Available at: www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed 10 June 2020).

Barocas S, Hardt M and Narayanan A (2017) Fairness in machine learning. *Video tutorial presented at 2017 Conference on Neural Information Processing Systems*, Long Beach, CA, 4–9 December. Video available at: https://fairmlbook.org/tutorial1.html (accessed 28 May 2019).

Barocas S and Selbst AD (2016) Big data's disparate impact. *California Law Review* 104(3): 671–732.

Bechmann A and Bowker GC (2019) Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media. *Big Data & Society* 6(1): 205395171881956.

Bowker GC and Star SL (2000) *Sorting Things Out: Classification and Its Consequences*. Cambridge: MIT Press.

Butler DJ, Wulff J, Stanley Gb, et al. (2012) A naturalistic open source movie for optical flow evaluation. In: Fitzgibbon A, Lazebnik S, Perona P, et al. (eds) *Computer Vision—ECCV 2012*. Lecture Notes in Computer Science. Berlin: Springer, pp. 611–625.

Carpenter J (2015) Google's algorithm shows prestigious job ads to men, but not to women. *The Independent*, 7 July. Available at: www.independent.co.uk/life-style/gadgets-and-tech/news/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-10372166.html (accessed 10 June 2020).

Casilli A (2019) *En Attendant Les Robots*. Paris: Le Seuil.

Charles J, Pfister T, Magee D, et al. (2013) Domain adaptation for upper body pose tracking in signed TV broadcasts. In: *Proceedings of the British machine vision conference*, Bristol, UK, 9–13 September, pp. 1–11. Norwich, UK: BMVA Press.

Chokshi N (2019) How surveillance cameras could be weaponized with A.I. *The New York Times*, 13 June. Available at: www.nytimes.com/2019/06/13/us/aclu-surveillance-artificial-intelligence.html (accessed 12 July 2019).

Constine J (2019) To automate bigger stores than Amazon, standard cognition buys Explorer.Ai. *TechCrunch*, 7 January. Available at: http://social.techcrunch.com/2019/01/07/autonomous-checkout/ (accessed 12 July 2019).

Cordts M, Omran M, Ramos S, et al. (2016) The cityscapes dataset for semantic urban scene understanding. In: *2016 IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, 26 June–1 July, pp. 3212–3223. New York: IEEE Press.

Courtland R (2018) Bias detectives: The researchers striving to make algorithms fair. *Nature* 558(7710): 357–360.

Crawford K (2021) *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.

Custers B, Calders T, Schermer B, et al. (2012) *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*. Berlin: Springer.

Dear P and Jasanoff S (2010) Dismantling boundaries in science and technology studies. *Isis; An International Review Devoted to the History of Science and Its Cultural Influences* 101(4): 759–774.

Deng L (2012) The MNIST database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine* 29(6): 141–142.

Domingos P (2015) *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. New York: Basic Books.

Fu H, Gong M, Wang C, et al. (2019) Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In: *2019 IEEE/CVF conference on computer vision and pattern recognition*, Long Beach, CA, 16–20 June, pp. 2422–2431. New York: IEEE Press.

Geiger A, Lenz P and Urtasun R (2012) Are We ready for autonomous driving? The KITTI vision benchmark suite. In: *2012 IEEE conference on computer vision and pattern recognition*, Providence, RI, 16–21 June, pp. 3354–3361. New York: IEEE Press.

Geiger RS, Yu K, Yang Y, et al. (2020). Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*, Barcelona, Spain, 27–30 January, pp. 325–336. New York: ACM Press.

Gershgorn D (2017) The Data That Transformed AI Research—and Possibly the World. *Quartz*, July 26. Available at: https://qz.com/1034972/the-data-that-changed-the-direction   -of-ai-research-and-possibly-the-world/ (accessed 10 July 2018).

Gibney E (2020) The battle for ethical AI at the world's biggest machine-learning conference. *Nature* 577(7792): 609–609.

Gitelman L (2013) *'Raw Data' Is an Oxymoron*. Cambridge: MIT Press.

Gray ML and Suri S (2019) *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Houghton Mifflin Harcourt.

Grosman J and Reigeluth T (2019) Perspectives on algorithmic normativities: Engineers, objects, activities. *Big Data & Society* 6(2): 2053951719858742.

Hache É (2011) *Ce à Quoi Nous Tenons. Propositions Pour Une Écologie Pragmatique*. Paris: La Découverte.

Haeringer AS and Pecqueux A (2020) La vulnérabilité comme ouverture à la contingence. Deux enquêtes situées. *SociologieS* [Online]. Dossiers, Du pragmatisme au méliorisme radical, posted online on May 2, 2020. Available at: http://journals.openedition.org/sociologies/14011 (accessed 15 May 2019).

Hao K (2019) YouTube is experimenting with ways to make its algorithm even more addictive. *MIT Technology Review*, 27 September. Available at: www.technologyreview.com/2019/09/27/132829/youtube-algorithm-gets-more-addictive/ (accessed 10 June 2020).

Hazelwood K, Bird S, Brooks D, et al. (2018) Applied machine learning at Facebook: A datacenter infrastructure perspective. In: *2018 IEEE international symposium on high performance computer architecture*, Vienna, Austria, 24–28 February, pp. 620–629. New York: IEEE Press.

Hennion A and Monnin A (2020) Du pragmatisme au méliorisme radical : enquêter dans un monde ouvert, prendre acte de ses fragilités, considérer la possibilité des catastrophes. Introduction au Dossier. *SociologieS* [Online].

Dossiers, Du pragmatisme au méliorisme radical, posted online on May 2, 2020. Available at: http://journals.openedition.org/sociologies/13931 (accessed 15 May 2019).

Henriksen A and Bechmann A (2020) Building truths in AI: Making predictive algorithms doable in healthcare. *Information, Communication & Society* 23(6): 802–816.

Hutchins E (1995) *Cognition in the Wild*. Cambridge: MIT Press.

Ionescu C, Papava D, Olaru V, et al. (2014) Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(7): 1325–39.

ISSCC Videos (2019) Yann LeCun. In: *International Solid-State Circuits Conference 2019: Deep Learning Hardware: Past, Present, and Future*, San Francisco, 17–21 February. Video available at: https://www.youtube.com/watch?v=YzD7Z2yRL7Y (accessed 19 May 2020).

James W (1912) *The Will to Believe*. New York: Longmans, Green, and Co.

Jaton F (2017) We get the algorithms of our ground truths: Designing referential databases in digital image processing. *Social Studies of Science* 47(6): 811–40.

Jaton F (2019) "Pardonnez cette platitude" : de l'intérêt des ethnographies de laboratoire pour l'étude des processus algorithmiques. *Zilsel* 5: 315–339.

Jaton F (2021) *The Constitution of Algorithms: Ground-Truthing, Programming, Formulating*. Cambridge: MIT Press.

Jaton F and Vinck D (2016) Processus frictionnels de mises en bases de données. *Revue d'anthropologie des connaissances* 10(4): 489–504.

Jaton F and Vinck D (submitted) Politicizing algorithms by other means: Toward inquiries for affective dissensions. Manuscript submitted in January 2021 to *Perspectives on Science*.

Jobin A, Ienca M and Vayena E (2019) The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1(9): 389–99.

Kant I (1998 [1785]) *Groundwork of the Metaphysics of Morals*. Cambridge: Cambridge University Press.

Latour B (1987) *Science in Action - How to Follow Scientists & Engineers through Society*. Cambridge: Harvard University Press.

Latour B (2004) *Politics of Nature*. Cambridge: Harvard University Press.

Latour B (2012) *Enquête Sur Les Modes D'existence : Une Anthropologie Des Modernes*. Paris: La Découverte.

Latour B (2017) *Où Atterrir ?* Paris: La Découverte.

Latour B, Mauguin P and Teil G (1992) A note on socio-technical graphs. *Social Studies of Science* 22(1): 33–57.

LeCun Y (2016) *Predictive Learning, NIPS 2016 | Yann LeCun, Facebook Research*. Available at: www.youtube.com/watch?v=Ount2Y4qxQo&t=1072s (accessed 19 May 2020).

Lehr D and Ohm P (2017) Playing with the data: What legal scholars should learn about machine learning. *UCDL Rev* 51: 653–717.

Liu P, Lyu M, King I, et al. (2019) SelFlow: Self-supervised learning of optical flow. In: *2019 IEEE/CVF conference on computer vision and pattern recognition*, Long Beach, CA, 16–20 June, pp. 4566–4575. New York: IEEE Press.

Liu Z, Luo P, Qiu S, et al. (2016) DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In: *2016 IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, 26 June–1 July, pp. 1096–1104. New York: IEEE Press.

Liu Z, Luo P, Wang X, et al. (2015) Deep learning face attributes in the wild. In: *2015 IEEE international conference on computer vision*, Araucuno Park, Chili, 11–18 December, pp. 3730–3738. New York: IEEE Press.

Lorenz D, Bereska L, Milbich T, et al. (2019) Unsupervised part-based disentangling of object shape and appearance. In: *2019 IEEE/CVF conference on computer vision and pattern recognition*, Long Beach, CA, 16–20 June, pp. 10947–10956. New York: IEEE Press.

Manders-Huits N and Zimmer M (2009) Values and pragmatic action: the challenges of introducing ethical intelligence in technical design communities. *International Review of Information Ethics* 10(2): 37–45.

Markoff J (2012) For Web Images, Creating New Technology to Seek and Find. *New York Times*, November 19. Available at: https://www.nytimes.com/2012/11/20/science/for-web-images-creating-new-technology-to-seek-and-find.html (accessed 10 July 2018).

McDade M and Testman A (2019) Tackling bias in AI. Available at: www.ibm.com/blogs/systems/tackling-bias-in-ai/ (accessed 27 September 2019).

Menze M and Geiger A (2015) Object scene flow for autonomous vehicles. In: *2015 IEEE conference on computer vision and pattern recognition*, Boston, MA, 7–12 June, pp. 3061–3070. New York: IEEE Press.

Mirowski P (2011) *Science–Mart – Privatizing American Science*. Cambridge, MA: Harvard University Press.

Mitchell TM (1980) *The Need for Biases in Learning Generalizations. CBM-TR-5-Ho*. New Brunswick: Rutgers University.

Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1(11): 501–507.

Mittelstadt B, Allo P, Taddeo M, et al. (2016) The ethics of algorithms: Mapping the debate. *Big Data & Society* 3(2): 2053951716679679.

Netzer Y, Wang T, Coates A, et al. (2011) Reading digits in natural images with unsupervised feature learning. In: Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira F and Weinberger KQ (eds) *Advances in Neural Information Processing Systems 24*. Red Hook, NY: Curran Associates, pp. 567–575.

Neyland D (2019). *The Everyday Life of an Algorithm*. London: Palgrave Pivot.

Noble SU (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

Obermeyer Z, Powers B, Vogeli C, et al. (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366(6464): 447–453.

Pasquale F (2016) *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge: Harvard University Press.

Portugal I, Alencar P and Cowan D (2018) The use of machine learning algorithms in recommender systems: a systematic review. *Expert Systems with Applications* 97: 205–227.

Radfar C (2019) Bias in AI: A problem recognized but still unresolved. *TechCrunch*, 25 June. Available at: http://social.techcrunch.com/2019/07/25/bias-in-ai-a-problem-recognized-but-still-unresolved/ (accessed 26 September 2019).

Richardson M, Prakash A and Brill E (2006) Beyond Pagerank: Machine learning for static ranking. In: *Proceedings of the 15th international conference on World Wide Web*, Edinburgh, Scotland, May, pp. 707–715. New York: ACM Press.

Shellenbarger S (2019) A crucial step for averting AI disasters. *Wall Street Journal*, 13 February. Available at: www.wsj.com/articles/a-crucial-step-for-avoiding-ai-disasters-11550069865 (accessed 26 September 2019).

Silberg J and Manyika J (2019) *Notes From the AI Frontier: Tackling Bias in AI (and in Humans)*. Chicago: McKinsey Global Institute.

Tompson J, Stein M, Lecun Y, et al. (2014) Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics* 33(5): 169.

Verma S and Rubin J (2018) Fairness definitions explained. In: *Proceedings of the international workshop on software fairness*, Gothenburg, Sweden, 29 May, pp. 1–7. New York: ACM Press.

Wagner B (2018) Ethics as an escape from regulation: From 'ethics-washing' to ethics-shopping? In: Bayamlıoğlu E, Baraliuc I, Janssens L, et al. (eds) *Being Profiled*. Amsterdam: Amsterdam University Press, pp. 84–89.

Wah C, Branson S, Welinder P, et al. (2011) *The Caltech-Ucsd Birds-200-2011 Dataset. Report*. Pasadena: California Institute of Technology.

Wan C, Probst T, Van Gool L, et al. (2019) Self-supervised 3D hand pose estimation through training by fitting. In: *2019 IEEE/CVF conference on computer vision and pattern recognition*, Long Beach, CA, 16-20 June, pp. 10845–10954. New York: IEEE Press.

Yang S, Luo P, Loy CC, et al. (2016) WIDER FACE: A face detection benchmark. In: *2016 IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, 26 June–1 July, pp. 5525–5533. New York: IEEE Press.

Zhang W, Sun J and Tang X (2008) Cat head detection – How to effectively exploit shape and texture features. In: Forsyth D, Torr P and Zisserman A (eds) *Computer Vision—ECCV 2008. Lecture Notes in Computer Science*. Berlin: Springer, pp. 802–816.

Zhang W, Zhu M and Derpanis KG (2013) From actemes to action: A strongly-supervised representation for detailed action understanding. In: *2013 IEEE international conference on computer vision*, Sydney, Australia, 8–12 April, pp. 2248–2255. New York: IEEE Press.