The Predictive Reframing of Machine Learning Applications: Good Predictions and Bad Measurements

Abstract

Supervised machine learning has found its way into ever more areas of scientific inquiry, where the outcomes of supervised machine learning applications are almost universally classified as predictions. I argue that what researchers often present as a mere terminological particularity of the field involves the consequential transformation of tasks as diverse as classification, measurement, or image segmentation into prediction problems. Focusing on the case of machine-learning enabled poverty prediction, I explore how reframing a measurement problem as a prediction task alters the primary epistemic aim of the application. Instead of measuring a property, machine learning developers conceive of their models as predicting a given measurement of this property. I argue that this *predictive reframing* common to supervised machine learning applications is epistemically and ethically problematic, as it allows developers to externalize concerns critical to the epistemic validity and ethical implications of their model's inferences. I further hold that the predictive reframing is not a necessary feature of supervised machine learning by offering an alternative conception of machine learning models as measurement models. An interpretation of supervised machine learning applications to measurement tasks as *automatically-calibrated model-based measurements* internalizes questions of construct validity and ethical desirability critical to the measurement problem these applications are intended to and presented as solving. Thereby, this paper introduces an initial framework for exploring technical, historical, and philosophical research at the intersection of measurement and machine learning.

Keywords

Machine Learning • Measurement • Prediction • Modeling • Conceptual Engineering

Acknowledgments

I would like to thank Shannon Vallor, Sabina Leonelli, Atoosa Kasirzadeh, and Arno Onken for their helpful comments and discussions, as well as Kevin Hoover and Marcel Boumans for feedback during the beginning stages of my thinking on the matter. Earlier versions of this paper have been presented at conferences of the European Philosophy of Science Association and the Society for Philosophy and Technology, as well as workshops at the University of Hannover and Edinburgh. I thank the participants at these events and one anonymous reviewer for their constructive criticism.

Funding

Work on this paper was supported by the Baillie Gifford PhD scholarship in AI ethics at the Centre for Technomoral Futures.

Competing Interests

The author declares no competing interests.

Alexander Martin Mussgnug
University of Edinburgh
0000-0002-5951-057X
a.mussgnug@ed.ac.uk

I Introduction

In the past decade, machine learning (ML) has evolved from a predominantly exploratory field to an increasingly established and more broadly used instrument of inquiry. Advancements in computational hardware, unprecedented volumes of data, and new statistical methods have enabled the deployment of ML in ever more areas. Today, ML models are applied to tasks as diverse as the automatic segmentation of plant images (Smith et al., 2020), unemployment rate forecasting (Chakraborty et al., 2021), or better modeling retinal sensory processing (Tanaka et al., 2019).

In light of the proliferation of ML applications within science, critical questions emerge regarding the epistemology of ML. Whereas issues such as the explanatory potential of ML models (e.g., Chirimuuta, 2021; López-Rubio & Ratti, 2019; Sullivan, 2019) or their ability to provide causal knowledge (e.g., Canali, 2016; Pietsch, 2016, 2021) have received initial philosophical consideration, relatively little has been said about the fact that ML applications are almost universally categorized as predictive. The overwhelming tendency to approach ML applications as predictive finds expression, for instance, in Agrawal et al.'s seminal book "Prediction Machines" on the economics of ML: "Because it [ML] is becoming cheaper it is being used for problems that were not traditionally prediction problems. Kathryn Howe, of Integrate.ai, calls the ability to see a problem and reframe it as a prediction problem 'AI Insight,' and, today, engineers all over the world are acquiring it." (2018, p. 23). What Agrawal et al. commend as "AI Insight," I take as motivation to critically analyze how exactly ML developers reframe problems as prediction tasks.

My argument will proceed in three parts. In section two of this paper, I will investigate the *predictive reframing* of ML applications. With predictive reframing, I seek to designate the transformation of tasks as diverse as measurement problems, image segmentation, or explanatory modeling into statistical predictions. Throughout the paper, I illustrate my argument by means of ML-enabled poverty prediction as an example of how developers reinterpret a socioeconomic measurement problem as a predictive task. I aim to shed light on the often-unacknowledged fact that the predictive reframing involved in applying supervised ML to measurement tasks alters their primary epistemic aim. In measurement, we seek to infer the value of a property of interest. However, when reframing measurement tasks as predictions, ML developers interpret their applications not as inferences of a property but, instead, as predicting a particular poverty metric.

In section three, I argue that this predictive reframing in ML is neither epistemically nor ethically neutral. Reframing and evaluating ML applications as predictions allows developers to leave questions central to the original measurement problem outside of the immediate scope of the predictive task. This includes, but is not limited to, the question of which measurement is the *right* measurement for the given purpose. Instead of engaging critically with the ethical desirability or epistemic validity of the model's inferences with respect to the property of interest, the evaluation of supervised ML models is done based on statistical correlation with a given measurement. However, just because a supervised ML model might display a high correlation with the known measurement values, i.e., reliably solve the predictive task, it does not necessarily follow that the application also adequately addresses the initial measurement problem the model is presented as solving. In other words, *a supervised ML model might provide good predictions but bad measurements*.

ML developers' tendency to nonetheless present their predictions as solutions to the original measurement problem, in conjunction with the fact that the predictive reframing obscures how discrepancies between the model's prediction and the reference measurement might correlate with real-world poverty, motivates me to propose interpreting some ML applications not as predictions but as measurements in the last section of this paper. More specifically, I argue in section four that one can interpret supervised ML applications to measurement tasks as *automatically calibrated model-based measurements* and thus, as being calibrated with the help of a reference procedure but ultimately *measuring* the property of interest. Such an understanding of some ML models as measurements internalizes questions of construct validity and ethical desirability critical to the measurement problem these applications are intended to and presented as solving. In other words, thinking of certain supervised

ML applications as model-based measurements rather than predictions brings them out of the frame of mere statistical calculus and into more critical engagement with the underlying theoretical and conceptual assumptions.

My argument also relates closely to scholarship on AI ethics, which often addresses precisely the questions that developers, through the predictive reframing of their application, externalize from the development and evaluation of ML models. By pointing towards the philosophy of measurement as an alternative perspective for understanding normative dimensions of ML applications, I illustrate one way in which the philosophy of science might help shed light on issues in AI ethics. My contribution to these questions is primarily introductory, highlighting the need for technical, historical, and philosophical research at the intersection of measurement and ML.

II Predictive Reframing

II.1 A Measurement Problem

As the United Nations General Assembly notes, "eradicating poverty in all its forms and dimensions, including extreme poverty, is the greatest global challenge and an indispensable requirement for sustainable development" (United Nations, 2015, p. 5). Effective policy intervention and research require reliable and timely information about the state of poverty at an appropriate spatial granularity. Through traditional survey-based measurements, such data is generally available for most industrialized countries. However, reliable measurements are scarce in developing regions where poverty relief is most needed. Even when available, wealth and consumption surveys can lie more than ten years apart and are often not provided at the spatial resolution required (Yeh et al., 2020).

In the absence of timely survey-based poverty metrics, economists must find alternative ways to provide information on the poverty of a region that can guide policy intervention and educate scholarship. What is needed are practical solutions to a *rather typical socioeconomic measurement problem,* namely that of providing timely and reliable measurements of poverty.

II.2 Predicting Poverty

To address this need for current poverty measurements, particularly in Africa, researchers have recently suggested relying on alternative forms of data. With the help of ML, an emerging field of research aims to infer poverty metrics in developing regions from data as diverse as mobile phone records and satellite imagery (e.g., Blumenstock et al., 2015; Jean et al., 2016; Pokhriyal & Jacques, 2017; Yeh et al., 2020). In addition to significantly lower costs, these methods can provide information on the poverty of a region at a much higher spatial granularity and frequency than survey-based measurements.

The developers, however, do not understand their applications as measurement. Instead, they speak of poverty prediction, as illustrated by titles such as "Combining satellite imagery and machine learning to predict poverty" (Jean et al., 2016) or "Predicting poverty and wealth from mobile phone metadata" (Blumenstock et al., 2015). Such an interpretation of the applications' outcomes as predictions might, upon closer inspection, appear somewhat curious. When "predicting," at least in its original sense, one seeks to anticipate or forecast ("Prediction," 2021). In other words, with the term prediction, we often associate inferences about the future.[1] However, ML applications to poverty prediction do not aim to predict the future but to infer the current state of poverty in a particular region by providing an alternative to often outdated and coarse survey measurements. *How do ML researchers then understand their models as "predicting poverty?"*

To comprehend how developers understand the outcomes of ML applications as predictions, one must look at their development. It is critical to distinguish two components of any machine learning

[1] Within scientific discourse, the term "prediction" is not limited to inferences about the future. For instance, in statistics, model outcomes are labeled predictions regardless of their temporality. As will become clear in the following paragraphs, my characterization of machine learning predictions, equally, is not dependent on their temporal relation to the event or property of interest.

application: the machine learning model and the learning algorithm. In machine learning, the learning algorithm is responsible for adjusting the parameters of an initial generic model based on data and, therethrough, "builds" the machine learning model from data (Zhou, 2021, p. 2). A useful heuristic is picturing the relationship between the machine learning model and algorithm analogously to that of a program and the programmer. Broadly speaking, one can differentiate between three kinds of machine learning algorithms. In unsupervised learning, models are trained upon a dataset of which they independently discover structural properties. Given instances of a random vector of data points x, unsupervised machine learning algorithms infer (aspects of) its probability distribution $p(x)$. Contrarily, supervised learning algorithms require a dataset containing an explicit label or target y. In supervised learning, one aims to infer unknown cases of y given x by approximating the probability distribution $p(y \mid x)$. A third category of machine learning algorithms, grouped under reinforcement learning, learns through external feedback (Goodfellow et al., 2016, pp. 104–106).

To "predict poverty," ML developers generally employ a supervised learning approach.[2] Therefore, developers collect data on the predictor variables (such as features extracted from satellite imagery or call data records) and known values of the target variable, usually in the form of survey-based poverty metrics. Upon data collection and preprocessing, the development of supervised ML models generally consists of two steps: model training and model testing.[3] During training, the model's parameters are adjusted based on differences between the initial model's inference given the feature data and the known values of the target variables. Pokhriyal & Jacques (2017), for instance, train Gaussian Process models (a specific ML model) on input data consisting of environmental data (on economic activity, access to facilities, and metrics of food security) and call data records, as well as known values of the Multidimensional Poverty Index (MPI). Blumenstock et al. (2015) train elastic net and tree-based ensemble regression models on features extracted from mobile phone network records and known values of a composite wealth index. In a second step, the model is then tested based on how well it infers known values of the target variable from data it was not trained upon.

During development, given input data but not yet knowing the corresponding value of the poverty metric in the dataset, the model *predicts,* i.e., anticipates this value. Subsequently, the value of the target variable in the dataset is revealed to the model, and adjustments are made (training) or performance metrics are calculated (testing). This explains how, during development, one might understand the inferences of supervised ML models as predictions: the model predicts a value in a dataset that is later disclosed to it.

The application of these models, however, presents itself differently. Researchers propose to employ the trained models (or future iterations of them) in regions or at a level of spatial granularity where no up-to-date survey-based measurements are available. Nonetheless, the conception of the proposed supervised ML *applications* as "predicting poverty" is an extension of the nature of the inferences during the development of the model. When researchers propose that their models might "predict poverty" where no survey-based measurements are available, they more precisely suggest that their model anticipates the missing value of the poverty metric in the new dataset. Consequently, *developers understand the ML model not as "predicting poverty," but, instead, as predicting the hypothetical value of a particular poverty metric.* This distinction is rather subtle and is, if at all, only mentioned cursorily in the literature. For instance, Pokhriyal & Jacques (2017, p. 9784) briefly remark that "throughout the[ir] paper, 'poverty' refers to the Global MPI." "Predicting poverty," in this case, more precisely means predicting the Multidimensional Poverty Index of a region.

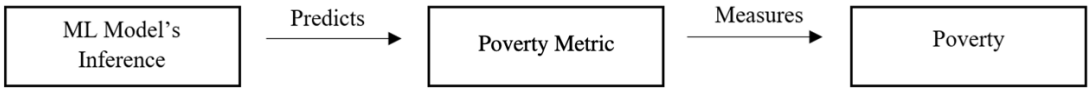
The understanding of ML applications as predictive is often presented as a mere terminological particularity of the field (e.g., Agrawal et al., 2020, Bell, 2014, p. 2, Hastie 2009, p. xi). Instead, I have argued that when presented with a measurement problem, but approaching it as a prediction task, ML developers commit to more than a mere act of lexical convention (measuring poverty vs. predicting

[2] Recently, a small number of papers have leveraged a semi-supervised learning approach (e.g., Perez et al., 2019; Zhao et al., 2020). I will leave an explicit treatment of semi-supervised learning for future research.
[3] In practice, the development is often more complex and involves iterative training, testing, and validation steps.

poverty). Instead, they actively reframe the primary epistemic aim from measuring the poverty of a region to predicting the value of a *given* measurement of poverty. In other words, through the predictive reframing, developers conceive of their supervised ML outcomes as second-order rather than first-order inferences (see Figure 1).[4]

Figure 1

The Predicting Reframing of the Original Measurement Task:



III Good Predictions and Bad Measurements

III.1 The Evaluation of Machine Learning Predictions

The predictive reframing of a measurement problem into a prediction task is neither epistemically nor ethically neutral. Considerations critical to the original problem are externalized when machine learning developers reframe the original measurement task into predictions of a particular poverty metric. This shows most clearly in how supervised learning applications are evaluated.

The initial problem is to provide information on the poverty of a region that can reliably guide policy and research. Consequently, poverty measurements are evaluated based on how accurately, coherently, or responsibly they capture the target property relative to a given application. Central to evaluating any potential solution to the initial problem is the question of whether any particular measurement is the *right* measurement for the context it is applied within. Therefore, measurement experts must assess whether any particular measurement can adequately inform, for instance, necessary policy decisions and highlight potential limitations.

The question of which measurement of poverty to use involves both epistemic and ethical dimensions. It implies ethical concerns regarding the fairness or social desirability of using a given metric (or any form of quantification), issues that should ideally be analyzed within a broader social and historical context. Furthermore, it raises closely related epistemic issues such as assuring the construct validity of a measurement, i.e., whether a given measurement conforms with theoretical hypotheses surrounding poverty and its relation to other properties (see, for instance, Mari et al., 2021, pp. 88–89). Far from mere statistical calculus, answering these questions requires a great degree of domain knowledge and critical judgment on often competing hypotheses (Boumans, 2015, Chapter 5) as the validity of particular poverty metrics is subject to disagreement even within the discipline.[5] As much as such evaluation is critical, it is also intricate and ideally involves addressing challenging contextual questions that lack clear-cut numerical answers (Alexandrova & Haybron, 2016).

However, the work on ML-enabled poverty prediction circumvents these critical questions regarding the validity, suitability, or social desirability of a particular measurement. Within the reframed predictive task, the model's inferences are no longer understood as measuring (dimensions of) poverty but rather, more narrowly, as predicting a given poverty metric. Consequently, when evaluating the model's predictions, *a particular way of measuring poverty is taken for granted*, leaving questions of how the model's inference relates to the property of interest outside the primary scope of the prediction task.

A closer look at the literature helps exemplify my point. For instance, in one of the landmark articles on ML-enabled poverty prediction, Blumenstock et al. (2015) predict a composite wealth index based on features extracted from call data records provided by Rwanda's near-monopoly mobile network

[4] I thank Michał Wieczorek for suggesting this formulation.

[5] See, for instance, the debate between Nájera Catalán & Gordon (2020) and Santos & Villatoro (2020) or research on the discrepancies between income and multidimensional poverty (e.g., Wang et al., 2016).

provider. Only when reading the supplementary materials, does one find out that the composite wealth index is the first principal component of eight questions asked in a phone survey (Blumenstock et. al, 2015, Suppl. 1D). These questions are chosen based on their correlation with the first principal component of a broader set of asset and household-related questions from a representative national Demographic and Health Survey (DHS) and practical constraints during the researchers' phone survey.[6] Ultimately, what the ML model predicts is an aggregate measure of phone survey answers to whether the respondent owns a refrigerator, radio, television, bicycle, or motor scooter, as well as the size of the household, access to electricity, and the number of children. Critically, the authors provide no validation of this composite wealth index and do not address how their measure relates to the concept of poverty or its adequacy for policy decisions. In other words, the developers leave wide open to what extent the trained model really predicts *poverty.*

Blumenstock et al.'s (2015) article is no exception. Generally, developers offer little justification for the choice of a particular poverty metric, nor do they mention critical consideration of alternative poverty metrics (e.g., Jean et al., 2016; Pokhriyal & Jacques, 2017). Instead, the evaluation of ML *predictions* is often limited to the computation of correlation metrics with known values of a given measurement. However, just because a supervised ML model might display a high correlation with recorded values of a particular poverty metric, i.e., reliably solve the predictive task, it does not necessarily follow that the application also adequately addresses the initial measurement problem. Given high correlation, it can be reasonably assumed that the ML model's outcomes somewhat accurately reproduce the reference measurement. However, such evaluation leaves unaddressed whether a given reference poverty metric is the right measurement by critically addressing its suitability for the particular purpose and theoretical validity.[7] Consequently, it does not necessarily follow that the ML models also produce adequate information regarding the poverty of a region. Even when the model might achieve high predictive accuracy, it might not provide a reliable solution to the original aim of the researchers. In other words, *a supervised ML model might provide good predictions but bad measurements*.

III.2 Motivating the Need to Move Beyond the Predictive Reframing

At this point, I might not have fully convinced the reader that machine learning developers' focus on predicting the ground truth data provided by a reference measurement is necessarily problematic. After all, policymakers and researchers could, in a separate step, assess the validity or adequacy of the respective poverty metric predicted to ensure the machine learning model can provide reliable solutions to the initial problem.

I respond to such a position in two ways. First, I highlight that, in practice, machine learning developers often fail to clearly communicate the limitations of their approach, effectively marketing their applications as solutions to the original measurement problem. Second, I argue that even if machine learning developers were to more explicitly acknowledge the need to further validate the reference measurement, fundamental shortcomings remain. In principle, the non-transitivity of the relationship between prediction, reference measurement, and construct, as well as considerations regarding the broader purpose of scientific research, motivate moving beyond the predictive reframing in supervised machine learning.

III.2.1 Communicating limitations

The predictive reframing and the resulting avoidance of questions critical to the initial measurement problem are hardly communicated. Most often, developers do not explicitly mention the reframing into a prediction task at all. Even if the researchers cursorily note it, they do so in a manner that obscures rather than illuminates the distinction: for instance, when developers briefly remark that "throughout the paper, 'poverty' refers to the Global MPI" and subsequently equate a measurement of a

[6] According to the authors, respondents were unable to quickly answer some questions from the DHS (Blumenstock et. al, 2015, Suppl. 1D).

[7] I am not committed to any particular notion of measurement validity or adequacy as a detailed treatment of the philosophical literature on measurement validation and adequacy (e.g., Alexandrova, 2017, Chapter 6; Alexandrova & Haybron, 2016; Bokulich & Parker, 2021; Feest, 2020) lies outside the scope of this paper.

property (Global MPI) with that property (poverty) (Pokhriyal & Jacques, 2017, p. 9784). In other cases, central aspects of their work, such as the composition of the particular poverty metric the model predicts, are only provided in the supplementary materials (e.g., Blumenstock et al., 2015, Suppl. 1D).

In this manner, *supervised ML applications are often presented (and possibly intended) as solving one task but developed and evaluated as solving a different problem*. ML-enabled poverty estimation is presented as addressing the demand for reliable measurements of poverty in developing regions but approached and evaluated as predictions of a given poverty measurement. In other words, the predictive reframing and its lack of communication enable developers to focus on a much more limited and straightforward problem while marketing their applications as solutions to a more substantial and complex one.

Even when developers might be aware of the limitations of their approach, these consequences might not be as evident to journalists or policymakers interpreting and possibly adopting their research. This is underscored by the often-made claim in popular discourse and media that ML algorithms offer some immediate (or even naïvely objective) insight into a particular phenomenon. The predictive reframing, however, understands the model as merely predicting the missing value in a dataset. This interpretation does not necessarily relate the model's inferences to some feature of reality but, instead, often remains firmly within the, one step removed, realm of data.

A first step would be for developers to explicitly acknowledge the implications of the predictive reframing and, thereby, appropriately relativize the ML model's predictions. However, expecting ML developers to highlight a fundamental limitation of their application runs counter to the natural tendency to present their work as favorably as possible. Given the prevalence of predictive reframing within ML research, any deviation from it might further conflict with financial incentives such as attracting funding. As a discipline heavily driven by industry (Hagendorff & Meding, 2021), the marketing of results arguably plays an even more significant role within ML.

III.2.1 Residuals – Error or Improvement?

Moreover, I argue that even if machine learning developers were to communicate the limitations of their approach, problems with the predictive reframing persist. Acknowledging that the epistemic target and standard of evaluation is a particular poverty metric, highlights the need to further evaluate the adequacy or validity of this reference measurement. Performance evaluation would then become a two-step process ensuring that the model's inferences reliably reproduce the reference measurement and, in a separate step, that the reference measurement is the appropriate metric given the application. In the case of ML-enabled poverty inference, developers would, for instance, verify whether the model accurately reproduces the metric, whereas policymakers and researchers (possibly together with developers) critically question whether the poverty metric predicted serves as the right measurement of poverty for the application.

I argue that such evaluation would still face a fundamental shortcoming: the relation between the ML model's prediction, the reference measurement, and the latent property of interest is not necessarily transitive. Even when the model's inferences are good enough predictions of the poverty metric and the poverty metric is a good enough measurement of poverty, the model's inferences may not ultimately provide reliable information on the poverty of a region. This is because the question of how discrepancies between the reference measurement and the ML model's inference (residuals) relate to the property of interest lies outside the scope of such a two-step evaluation. Given that the supervised ML models proposed for poverty prediction do not perfectly reproduce the poverty measurements they predict, even numerically small residuals might correlate with real-world poverty in ways that undermine the prediction's adequacy and validity.

Curiously, it is also conceivable that what might constitute an error in the prediction problem might be an improvement in measuring poverty. Imagine a case where, because some conceptual commitments of the operationalization of the reference measurement translate less to a particular region, a survey-based poverty metric might not measure the poverty of this region adequately. In this case, the supervised ML model not relying on those survey responses but on satellite imagery and call data to predict the poverty metric might more accurately capture the poverty of this given region. While such a

discrepancy between the model's prediction and the measurements would constitute an error when we conceive of its primary epistemic aim as predicting the poverty metric, it might provide an improvement with respect to our ultimate goal of measuring poverty.

The point, however, is broader than merely the evaluation of individual applications — an understanding of supervised ML applications as predictive limits the advancement of the discipline as a whole. As long as applications are conceived of and developed to predict or reproduce the results of a reference procedure, they are limited to the accuracy and validity of this procedure. The primary value of supervised ML is thus restricted to gains in efficiency or applicability (see also Ratti, 2020). This is because as long as the primary epistemic aim of these applications is interpreted as predicting the results of a given reference procedure, improvements in accuracy relative to the ultimate property of interest are errors with respect to the primary epistemic aim.[8]

IV How machine learning measures

Here I propose an alternative conceptualization of supervised ML that avoids the predictive reframing and understands applications of supervised ML to measurement problems as measurements. Thereby, I hope to not merely avoid some of the problematic aspects of the predictive reframing but also situate supervised ML applications within the appropriate frame of reference for the task they are presented as solving. When ML applications are presented as solutions to the problem of measuring poverty, they should also be developed and evaluated in the context of measurement rather than prediction. Again, I will illustrate this point by showing how machine-learning enabled poverty inferences presented as solutions to the original problem of measuring poverty can be interpreted as measuring poverty rather than reframing them as predicting a given poverty metric.

I argue that when supervised ML is applied to measurement tasks, one can conceive of these applications as *automatically calibrated model-based measurements.* By understanding supervised ML models as measurement models, one can account for the fact that models are calibrated towards a reference procedure and, nonetheless, understand the ultimate property of interest and not the hypothetical results of a reference procedure as their primary epistemic target. This is achieved by separating the goal during the development of ML models from the epistemic target of their application. The former I propose to understand as a predictive activity, namely (steps of) the calibration of a measurement model, whereas the latter I suggest understanding as the *measurement* of a specific property.

Let me briefly illustrate my point. One might calibrate a watch towards the measurements of reference instruments, such as the network of atomic clocks or calibrate a psychological test towards a set of reference scores. However, once in use, one would not commonly conceive of these instruments as *predicting* the reference (such as predicting the hypothetical measurement of the atomic clock or predicting the performance relative to the set) but as *measuring* the property of interest. One employs the calibrated clock to measure time and the psychological test to measure cognitive ability.

I propose understanding supervised ML applications to measurement problems in the same sense. Referring to the example of poverty metrics discussed above, supervised ML models are calibrated towards some reference poverty metric, but when intended to be deployed, they should be understood and evaluated as *measuring poverty.* This conceptualization of supervised ML applications I take to be both descriptively accurate as it captures better the original epistemic task, as well as instrumentally useful as it thereby avoids some of the problematic aspects of the predictive reframing I have touched upon.

Below, I lay out in some detail a first exploration of how some supervised ML applications might be interpreted as *automatically calibrated model-based measurements*. To this end, I provide a brief

[8] One might hold that it is in the nature of supervised ML to be merely reproductive. However, the extensive literature on fairness in machine learning illustrates how conceptual and ethical considerations regarding the target construct (e.g., that an applicant's fit for a job is not influenced by gender or race) can be implemented to avoid reproducing bias (Mehrabi et al., 2022, Section 5).

account of both measurement calibration and model-based measurement to then illustrate how we might, analogously, understand supervised ML as the automatic calibration of a model-based measurement.

IV.1 Measurement Calibration

Recent accounts of measurement in both the social (Boumans, 2007, p. 200) and the natural sciences (e.g., Morrison, 2009, pp. 49–55; Tal, 2017a, 2017b) equally emphasize the role of models in measurement practice. According to such a reading, measurement involves an empirical process, as well as theoretical and statistical modeling. In the social sciences, the empirical process can often be limited to passive observation. Consequently, researchers have little control over factors other than the property of interest that might influence their observations. Modeling provides a means to account for these confounding factors to reliably infer the (often latent) property of interest from observations. Within the natural sciences, and particularly in laboratory settings, metrologists generally have more control over the measurement operation. Nonetheless, when aiming at a particularly accurate measurement outcome, modeling remains critical even in the natural sciences to abstract from the measurement any remaining influencing factors and bring the outcomes of the specific operation into coherence with other measurements and background theory.

Measurement experts do so in a process called measurement calibration, which finds one of its most definite philosophical expressions in Eran Tal's work on precision metrology (e.g., 2017a, 2017b). According to Tal, the first step towards understanding measurement calibration lies in distinguishing between instrument indications and measurement outcomes. An instrument indication "is a property of a measuring instrument in the final state after the measurement is completed" (Tal, 2017b, p. 34). Examples of instrument indications are the level of a liquid in a thermometer or the change of a pH test strip's color. These instrument indications must be understood as parameters of the particular measuring operation, including the specific instrument, environment, and operators. In contrast, measurement outcomes are knowledge claims about the state of the object under measurement, such as its temperature or acidity. To arrive at a measurement outcome, one must abstract from the measurement the idiosyncrasies of the specific operation to bring it into coherence with a broader network of measurement instruments. This inference from instrument indications to measurement outcomes is *calibration,* which Tal proposes to understand as a two-part modeling practice.

The first step, Tal labels forward calibration. Forward calibration iteratively determines the relationship between quantity values provided by a reference procedure and the indications of the instrument being calibrated. In a second step, termed backward calibration, one then aims to infer the measurement outcome from an instrument indication. Tal proposes to understand both inferences as model-based. Background knowledge, as well as statistical and theoretical assumptions, relate the quantity values provided by a reference to the instrument indications and later to the measurement outcome through iteratively adjusted functional relationships.

To illustrate this point, Tal presents the example of calibrating a caliper by placing gauge blocks between its jaws. The gauge blocks provide references in the form of quantity values that can be mapped to the instrument's indications. However, various other factors affect the instrument's final state, such as the temperature of the environment. To arrive at a precise forward calibration, one must *model the measuring process* as a function of the quantity value provided by the reference object and other parameters influencing the reading. This calibration "*involves iterative modifications to the model of the apparatus* as well as to the apparatus itself" (p. 38, emphasis added). Once a satisfactory level of accuracy is achieved, the calibration function provides us with a model-based estimate of the instrument indication given a quantity value and additional parameters involved in the measurement operation.

The instrument indication, in contrast to a measurement outcome, cannot immediately be attributed to the object of measurement. To infer the measurement outcome, one again relies on a model of the measurement operation: the backward calibration function. The measurement outcome is the best estimate of the object's diameter given the observed instrument indication, the backward calibration function, and the parameters of other potential influences on the measurement. If the circumstances between forward calibration and measurement are sufficiently similar, and no additional factors need to be considered, the backward calibration function is but the inverse of the forward calibration function. This modeling and its underlying statistical and theoretical assumptions then warrant the claim that the ultimate measurement outcome pertains to the object rather than the specific measurement operation.

In summary, Tal's model-based account of measurement calibration interprets measurement outcomes not as the immediate results of a physical operation but as pertaining to an iteratively calibrated

model of that system. They are estimates based on a two-way process of inference. First, calibration iteratively determines a functional relationship between quantity values provided by a reference and the measurement instrument's indications. Second, measurement outcomes are deduced as best predictors of observed instrument indications given the theoretical-statistical model of the system.

IV.2 Model-Based Measurement

Tal's account highlights the significance of modeling in the calibration of measurements in precision metrology. To arrive at the most accurate measurement outcome, the data obtained from the instrumented measurement operation must be modeled to account for confounding factors and the idiosyncrasies of the specific operation. In other cases of measurement, however, one lacks a measurement instrument in the first place. Measurements in the social sciences often rely on data coming from official statistics or field observations, not from controlled instrumented measurement operations.[9] Rather than calibrating a measurement instrument, in the social sciences, the measurement model itself relates the observations to the property of interest. In other words, the measurement model also functions as the measurement instrument. For this reason, Marcel Boumans (2007) terms these inferences model-based measurements.

Similarly, I propose to understand supervised ML applications as measurements or, more specifically, as *automatically calibrated model-based measurements*. Supervised ML models are models intended to measure a given property that are automatically (the "learning" in machine learning) calibrated towards a reference measurement. Nonetheless, one can ultimately interpret their application as *measuring* a property of interest rather than *predicting* a hypothetical measurement of it.

IV.3 Supervised Machine Learning as Automatically Calibrated Model-Based Measurement

IV.3.1 Model Choice in Machine Learning and Measurement

Similar to measurement calibration, one of the first steps in any supervised ML application is deciding on an initial model architecture. In the case of ML-enabled poverty inference, the models employed range from Gaussian Process Regression (GPR) models (Pokhriyal & Jacques, 2017) and Convolutional Neural Networks (CNN) (Xie et al., 2016) to regression models (Blumenstock et al., 2015). The model choice depends on theoretical and statistical assumptions such as whether the developers face a classification or regression problem, the presumed functional relationship between the target variable and the feature vector, computational requirements, and the nature of the input data. CNNs, for instance, are primarily applied when models are trained upon visual information. Xie et al. (2016) rely on a CNN within a more complex transfer learning architecture to extract spatial information from satellite image data. In contrast, Pokhriyal & Jacques (2017) employ a GPR model to infer the poverty of a region based on environmental data and call records as these models do not merely output a singular value but also an associated uncertainty.

However, the model-choice in supervised ML is more complex than that, at times relying heavily on intuition or trial and error. Given a model, the developers specify the learning algorithm, kernel function, or hyperparameters of the model. For CNNs, the developers decide, for instance, on the number of layers, the learning algorithm, or the size of the convolutional filters. For GPR models, the researchers must specify the nature of the kernel function that can be interpreted as a similarity measure of individual data points. Pokhriyal & Jacques (2017), for example, employ a kernel function that separately accounts for nonlinear dependencies in the feature and geographic space of the input data.

This initial model choice and specification involved in supervised ML mirrors that in model-based measurement. Similar to ML developers, measurement experts must first specify an initial measurement model in the form of one or multiple functions. Thereby, theoretical or statistical background knowledge and the degree of precision needed, guide the initial specification of the model. Granted, the functions underlying a measurement might appear more formally representative of the situation being modeled than is the case in ML. Before training, ML models are often vastly more generic than many of the models employed in traditional measurements.

[9] Similarly, information as diverse as expert judgment or theoretical commitments is used as a reference when calibrating measurement models (Boumans, 2015, Chapter 5).

IV.3.2 Training & Testing as Calibration

However, in both cases, the initial model is only a tentative starting point. To move from an initially somewhat generic model to one that can reliably infer the poverty of regions, the model is iteratively adjusted, based on both input data and known target values. This "training" in supervised ML, I argue, can be understood as the calibration of a measurement model.

First, one might look at the data upon which the supervised ML models are calibrated. In the case of supervised ML-enabled poverty estimation, the most common data sources used are satellite data, environmental data, and call data records. The empirical processes producing the data (satellite photography, monitoring of environmental conditions, etc.) are generally forms of passive observation not specifically designed for the application. This is similar to measurements in the field and social sciences, which often rely on secondary data such as government statistics or market data. Identical to traditional measurement calibration, the supervised ML applications do not merely require the input data but also the known reference values the model is calibrated towards. In the case of supervised ML-enabled poverty prediction, the models are calibrated towards known survey-based poverty metrics.

Based on these data, iterative adjustments are made to parameters in an often statistical model. In supervised ML, a model is trained upon both the input data x (obtained through one or many empirical procedures) and a corresponding target variable y (obtained through a reference procedure), and thereby "learns" to approximate a particular functional relationship. The precise ways in which models are trained based on these data vary. Broadly speaking, a learning algorithm, similar to the job of the measurement expert, adjusts parameters in an initial model based on differences between the reference data and the model's prediction.

What is more, the same statistical principle often underlies how adjustments are made to supervised ML models and measurement models. In poverty estimation applications, and in supervised ML more generally, adjustments to the parameters are often made based on the statistical principle of maximum likelihood estimation (Goodfellow et al., 2016, Chapter 5). Thereby, one adjusts the parameters of the model to maximize the likelihood so that, under the assumed statistical model, the observed data is most probable. The "learning" in supervised ML refers to the fact that the learning algorithm performs these adjustments *automatically*. While not necessarily executed automatically, the same statistical principle of maximum likelihood estimation finds application when adjusting measurement models, for instance, in econometrics (e.g., Holston et al., 2017).

Upon this first step of calibration, both model-based measurements and supervised ML models are then tested. To test ML models, programmers generally either set a part of the data aside before training or use cross-validation methods (Kubat, 2017, pp. 224–225). The goal of testing the ML model or model-based measurements is to assess its performance and to ensure that the functional relationship generalizes to the overall domain of interest. That is, the developers seek to test whether the model's inferences also cohere with reference measurements the model has not "seen" before. If the inferences prove to be unsatisfactory, both the ML developers and measurement experts go back to implement changes to the initial model specification, the previous calibration procedure, or the empirical process.

IV.3.3 Application as Measurement

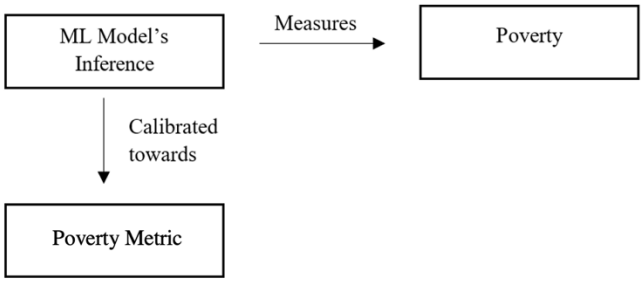
The trained and tested supervised ML models can then be employed to infer yet unknown values of poverty and, thereby, support policy decisions and further research. The underlying assumption is that dependent on sufficiently similar circumstances, the functional relationship between x and y that the model learned from the training data generalizes to the new instances. Given, for instance, satellite imagery or call data records and the probability distribution approximated by the trained and tested supervised ML model, one can then infer the poverty of regions beyond the ones the model was trained upon. This inference, I argue, is analogous to the second part of the inferential framework by which Tal characterizes measurement. Based on data obtained by one or many empirical processes (such as an instrumented measurement operation, surveying, or passive observation) and the measurement model previously established, one can, given sufficiently similar circumstances, *measure* unknown values of a property of interest.

Both calibrated measurement and ML-enabled poverty inference rely on a two-part inferential process that infers values of a property of interest that best cohere with the data under a model previously calibrated and thereby, indirectly, with the values of the property of interest previously provided by a reference procedure. Nonetheless, one commonly interprets measurements not as predicting hypotheticals

of the reference procedure but as *measurements* of the property of interest. As mentioned before, one might calibrate a watch towards the network of atomic clocks or a temperature sensor towards a reference thermometer. However, once in use, one would not conceive of these instruments as *predicting* the reference measurement but as *measuring* time or the temperature of an object. It is in similar fashion that I propose to understand supervised ML applications, by separating the object of calibration of supervised ML models from the epistemic aim of their application. Developers might *calibrate* their model towards a reference measurement of poverty. Once deployed, however, the model is intended to *measure* poverty rather than *predicting (hypothetical) values of a given poverty metric* (see Figure 2). For that reason and for all the similarities previously outlined, I argue for a perspective from which to interpret *supervised ML as automatically calibrated model-based measurement.*

Figure 2

Supervised ML as Automatically Calibrated Model-Based Measurement:



IV.3.4 Automatically Calibrated Model-Based Measurements

This difference in interpretation, I argue, is decisive and can help address the problems arising out of the predictive reframing of measurement tasks within ML previously outlined. Placing supervised machine learning applications within the frame of reference of the measurement problem that they are originally intended to and presented as solving internalizes precisely those concerns the predictive reframing circumvents. Thereby, the onus of ensuring the adequacy and validity of the machine learning model's outcomes relative to the ultimate property of interest is (at least partly) placed on the developers.

Once we conceive of these applications as measurements rather than predictions, the accuracy of the model's inferences can no longer be reduced to statistical correlation with a given dataset. Thinking of certain supervised ML applications as *automatically calibrated model-based* measurements rather than predictions brings them out of the frame of mere statistical calculus and into more critical engagement with the underlying theoretical and conceptual assumptions. This involves ethical and epistemic concerns about whether the model is the *right* measurement of the property of interest (in our example, poverty), for a given application. Consequently, machine learning developers must critically deliberate about what poverty index is appropriate, compare and combine multiple measurements, and clearly communicate potential limitations.

However, ideally one does not stop at ensuring the validity and adequacy of the reference measurement. Instead, developers must take into account the non-transitivity of the predictions, reference measurement, and construct to assess how the model's inferences, themselves, including differences between the model's measurements and the reference measurement, relate to the ultimate property of interest: in our example, poverty. Doing so is likely complicated by the complexity and limited mechanistic explanation provided by machine learning models (Chirimuuta, 2021; López-Rubio & Ratti, 2019). At the same time, is precisely for this reason that the developers themselves, possessing the most detailed understanding of their applications, are in the best position to do so.

Rather than merely illuminating problematic aspects of current practice in machine learning, a closer association of supervised machine learning and metrology can help expand machine learning developers' toolkit when working toward epistemically and ethically more successful applications. To evaluate how the machine learning outcomes, themselves, relate to the construct of interest, developers

could interact with the literature on validation in many domains of measurement.[10] Taking inspiration from methodologies in measurement, such as various forms of validation, can help facilitate closer engagement with theoretical and conceptual considerations of their applications.

Lastly, an interpretation of the results of supervised ML models as measurements asks ML developers to live up to their own claims — not only with respect to the proposed uses of their models but also to the self-understanding of the discipline as a whole. If ML claims to be more than glorified statistics (Davison, 2019), it must be willing to understand and evaluate its results as more than mere statistical predictions. This requires no less than a fundamental rethinking of the development and evaluation of supervised ML applications to measurement problems. Involved in that is more than I have covered here. Differences in the more implicit epistemic virtues, best practices, and the public's perception of measurement and prediction arguably factor into this proposed conceptual shift as much as a greater need for domain knowledge and expertise when developing and evaluating ML models.

V Conclusion

In this paper, I hope to have sketched (i) how ML developers reframe measurement tasks into prediction problems, (ii) how this predictive reframing can lead to problems by avoiding questions critical to the original measurement task, and (iii) offered an alternative interpretation of supervised ML applications as *automatically calibrated model-based measurements*. I argue that conceptualizing some supervised ML applications as measurements correctly identifies the epistemic aim of the original measurement task, places their development into the appropriate frame of reference, and properly internalizes critical questions, such as whether a certain model is the *right* measurement of a property.

While I have focused on the example of ML-enabled poverty "prediction," I believe that the predictive reframing of supervised ML applications is a rather common feature across disciplines. Other machine learning applications in the social sciences, such as psychometrics, are marketed as predicting anxiety, depression, and stress while only engaging minimally with the limitations of the questionnaire used (e.g., Priya et al., 2020). Similarly, machine learning models trained on videos labeled through Amazon's online crowdsourcing marketplace are presented as predicting personality traits (e.g., Ponce-López et al., 2016). The application of machine learning to measurement problems in the natural sciences might be subject to distinct epistemic and ethical considerations, necessitating its own philosophical examination. Nonetheless, the predictive reframing can also be observed here. In the earth sciences, machine learning applications trained on tracer-derived proxy measurements or simulated data, "predict" mean ages of shallow well samples (Green et al., 2021) or Mesozoic-Cenozoic precipitation (Chandra et al., 2021). In these and many other cases, an understanding of certain supervised machine learning applications as measurements might provide a helpful change in perspective.

Moreover, my argument relates in critical ways to current debates on epistemic and normative challenges surrounding ML. For instance, the predictive reframing is not merely informative for ML itself, but also provides an explanation for the importance of AI ethics research. Currently, a significant part of AI ethics addresses precisely the questions that developers, through the predictive reframing of their application, externalize from the development and evaluation of ML models. The proposed reconceptualization of some supervised ML applications as measurements, thus, might also shed light on how developers can internalize and integrate normative and epistemic issues critical to AI ethics.

What I set out to do in this paper can best be described as exploratory and introductory. I believe that much more can be gained by exploring analogies between measurement and ML within specific disciplines, by a more detailed analysis of the nature of calibration, measurement uncertainty, and evaluation, and by the transfer of historical insight on measurement to AI Ethics. With this paper, I hope to have introduced an initial framework for exploring these issues.

[10] For a particularly influential article on measurment validation, see Messick (1987). Recently, ML developers have started to engage with measurement and construct validation in the context of AI fairness (Jacobs, 2021; Jacobs & Wallach, 2021).

VI References

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Press.

Agrawal, A., Gans, J., & Goldfarb, A. (2020, September 1). How to Win with Machine Learning. *Harvard Business Review*. https://hbr.org/2020/09/how-to-win-with-machine-learning

Alexandrova, A. (2017). *A Philosophy for the Science of Well-being*. Oxford University Press.

Alexandrova, A., & Haybron, D. M. (2016). Is Construct Validation Valid? *Philosophy of Science*, *83*(5), 1098–1109. https://doi.org/10.1086/687941

Bell, J. (2014). *Machine Learning.* Wiley.

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, *350*(6264), 1073–1076. https://doi.org/10.1126/science.aac4420

Bokulich, A., & Parker, W. (2021). Data models, representation and adequacy-for-purpose. *European Journal for Philosophy of Science*, *11*(1), 31. https://doi.org/10.1007/s13194-020-00345-2

Boumans, M. (2007). Invariance and calibration. In M. Boumans (Ed.), *Measurement in Economics: A Handbook*. Emerald Group Publishing Limited.

Boumans, M. (2015). *Science Outside the Laboratory: Measurement in Field Science and Economics*. Oxford University Press.

Canali, S. (2016). Big Data, epistemology and causality: Knowledge in and knowledge out in EXPOsOMICS. *Big Data & Society*, *3*(2), 2053951716669530. https://doi.org/10.1177/2053951716669530

Chakraborty, T., Chakraborty, A. K., Biswas, M., Banerjee, S., & Bhattacharya, S. (2021). Unemployment Rate Forecasting: A Hybrid Approach. *Computational Economics*, *57*(1), 183–201. https://doi.org/10.1007/s10614-020-10040-2

Chandra, R., Cripps, S., Butterworth, N., & Muller, R. D. (2021). Precipitation reconstruction from climate-sensitive lithologies using Bayesian machine learning. *Environmental Modelling & Software*, *139*, 105002. https://doi.org/10.1016/j.envsoft.2021.105002

Chirimuuta, M. (2021). Prediction versus understanding in computationally enhanced neuroscience. *Synthese*, *199*(1), 767–790. https://doi.org/10.1007/s11229-020-02713-0

Davison, J. (2019, September 2). *No, Machine Learning is not just glorified Statistics*. Medium. https://towardsdatascience.com/no-machine-learning-is-not-just-glorified-statistics-26d3952234e3

Feest, U. (2020). Construct Validity in Psychological Tests? The Case of Implicit Social Cognition. *European Journal for Philosophy of Science*, *10*(1), 1–24. https://doi.org/10.1007/s13194-019-0270-8

Flach, P. (2012). *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, Massachusetts : The MIT Press.

Green, C. T., Ransom, K. M., Nolan, B. T., Liao, L., & Harter, T. (2021). Machine learning predictions of mean ages of shallow well samples in the Great Lakes Basin, USA. *Journal of Hydrology*, *603*, 126908. https://doi.org/10.1016/j.jhydrol.2021.126908

Hagendorff, T., & Meding, K. (2021). Ethical considerations and statistical analysis of industry involvement in machine learning research. *AI & SOCIETY*. https://doi.org/10.1007/s00146-021-01284-z

Hastie, T. J. (2009). *The elements of statistical learning: Data mining, inference, and prediction / Trevor Hastie, Robert Tibshirani, Jerome Friedman.* (Second edition.). Springer.

Holston, K., Laubach, T., & Williams, J. C. (2017). Measuring the natural rate of interest: International trends and determinants. *Journal of International Economics*, *108*, S59–S75. https://doi.org/10.1016/j.jinteco.2017.01.004

Jacobs, A. Z. (2021). Measurement as governance in and for responsible AI. *ArXiv:2109.05658 [Cs]*. http://arxiv.org/abs/2109.05658

Jacobs, A. Z., & Wallach, H. (2021). Measurement and Fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 375–385. https://doi.org/10.1145/3442188.3445901

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790–794. https://doi.org/10.1126/science.aaf7894

Kubat, M. (2017). *An Introduction to Machine Learning* (2nd ed. 2017.). Springer International Publishing.

López-Rubio, E., & Ratti, E. (2019). Data Science and Molecular Biology: Prediction and Mechanistic Explanation. *Synthese*, *4*, 1–26. https://doi.org/10.1007/s11229-019-02271-0

Mari, L., Wilson, M., & Maul, A. (2021). *Measurement across the Sciences: Developing a Shared Concept System for Measurement*. Springer International Publishing. https://doi.org/10.1007/978-3-030-65558-7

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A Survey on Bias and Fairness in Machine Learning. *ArXiv:1908.09635 [Cs]*. http://arxiv.org/abs/1908.09635

Messick, S. (1987). Validity. *ETS Research Report Series*, *1987*(2), i–208. https://doi.org/10.1002/j.2330-8516.1987.tb00244.x

Morrison, M. (2009). Models, measurement and computer simulation: The changing face of experimentation. *Philosophical Studies*, *143*(1), 33–57. https://doi.org/10.1007/s11098-008-9317-y

Nájera Catalán, H. E., & Gordon, D. (2020). The Importance of Reliability and Construct Validity in Multidimensional Poverty Measurement: An Illustration Using the Multidimensional Poverty Index for Latin America (MPI-LA). *The Journal of Development Studies*, *56*(9), 1763–1783. https://doi.org/10.1080/00220388.2019.1663176

Perez, A., Ganguli, S., Ermon, S., Azzari, G., Burke, M., & Lobell, D. (2019). Semi-Supervised Multitask Learning on Multispectral Satellite Images Using Wasserstein Generative Adversarial Networks (GANs) for Predicting Poverty. *ArXiv:1902.11110 [Cs]*. http://arxiv.org/abs/1902.11110

Pietsch, W. (2016). The Causal Nature of Modeling with Big Data. *Philosophy & Technology*, *29*(2), 137–171. https://doi.org/10.1007/s13347-015-0202-2

Pietsch, W. (2021). *On the Epistemology of Data Science: Conceptual Tools for a New Inductivism*. Springer Nature.

Pokhriyal, N., & Jacques, D. C. (2017). Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences*, *114*(46), E9783–E9792. https://doi.org/10.1073/pnas.1700319114

Ponce-López, V., Chen, B., Oliu, M., Corneanu, C., Clapés, A., Guyon, I., Baró, X., Escalante, H. J., & Escalera, S. (2016). ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results. In G. Hua & H. Jégou (Eds.), *Computer Vision – ECCV 2016 Workshops* (pp. 400–418). Springer International Publishing. https://doi.org/10.1007/978-3-319-49409-8_32

Prediction. (2021). In *OED Online*. Oxford University Press. http://www.oed.com/view/Entry/149860

Priya, A., Garg, S., & Tigga, N. P. (2020). Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms. *Procedia Computer Science*, *167*, 1258–1267. https://doi.org/10.1016/j.procs.2020.03.442

Ratti, E. (2020). What Kind of Novelties Can Machine Learning Possibly Generate? The Case of Genomics. *Studies in History and Philosophy of Science Part A*, *83*, 86–96. https://doi.org/10.1016/j.shpsa.2020.04.001

Santos, M. E., & Villatoro, P. (2020). The Importance of Reliability in the Multidimensional Poverty Index for Latin America (MPI-LA). *The Journal of Development Studies*, *56*(9), 1784–1789. https://doi.org/10.1080/00220388.2019.1663177

Smith, A. G., Han, E., Petersen, J., Olsen, N. A. F., Giese, C., Athmann, M., Dresbøll, D. B., & Thorup-Kristensen, K. (2020). *RootPainter: Deep Learning Segmentation of Biological Images with Corrective Annotation* (p. 2020.04.16.044461). https://doi.org/10.1101/2020.04.16.044461

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677.

Sullivan, E. (2019). Understanding from Machine Learning Models. *The British Journal for the Philosophy of Science*, *axz035*, Article axz035. https://doi.org/10.1093/bjps/axz035

Tal, E. (2017a). A Model-Based Epistemology of Measurement. In N. Mößner & A. Nordmann (Eds.), *Reasoning in Measurement* (pp. 233–253). Taylor & Francis.

Tal, E. (2017b). Calibration: Modelling the measurement process. *Studies in History and Philosophy of Science Part A*, *65–66*, 33–45. https://doi.org/10.1016/j.shpsa.2017.09.001

Tanaka, H., Nayebi, A., Maheswaranathan, N., McIntosh, L., Baccus, S. A., & Ganguli, S. (2019). From deep learning to mechanistic understanding in neuroscience: The structure of retinal prediction. *ArXiv:1912.06207 [Physics, q-Bio]*. http://arxiv.org/abs/1912.06207

United Nations. (2015). *Transforming Our World: The 2030 Agenda for Sustainable Development*. https://sustainabledevelopment.un.org/content/documents/21252030%20Agenda%20for%20Sustainable%20Development%20web.pdf

Wang, X., Feng, H., Xia, Q., & Alkire, S. (2016). On the relationship between income poverty and multidimensional poverty in China. *OPHI Working Papers*, *101*. https://ora.ox.ac.uk/objects/uuid:b520ab4c-1a5d-4440-aab2-a89a289f89aa

Xie, M., Jean, N., Burke, M., Lobell, D., & Ermon, S. (2016). Transfer Learning from Deep Features for Remote Sensing and Poverty Mapping. *ArXiv:1510.00098 [Cs]*. http://arxiv.org/abs/1510.00098

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S., & Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, *11*(1), 2583. https://doi.org/10.1038/s41467-020-16185-w

Zhao, T., Huang, H., Yao, X., Luo, J., & Fu, X. (2020). Predicting individual socioeconomic status from mobile phone data: A semi-supervised hypergraph-based factor graph approach. *International Journal of Data Science and Analytics*, *9*(3), 361–372. https://doi.org/10.1007/s41060-019-00195-z

Zhou, Z.-H. (2021). *Machine Learning*. Springer Nature.