The Prediction-Explanation Fallacy:
A Pervasive Problem in Scientific Applications of Machine Learning

Marco Del Giudice

University of New Mexico

[Preprint date: December 15, 2021 (v.3)]

Address correspondence to Marco Del Giudice, Department of Psychology, University of New Mexico. Logan Hall, 2001 Redondo Dr. NE, Albuquerque, NM 87131, USA; email: marcodg@unm.edu

Abstract

In this paper, I highlight a problem that has become ubiquitous in scientific applications of machine learning methods, and can lead to seriously distorted inferences about the phenomena under study. I call it the *prediction-explanation fallacy*. The fallacy occurs when researchers use prediction-optimized models for explanatory purposes, without considering the tradeoffs between explanation and prediction. This is a problem for at least two reasons. First, prediction-optimized models are often deliberately biased and unrealistic in order to prevent overfitting, and hence fail to accurately explain the phenomenon of interest. In other cases, they have an exceedingly complex structure that is hard or impossible to interpret, which greatly limits their explanatory value. Second, different predictive models trained on the same or similar data can be biased in different ways, so that multiple models may predict equally well but suggest conflicting explanations of the underlying phenomenon. In this note I introduce the tradeoffs between prediction and explanation in a non-technical fashion, present some illustrative examples from neuroscience, and end by discussing some mitigating factors and methods that can be used to limit or circumvent the problem.

Keywords: bias-variance tradeoff; explanation; machine learning; prediction; Rashomon effect.

In this paper, I want to highlight a problem that has become ubiquitous in scientific applications of machine learning (ML) methods. As far as I know, this problem has not yet been singled out for discussion in the literature; but it deserves to be named, clearly described, and widely understood by researchers, who are increasingly relying on ML techniques without always appreciating their limits and constraints.

In a nutshell, the *prediction-explanation fallacy* occurs when researchers use prediction-optimized models for explanatory purposes, without considering the tradeoffs between prediction and explanation. This is a problem for at least two connected reasons. First, in many typical applications of ML techniques, prediction-optimized models are deliberately biased and unrealistic in order to prevent overfitting, and hence may fail to accurately explain the phenomenon of interest (see Breiman, 2001; Shmueli, 2010; Yarkoni & Westfall, 2017). In other cases, the models have an exceedingly complex structure that is hard or impossible to interpret, which greatly limits their explanatory value. Second, different predictive models trained on the same or similar data can be biased in different ways, with the result that multiple models may predict equally well but suggest conflicting, mutually inconsistent explanations of the underlying phenomenon (Breiman, 2001; Hancox-Li, 2020).

The prediction-explanation fallacy can lead to distorted and misleading conclusions—not only about the results of a single analysis, but also about the robustness and replicability of the phenomenon under study. In what follows, I first introduce the tradeoffs between prediction and explanation in a non-technical fashion. I continue by presenting some illustrative examples from the neuroscience literature, and end with a brief discussion of mitigating factors and methods that can be used to limit or circumvent the problem.

Prediction ≠ Explanation

Researchers across disciplines are expanding their data analytic practices to include a variety of ML methods, from relatively simple techniques such as regularization and cross-validation to complex algorithms such as deep learning (for introductions see Berk, 2016; James et al., 2021; for deeper treatments see Efron & Hastie, 2016; Hastie et al., 2009). In the fields of psychology and neuroscience, ML tutorials and easy-to-use packages are multiplying due to high demand by researchers (e.g., Koul et al., 2018; Kumar et al., 2020; Rosenbusch et al., 2020; Yarkoni & Westfall, 2017). A key reason for the success of these methods is the fact that they often outperform classical statistical procedures when the goal is to predict new outcomes, generalizing beyond the particular dataset used for training (out-of-sample prediction). In comparison, classical procedures—which focus on building accurate models of the data-generating process and minimizing bias—tend to yield models that overfit the data at hand, and perform badly or fail to replicate when tested in new datasets (Breiman, 2001; Rosenbusch et al., 2020; Yarkoni & Westfall, 2017).

It goes without saying that predictive accuracy is a key advantage of ML techniques, and that predictive modeling can be remarkably useful in a variety of tasks. However, researchers who employ ML in their studies—or use the results of those studies as primary sources, often without a deep understanding of the methods—can easily forget that superior predictive performance and comes at a cost. The cost is that maximizing the predictive accuracy of a statistical model tends to sacrifice its ability to represent the underlying phenomenon in an accurate and interpretable fashion. Indeed, when accurate prediction is the only criterion of success, the correspondence (or lack thereof) between statistical models and reality becomes irrelevant; "for all practical purposes *there is no model responsible for the data*" (Berk, 2016, p. 25; emphasis mine). Prediction and explanation are conceptually different tasks, and often require different modeling approaches (Shmueli, 2010; see also Bzdok & Ioannidis, 2019); the resulting tradeoffs should be taken into account whenever predictive models are used in scientific applications.

Before continuing, note that here I use the term "explanation" in a broad sense, to include not only causal relations between variables, but also patterns of association that could be regarded as more "descriptive" than strictly explanatory (e.g., Mõttus et al., 2020; Shmueli, 2010). For example, studies that measure sex differences in a given domain or describe networks of correlations among traits (see below) are rarely guided by specific causal hypotheses, and may be largely or entirely agnostic as to the exact causal mechanisms involved. While causal explanation is typically the ultimate goal of scientific research, the careful description of association patterns is often a crucial intermediate step in the process. From a statistical point of view, both causal inference and pattern description depend on the correct specification of the underlying phenomenon; from the standpoint of theory development, they both benefit from model transparency and interpretability. For the purposes of this paper, I include both of them under the umbrella of (broad-sense) explanation.

Tradeoffs Between Prediction and Explanation in Machine Learning

To simplify a complicated issue, good scientific explanations should be accurate (i.e., they should correctly represent the structure of the underlying phenomenon) as well as interpretable (i.e., they should be transparent and parsimonious). Of course, these two properties may be in tension with each other, to the extent that simplifications and approximations make explanations more cognitively tractable. In this section I consider how both of them trade off with predictive accuracy in the context of ML applications (Fig. 1).

First and most relevant to the topic of this paper, tradeoffs between predictive and explanatory accuracy occur when better predictive performance is achieved at the cost of increased model bias. Models optimized to provide accurate explanations of a phenomenon tend to overfit the training data, and hence generalize poorly to new samples; conversely, strategies designed to avoid overfitting (e.g., regularization) inevitably introduce particular types of biases, which can improve a model's performance if appropriately chosen (Schaffer, 1993). Thus, "the 'wrong' model can sometimes predict better than the correct one" (Shmueli, 2010, p. 293), and:

> "From a statistical standpoint, it is simply not true that the model that most closely approximates the data-generating process will in general be the most successful at predicting real-world outcomes […] a biased, psychologically implausible model can often systematically outperform a mechanistically more accurate, but also more complex, model" (Yarkoni & Westfall., 2017, p. 1100).

This kind of tradeoff is most acute in *underparametrized* models, i.e., models that have fewer parameters than the number of data points in the training set. To understand why, it is useful to refer to a key concept of predictive modeling, the *bias-variance tradeoff*. As models become more flexible and complex (relative to the size of the dataset), they increasingly tend to overfit the training data; accordingly, their predictions become less biased but also more variable across datasets. Depending on the shape of the bias and variance functions, maximizing prediction accuracy may require a compromise between the two sources of error, and hence the introduction of biases that reduce the explanatory accuracy of the model (see e.g., James et al., 2021; Shmueli, 2010; Yarkoni & Westfall, 2017).

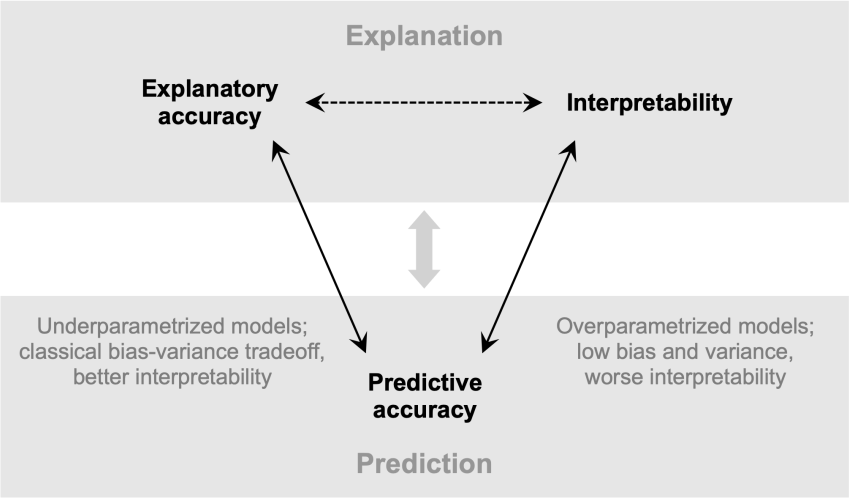
Explanation
Explanatory accuracy
Interpretability
Underparametrized models; classical bias-variance tradeoff, better interpretability
Predictive accuracy
Overparametrized models; low bias and variance, worse interpretability
Prediction

Figure 1. Tradeoffs between prediction and explanation in machine learning. Tradeoffs between predictive and explanatory accuracy are most acute in underparametrized models subject to the "classical" bias-variance tradeoff. Tradeoffs between predictive accuracy and interpretability become especially severe in overparametrized models. While not the main focus of this paper, explanatory accuracy and interpretability can also be in tension with each other (dashed arrow).

Tradeoffs between predictive accuracy and interpretability occur when models grow so complex that they are difficult (when not impossible) to parse and understand (see Breiman, 2001; James et al., 2021). While underparametrized models can become fairly large and intricate, tradeoffs of this kind are the rule when dealing with *overparametrized* models, which have more parameters than the number of training data points. As it has become clear over the last few years, highly overparametrized models—most notably neural networks, random forests, and other ensemble algorithms—often manage to "escape" the classical bias-variance tradeoff, and achieve low levels of bias while also avoiding overfitting. This happens in cases where, as models grow more complex, the prediction variance first increases—as in the classical scenario—but then begins to decrease after the number of parameters exceeds that of data points (see e.g., Belkin et al., 2019; Hastie et al., 2019; James et al., 2021; Yang et al., 2020).

To the extent that overparametrized models achieve low bias, they can be said to possess explanatory accuracy; however, their structure tends to be opaque and inscrutable, to the point that one needs specialized techniques to extract usable information and try to explain how they work (e.g., Biecek & Burzykowski, 2021; Linardatos et al., 2021; Molnar, 2019). Stated differently, these models manage to internally represent the structure of the underlying phenomenon, but typically do so in a convoluted and massively unparsimonious form. This greatly limits their explanatory value *as models of the phenomenon*, even when they can be used to extract usable information by indirect means. It is also the case that, in domains characterized by well-defined and meaningful variables, simple and interpretable models (e.g., logistic regression) often perform similarly to neural networks and other overparametrized "black boxes". Tradeoffs between predictive accuracy and interpretability are by no means inevitable, and tend to arise more often with certain types of models and problems than others (Rudin, 2019).

From a broader perspective, a variable may contribute to accurate prediction for reasons that have nothing to do with its theoretical importance or causal role in the explanation of a phenomenon. To illustrate, some variables may contribute to prediction more than others merely because they are measured with less error. Or, it is entirely possible for the same variable to improve the performance of a predictive model, but seriously distort the results of an explanatory model (for example because it acts as a collider for the effect of interest; see Elwert & Winship, 2014; Rohrer, 2018). In these and similar cases, predictive and explanatory accuracy come at each other's expense, regardless of the complexity and number of parameters of the models employed.

Researchers commit the prediction-explanation fallacy when they overlook the tradeoffs between prediction and explanation, and uncritically use prediction-optimized models as explanations of the underlying phenomena. Note the word "uncritically": using predictive models for explanatory purposes is not *necessarily* a problem; there are circumstances in which this approach is justified, and methods that allow researchers to circumvent the problem or at least reduce its severity (more on this below). For the same reasons, committing the fallacy does not automatically invalidate one's analysis, nor does it mean that one's interpretation of the results is necessarily wrong. The point is the tension between prediction and explanation should be explicitly taken into account by researchers, and addressed on a case-by-case basis.

Regularization as a Source of Bias

The fact that biased, oversimplified models such as unit-weighted regression can outperform their classical counterparts when used for prediction has been known for a long time (see e.g., Dawes, 1979; Hagerty & Srinivasan, 1991). Newer regularization techniques (for example the LASSO and elastic net; see Hastie et al., 2009; James et al., 2021) offer powerful ways to shrink (i.e., strategically bias) the model coefficients while simultaneously selecting an optimal subset of variables. They do so by making sparsity assumptions, with the expectation that sparse models (i.e., models with relatively few nonzero coefficients) will perform well in prediction even if the data-generating process is not actually sparse (this is known as the "bet on sparsity"; see Hastie et al., 2009). Of course, if the data-generating process *does* happen to be sparse, these techniques can effectively filter out sampling noise and yield models with high predictive *and* explanatory accuracy;[1] but this is not true in general, and researchers routinely

[1] This is directly related to the Bayesian interpretation of regularization methods; for example, the LASSO is equivalent to a Laplace prior on the distribution of model parameters, while ridge regularization is equivalent to a

apply regularization to domains in which sparsity is not a plausible assumption. In sum, regularization (especially when it involves sparsity) can be a major source of bias in predictive models, and this should be taken into account when evaluating their adequacy as explanations.

The same logic applies to models whose purpose is not to predict a specific outcome but to describe relations among sets of variables. For example, network models have become quite popular in psychopathology and personality psychology, where they are used to investigate the relations between multiple symptoms, traits, and/or behaviors (se Costantini et al., 2015; Epskamp et al., 2018; McNally, 2016). Researchers in these fields typically use specialized versions of the LASSO to reduce the number of nonzero connections (edges) in the estimated networks. This is done with two distinct purposes: the first is to eliminate "spurious" edges and simplify the interpretation of the results (i.e., increase explanatory accuracy and interpretability); the second is to improve the generalizability of the estimated networks across different samples (analogous to out-of-sample prediction; see Epskamp & Fried, 2018; Epskamp et al., 2017). Unfortunately, these goals can be in conflict with one another. Since the LASSO is based on the assumption of sparsity, it tends to return sparse networks regardless of the underlying structure of the variables, especially when sample size is small for the number of parameters in the model; thus, finding a sparse network after regularization is not convincing evidence that the true structure is actually sparse (Epskamp & Fried, 2018; Epskamp et al., 2017). Regularized networks can successfully recover the underlying structure of the variables even in relatively small samples (and thus achieve high explanatory accuracy in addition to generalizability) *if* that structure happens to be sparse; otherwise, they may introduce significant biases and suggest distorted explanations of the phenomenon under study. Failing to understand this problem can lead researchers to commit a variant of the prediction-explanation fallacy, as they use regularized networks to describe the structure of the variables without considering the biases they introduce.

Many Models, Many Explanations: The Rashomon Effect

The tension between prediction and explanation has an important corollary, known as the *Rashomon effect* (Breiman, 2001): for a given prediction problem, there is often a multitude of models that predict about equally well, but each models tells a different story about which variables are important and/or how they are related. In other words, the models are equally good for prediction, but suggest different, mutually inconsistent explanations of the same phenomenon (Dong & Rudin, 2020; Hancox-Li, 2020). This should not come as a surprise; the point is that different models may achieve the same predictive performance by implementing different biases (i.e., finding different but equally useful ways to be "wrong").

Gaussian prior (see James et al., 2021; McElreath, 2020). In some disciplines, researchers who use Bayesian methods are increasingly choosing weakly regularizing priors over diffuse priors, with the goal of increasing model robustness and avoiding overfitting (e.g., Lemoine, 2019). Besides improving prediction (see McElreath, 2020), regularizing priors can improve the explanatory accuracy of the models to the extent that they introduce realistic assumptions (i.e., they match the actual structure of the phenomenon). However, there is a risk of smuggling in a version of the prediction-explanation fallacy if researchers follow this practice unthinkingly, without understanding and discussing the relevant tradeoffs. In passing, note that "bias" is an intrinsically frequentist concept; in Bayesian statistics, parameters are not fixed quantities but have probability distributions. The tradeoffs between explanation and prediction do not disappear in Bayesian statistics, but they have to be framed in somewhat different terms.

The Rashomon effect comes in two flavors. On the one hand, models based on different algorithms and functional forms (e.g., logistic regression, classification trees, neural networks) can perform very similarly to one another when trained and tested on the same data. On the other hand, even using a single type of model can yield unstable results, because small changes in the data or in the tuning parameters can have a dramatic impact on which variables get selected and/or on the model coefficients (see Breiman, 2001). For example, regularization techniques deal with multiple redundant variables by excluding some of them from the model, or shrinking their coefficients by a large amount; but precisely *which* variables end up being excluded or deemphasized in a particular model may depend on minor fluctuations of the data.

A notable consequence of the Rashomon effect is that when different types of models are trained on the same dataset, they may easily identify different sets of variables as being "important" for prediction. Even training the same type of model on similar datasets may yield contradictory accounts of the importance of variables—not because the phenomenon under study lacks consistency, but because prediction-optimized models tend to be unstable (in the sense explained earlier). When models trained on the same or similar data appear to suggest markedly inconsistent explanations of a phenomenon, it can be tempting to conclude that the phenomenon *itself* is not robust; however, this is just an insidious manifestation of the prediction-explanation fallacy in the context of multiple models.

Explanation and Prediction in Surrogate Models

In the attempt to understand the workings of opaque black box models, researchers sometimes use global "surrogate models" (see e.g., Molnar, 2019). A global surrogate is just a simpler, interpretable model that is trained to *predict the predictions* of the original model. For example, imagine that a complex neural network was trained on a dataset to predict a binary outcome. Researchers could then train a logistic regression model on the same dataset, but instead of predicting the original outcome, the surrogate would be trained to predict the predictions made by the neural network. This simpler model could then be probed for insights into the workings of the original model. A global surrogate aims to reproduce the output of an entire model, in contrast with "local surrogates" that focus on individual predictions and try to explain how the model arrived at them (Biecek & Burzykowski, 2021; Linardatos et al., 2021; Molnar, 2019). Of course, the success of this strategy depends on the ability of the surrogate to approximate the functional relations between variables in the original model.

What is easily missed is that surrogate models are subject to all the tradeoffs discussed in this section; thus, improving the predictive accuracy of a surrogate model—for example by regularization and/or cross-validation—will tend to make it less accurate as an explanation of the original model. This is a problem because the purpose of surrogate models is intrinsically explanatory. Moreover, training multiple surrogate models on the same or similar data can lead to the Rashomon effect: different surrogates may seem to explain the original models about equally well, but suggest multiple, inconsistent explanations of how they work. Overlooking these issues can lead to "second-order" instances of the prediction-explanation fallacy, which may be particularly hard to detect and correct.

Illustrative Examples

Sex Differences in Brain Structure

In recent years, there has been an explosion of studies employing predictive ML methods to distinguish males and females based on their brain anatomy (e.g., data on cortical volume, thickness, or three-dimensional morphology). Classification accuracy is typically above 90%, but drops to 60-70% when differences in total brain volume are controlled for (reviewed in Eliot et al., 2021). In view of the high predictive accuracy achieved by these models, it can be tempting to use them to determine which regions of the brain contribute the most to differentiating males and females—a fertile ground for the prediction-explanation fallacy in all its forms.

For a clear-cut example of the fallacy, consider the study by Luo et al. (2019). These authors aimed to answer two questions: "(a) can gender be discriminated with a high accuracy using cortical 3-D morphology? (b) What is the most discriminative region of gender in cortical 3-D morphology?" (p. 2). To this end, they trained a hierarchical sparse representation classifier on cortical morphology data, achieving 97% accuracy. Then, they used bootstrapped model weights to identify "important 3-D morphological features in gender discrimination" (p. 7). A brain map of the discriminative regions (their Fig. 4) showed a highly sparse configuration; this is not surprising, given the strong sparsity assumptions built into the model (pp. 4-5). This study exemplifies the fallacy because the authors went straight from training a predictive model to making statements about the most important differences between male and female brains, e.g.: "The main morphology difference for gender exists mainly in the frontal lobe and the limbic lobe, others scattered in the parietal lobe, the temporal lobe, the corpus callosum and the precuneus" (p. 7). They did not discuss how their modeling decisions might have biased the analysis, or caution readers against incorrect interpretations of their results. It can be useful to restate that committing the prediction-explanation fallacy does not automatically invalidate the results of a study; however, it does raise questions about their interpretation, and may challenge the validity of the inferences drawn by the authors.

A neuroimaging study by Anderson et al. (2019) illustrates a diametrically opposite approach to the tradeoffs between prediction and explanation. These authors applied independent component analysis (ICA) to cortical volume and density data, and used the resulting components to train and compare a number of predictive models. Logistic regression and support vector machines (SVM) performed best, with a classification accuracy of 93%. However, the authors did not plot the model weights or use them to identify brain regions that discriminate between males and females; instead, they presented descriptive maps of sex-differentiated regions based on the results of ICA (their Fig. 1 and 2). This study avoided the prediction-explanation fallacy by restricting the use of predictive models to the classification task. Note that this is not necessarily the optimal strategy; depending on context, careful consideration of the weights of predictive models can provide useful information and complement the results of more descriptive analyses. Another possibility is to deliberately fit different types of models to the data, some optimized for prediction and others for accurate description/explanation. For example, Sepehrband et al. (2018) analyzed sex differences in cortical structure with two models—a SVM and a standard general linear model (GLM)—and explicitly compared their results, while taking care to note the different goals of the two analyses. Although model weights were generally concordant, several regions showed high discriminatory power in the SVM but no significant sex differences in the GLM (see their Tab. 2). Those regions could be promising candidates for follow-up analyses, because the SVM algorithm might have picked up complex interaction patterns that would have been missed by the simple GLM used in the study.

As I noted earlier, the prediction-explanation fallacy does not only apply to the results of individual models, but also to comparisons between multiple models and studies. As part of their critical review of the research on sex differences in the brain, Eliot et al. (2021) compiled a dozen of studies that had used ML methods to predict a person's sex from patterns of brain structure and function (see their Tab. 7). They noted:

> "[T]he studies […] differ strikingly in features found to be most important for [sex/gender] classification accuracy. Of course, one would not expect similar features to emerge between studies using qualitatively different data, such as rsfMRI activity versus regional gray matter volumes. But even among studies that relied exclusively on structural measures, we see a lack of replication among the brain regions identified as most important for male/female classification across studies" (p. 681).
>
> And concluded:
>
> "[T]his discrimination is largely based on brain size and there is no agreement about local features that are most important for distinguishing male versus female types. The lack of hallmark 'male' versus 'female' brain features is likely because each algorithm was custom-developed for its particular dataset. […] These findings challenge the notion that there exists a discrete set of variables that capture core differences between male and female brains across the human species." (p. 681).

However, variability in the "important features" identified across models and samples—even with similar data and similar levels of predictive accuracy—is expected as a manifestation of the Rashomon effect, and cannot be used to draw simple conclusions about the existence (or lack thereof) of reliable differences between male and female brains. While local overfitting may have contributed to inflate the variability of these findings (as hinted at in the passage above), one should not expect high levels of consistency to begin with; in fact, more aggressive strategies to reduce overfitting may even *exacerbate* the Rashomon effect instead of reducing it (see Breiman, 2001; Schaffer, 1993). Note that I am not arguing that Eliot et al.'s conclusions are necessarily *false*; my point is that they are not warranted by the observation that different models rely on different sets of brain features for prediction.

My last example for this section is an interesting, widely circulated preprint by Sanchis-Segura et al. (2021). These authors explored the issue of sex differences in brain structure with a variety of descriptive and predictive methods. In one of the analyses, they used a dataset of gray matter volume to train five different classifiers: two "classical" models fit without regularization or cross-validation (logistic regression and linear discriminant analysis [LDA]), and three prediction-optimized models (SVM, random forests, and multiple adaptive regression splines [MARS]). The classifiers achieved similar levels of accuracy (86-90% without correcting for total brain volume, 59-66% in a volume-corrected dataset), and were used to generate five estimates of the “probability of being classified as male” (PCAM) for each participant. Then, the authors trained another set of predictive models (boosted beta regression) that used regional volumes to predict the PCAM scores generated by the classifiers, and compared regression weights across the resulting models. In other words, the authors used beta regression models as global surrogates, to identify the brain features that contributed most to prediction in the original classifiers and quantify their relative importance. Despite the high correlations between the five PCAM scores (the average correlation was .87 without correcting for total brain volume, .70 in a corrected dataset; see their Fig. 7D), the relative importance of different brain regions varied substantially across classifiers, yielding low levels of consistency by most measures (pp. 6-7; see their Fig. 5 and 6).[2]

The authors correctly noted that “because they differ in their statistical assumptions and operations, distinct algorithms rely on distinct brain features […] and assign different PCAM scores to the same subjects” (p. 7). But then they went on to write:

> “Therefore, it is apparent that—despite working with identical data from the same individuals—the different algorithms tested in the present study do not provide directly exchangeable outcomes or identify a single, coherent, and reproducible subset of brain features as the source of the males-females multivariate differences […] Together, these sources of empirical evidence directly challenge the binary sex views of human brains […] these views assume that, because distinct ML algorithms are able to correctly ‘predict’ sex from neuroanatomical features in 80–90% of the cases, all these algorithms must be identifying two distinct brain types in the human species, one typical of males and the other typical of females […]. However, these universal ‘brain types’ do not seem to really exist, given that different algorithms identify distinct brain features as the landmarks of ‘male/ female brains’ in different samples of females and males and when applied to the same subjects” (pp. 7-8).

In their interpretation of the results, the authors commit two instances of the prediction-explanation fallacy. First and more obviously, they read the lack of concordance among models as evidence against the existence of universal male/female “brain types”; because different classifiers can be *expected* to rely on different sets of predictors, even when trained on the same data, this inference is unwarranted.[3] The second fallacy concerns the use of boosted beta regression to infer the relative importance of brain features according to different classifiers. This algorithm uses gradient boosting and cross-validation to select an optimal subset of variables for prediction (Schmid et al.., 2013). The resulting surrogate models can be expected to maximize predictive performance at the expense of explanatory accuracy, and to show a degree of instability when faced with many redundant predictors. In other words, different regression

[2] Of note, the authors also calculated aggregate levels of consistency, identified a list of top predictors across classifiers (see their Fig. 5 and 6), and suggested that comparing the results of several algorithms should lead to more valid conclusions than focusing on any single one of them (p. 9). These are all useful strategies to reduce the problems associated with prediction-explanation tradeoffs.

[3] Of course, there are many other reasons—theoretical as well as empirical—to reject the simplistic idea that there are only two homogeneous “brain types”, one for male and one for females; see for example Del Giudice (2021), Joel (2021), and other findings in Sanchis-Segura et al. (2021).

models trained on similar PCAM scores may select somewhat different sets of "important predictors" for reasons that have nothing to do with a lack of consistency between the original classifiers. It remains unclear to what extent the discordance on display in Sanchis-Segura et al.'s Fig. 5 and 6 is due to actual differences among the classifiers, or to instability in the regression models used to explain their functioning. These surrogates identified similar sets of important predictors for logistic regression and LDA, which is encouraging given the strong similarity between these algorithms (see James et al., 2021). At the same time, the PCAM scores produced by the two classifiers correlated at .99, making this a limit case of almost perfect consistency.

Neural Correlates of Emotions

Most biological theories of emotions (e.g., Ekman, 1999; Panksepp, 1998) postulate the existence of brain mechanisms specialized to produce specific emotional response. According to these theories, the experience of different emotions—such as happiness, anger, or disgust—should correlate with somewhat distinctive patterns of brain activity ("neural signatures"). The existence of such signatures should make it possible to accurately predict the emotional state of a person from measures of his/her brain activity. By applying ML methods to functional neuroimaging data, researchers have been able to classify participants' emotional states into discrete categories with significant accuracy (e.g., Kassam et al., 2013; Kragel & LaBar, 2015; Saarimäki et al., 2016).[4]

Some of the studies in this area can serve to illustrate the prediction-explanation fallacy in its subtler forms. For example, Kragel and LaBar (2015) trained a set of partial least square discriminant analysis models, and employed cross-validation to select the number of latent variables. They then used model coefficients to identify the brain voxels that contributed most strongly to predicting each specific emotion (see their Fig. 3). The authors noted that the voxels selected by the predictive models overlapped only in part with those that showed significant differences in univariate GLMs. They noted that the GLMs were less sensitive to patterns spread across multiple voxels and more prone to overfitting noise at the single-voxel level (p. 1446), but did not discuss the respective biases and limitations of the two types of models in any detail. Most importantly, they presented their results in a way that blurred the line between prediction and explanation. For instance:

> "Our analysis of regression coefficients revealed that this information was contained within diverse patterns of activation, spanning a number of cortical and subcortical brain regions. […] Maps for contentment included precuneus, medial prefrontal, cingulate and primary somatosensory cortices […] Despite engaging partially overlapping neural substrates at the macro-scale, emotion-predictive patterns were largely non-overlapping at the voxel level. Such separability of emotional states at the voxel level may explain why meta-analytic works […] have associated neural activity with discrete emotions

[4] Studies in this area tend to be based on very small samples (*N* = 10 to 48 for the three studies cited here), which raises legitimate concerns about generalizability and replicability. Because (a) I am using these studies as illustrative examples of a conceptual point, and (b) the argument would still apply if the studies were based on larger samples, I will not discuss these issues further. Also, I will not discuss the other potential limitations of this kind of study (e.g., the pitfalls of trying to identify specific brain mechanisms based on folk labels such as "anger" or "fear"; see Scarantino, 2012).

> (e.g., correspondence between activation within the amygdala and fear or dorsal anterior cingulate and happiness), yet have failed to consistently identify emotion-specific neural substrates" (pp. 1444-1445).

In a similar study, Saarimäki et al. (2016) trained neural networks to classify emotional experiences into discrete categories, and used model weights to calculate and plot the "importance values" associated with each brain voxel (their Fig. 3). Even if they did not discuss the tradeoffs between prediction and explanation, these authors were careful to explicitly frame their results in terms of prediction, and managed to avoid confusions between the two domains throughout most of the paper. However, at various points of the discussion section they seemed to equivocate between the sets of predictive voxels used by the neural networks and the broader (explanatory) concept of neural signatures. For example:

> "Our results reveal that basic emotions are supported by discrete neural signatures within several brain areas, as evidenced by the high classification accuracy of emotions from hemodynamic brain signals. […] The distributed emotion-specific activation patterns may provide maps of internal states that correspond to specific subjectively experienced, discrete emotions […] In our study, the medial prefrontal and medial posterior regions […] contributed most significantly to classification between different basic emotions […]. Thus, local activation patterns within these areas differ across emotions and thus presumably reflect distinct neuronal signatures for different emotions" (pp. 7-8).

To the extent that this qualifies as a case of prediction-explanation fallacy, it is as mild and benign as it gets. However, lacking an explicit discussion of the explanatory limits of the analysis, readers less careful than the authors may easily look at the results and draw unwarranted conclusions about the "signatures" of different emotions and their localization.

In contrast with classical emotion theorists, constructivist scholars deny the existence of specialized emotion mechanisms in the brain. An influential example of this approach is the *theory of constructed emotion* (Barrett, 2017). According to the theory, emotions consist of complex and highly variable patterns of sensory inputs, interoceptive sensations, facial movements, and so on; these patterns do not *arise* from the activity of dedicated mechanisms, but instead get *categorized* as instances of a certain emotion through the incessant concept-forming activity of the brain. One implication of this view is that emotions should not be associated with specific neural signatures. In reviewing the empirical data in support of the theory, Barrett (2017) wrote:

> "Ironically, perhaps the strongest evidence to date for the theory comes from studies that use pattern classification to distinguish categories of emotion. Several recent articles taking this approach have reported success in differentiating one emotion category from another—a finding that is routinely construed as providing the long awaited support for the classical view (Kassam et al., 2013; Kragel and LaBar, 2015; Saarimäki et al., 2016). However, patterns that distinguish among the categories in one study do not replicate in the other studies" (p. 15).

The deeper irony of this passage is that the author's interpretation of the literature is a conspicuous example of the prediction-explanation fallacy. The regions/voxels identified as most predictive can be expected to vary from one study to the next, even if the underlying patterns of brain activity are stable and consistent. (Of course, the variability is going to be magnified if the studies are based on small samples, as in this case.) Discordances between prediction-optimized models are par for the course; it is a mistake to conclude that a phenomenon lacks robustness just because different predictive models fail to agree with one another. Naturally, this problem does not automatically falsify Barrett's theory; however, it does challenge the notion that the theory receives strong support from the sheer variability of the results of brain imaging studies.

Mitigating Factors and Strategies

Throughout this paper I have emphasized the differences and tradeoffs between prediction and explanation, but it is important to stress once again that these goals are not always or necessarily in tension. The biases introduced to maximize prediction accuracy can also improve explanatory accuracy if they happen to match the structure of the phenomenon under study. For example, algorithms that make sparsity assumptions can generate accurate explanations—and potentially achieve lower bias than more flexible algorithms—when the data-generating process is actually sparse (e.g., Epskamp et al., 2017). Thus, using a prediction-optimized model for explanatory purposes can be justified if there are theoretical and/or empirical reasons to believe that the assumptions of the model are closely matched to the structure of the underlying phenomenon. From a different angle, it may be possible to gain insights about the structure of a phenomenon precisely by comparing the performance of alternative models that incorporate different assumptions and biases (e.g., sparsity vs. density; see Yarkoni & Westfall, 2017). One should also consider that regularization has a stronger impact when sample size is small for the number of model parameters; thus, some of the tradeoffs I discussed here tend to become less severe when working with large datasets.

As I noted earlier, one strategy that can be used to avoid the prediction-explanation fallacy is to fit different types of models to the data, some optimized for prediction and other for accurate explanation. If done with care, comparisons between different types of models can be illuminating. Another option is to avoid focusing on a single best-performing model, and instead capitalize on the Rashomon effect by training a *set* of well-performing models, each with somewhat different explanatory biases. The set can then be examined as an ensemble (for example by calculating the average importance of each variable across all the models; see e.g., Sanchis-Segura et al., 2021), and used to draw explanatory inferences that are often going to be less biased and more accurate than those provided by any individual model. This strategy works best when researchers take steps to decorrelate the models in the set, purposefully making them different from one another so that their idiosyncratic biases will tend to cancel out in the aggregate. A well-known example is the random subspace method employed in random forests, in which each model is trained on a different subset of data *and* predictors (see Berk, 2016; James et al., 2021; Kostiantis, 2014). In contrast, training a large number of models on small variations of the same data (for example by bootstrapping the dataset without sampling from the predictors; e.g., Luo et al., 2019) will do relatively little to decorrelate them, and explanatory biases will usually fail to cancel out.

Finally, ML researchers are developing specialized methods to explore model variability in a more systematic fashion. A recent example is the work on *variable importance clouds* (Dong & Rudin, 2020), a visualization technique that maps the importance of each variable across the "Rashomon set", that is, the full set of (approximately) equally accurate predictive models of a given class. Variable importance clouds go beyond aggregate estimates of importance and can reveal the existence of explanatory tradeoffs between variables, so that when one of the variables has high importance in a model, the other tends to have low importance (and vice versa).

Conclusion

Scientific applications of ML techniques can be extremely powerful; they also raise new problems and complications, both in their use and in the interpretation of their results. One of these problems—which seems to be particularly widespread—is the uncritical use of prediction-optimized models for explanatory purposes. Here I tried to pinpoint this fallacy, explain it in simple terms, and give it a convenient and descriptive name. The prediction-explanation fallacy can take a number of related forms; it can range from mild, ambiguous cases to glaring misinterpretations of the information provided by predictive models. The solution is not to prescribe that one should *never* use predictive models for explanation; that would be just a different sort of fallacy. Instead, researchers should explicitly address the tension between explanation and prediction in their analyses, consider potential mitigating factors, and—when feasible—use appropriate strategies to limit or circumvent the problem. I hope this note will contribute to improve the applied use of ML by helping researchers run more transparent, informative analyses and avoid drawing misleading conclusions from the data.

Acknowledgments

I wish to thank Andrew Cutler and Carla Sanchis-Segura for their generous feedback on earlier drafts of this paper.

References

Anderson, N. E., Harenski, K. A., Harenski, C. L., Koenigs, M. R., Decety, J., Calhoun, V. D., & Kiehl, K. A. (2019). Machine learning of brain gray matter differentiates sex in a large forensic sample. *Human Brain Mapping, 40,* 1496-1506. https://doi.org/10.1002/hbm.24462

Barrett, L. F. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience, 12,* 1-23. https://doi.org/10.1093/scan/nsw154

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences USA, 116,* 15849-15854. https://doi.org/10.1073/pnas.1903070116

Berk, R. A. (2016). *Statistical learning from a regression perspective*. Springer.

Biecek, P., & Burzykowski, T. (2021). *Explanatory model analysis: explore, explain, and examine predictive models.* CRC Press.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16,* 199-231. https://doi.org/10.1214/ss/1009213726

Bzdok, D., & Ioannidis, J. P. (2019). Exploration, inference, and prediction in neuroscience and biomedicine. *Trends in Neurosciences, 42,* 251-262. https://doi.org/10.1016/j.tins.2019.02.001

Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mõttus, R., Waldorp, L. J., & Cramer, A. O. (2015). State of the aRt personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality, 54,* 13-29. https://doi.org/10.1016/j.jrp.2014.07.003

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571-582. https://doi.org/10.1037/0003-066X.34.7.571

Del Giudice, M. (2021). Binary thinking about the sex binary: A comment on Joel (2021). *Neuroscience and Biobehavioral Reviews, 127,* 144-145. https://doi.org/10.1016/j.neubiorev.2021.04.020

Dong, J., & Rudin, C. (2020). Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence, 2,* 810-824. https://doi.org/10.1038/s42256-020-00264-0

Efron, B., & Hastie, T. (2016). *Computer age statistical inference*. Cambridge University Press.

Ekman, P. E. (1999). Basic emotions. In T. Dalgleish & T. Power (Eds.), *Handbook of cognition and emotion* (pp. 45–60). Wiley.

Eliot, L., Ahmed, A., Khan, H., & Patel, J. (2021). Dump the "dimorphism": Comprehensive synthesis of human brain studies reveals few male-female differences beyond size. *Neuroscience & Biobehavioral Reviews, 125,* 667-697. https://doi.org/10.1016/j.neubiorev.2021.02.026

Elwert, F., & Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology, 40,* 31-53. https://doi.org/10.1146/annurev-soc-071913-043455

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods, 50,* 195-212. https://doi.org/10.3758/s13428-017-0862-1

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods, 23*, 617-634. https://doi.org/10.1037/met0000167

Epskamp, S., Kruis, J., & Marsman, M. (2017). Estimating psychopathological networks: Be careful what you wish for. *PLoS ONE, 12,* e0179891. https://doi.org/10.1371/journal.pone.0179891

Hagerty, M. R., & Srinivasan, V. (1991). Comparing the predictive powers of alternative multiple regression models. *Psychometrika, 56,* 77-85. https://doi.org/10.1007/BF02294587

Hancox-Li, L. (2020). Robustness in machine learning explanations: Does it matter? In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 640-647). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372836

Hastie, T., Montanari, A., Rosset, S., & Tibshirani, R. J. (2019). Surprises in high-dimensional ridgeless least squares interpolation. *arXiv,* 1903.08560. https://arxiv.org/abs/1903.08560v4

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning with applications in R* (2nd ed.). Springer.

Joel, D. (2021). Beyond the binary: Rethinking sex and the brain. *Neuroscience and Biobehavioral Reviews, 122,* 165–175. https://doi.org/10.1016/j.neubiorev.2020.11.018

Kassam, K. S., Markey, A. R., Cherkassky, V. L., Loewenstein, G., & Just, M. A. (2013). Identifying emotions on the basis of neural activation. *PLoS ONE, 8,* e66032. https://doi.org/10.1371/journal.pone.0066032

Kotsiantis, S. B. (2014). Bagging and boosting variants for handling classifications problems: A survey. *The Knowledge Engineering Review, 29,* 78-100. https://doi.org/10.1017/S0269888913000313

Koul, A., Becchio, C., & Cavallo, A. (2018). *PredPsych:* A toolbox for predictive machine learning-based approach in experimental psychology research. *Behavior Research Methods, 50,* 1657-1672. https://doi.org/10.3758/s13428-017-0987-2

Kragel, P. A., & LaBar, K. S. (2015). Multivariate neural biomarkers of emotional states are categorically distinct. *Social Cognitive and Affective Neuroscience, 10,* 1437-1448. https://doi.org/10.1093/scan/nsv032

Kumar, M., Ellis, C. T., Lu, Q., Zhang, H., Capotă, M., Willke, T. L., ... & Norman, K. A. (2020). BrainIAK tutorials: User-friendly learning materials for advanced fMRI analysis. *PLoS Computational Biology, 16,* e1007549. https://doi.org/10.1371/journal.pcbi.1007549

Lemoine, N. P. (2019). Moving beyond noninformative priors: Why and how to choose weakly informative priors in Bayesian analyses. *Oikos, 128,* 912-928. https://doi.org/10.1111/oik.05985

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy, 23,* 18. https://dx.doi.org/10.3390/e23010018

Luo, Z., Hou, C., Wang, L., & Hu, D. (2019). Gender identification of human cortical 3-D morphology using hierarchical sparsity. *Frontiers in Human Neuroscience, 13*, 29. https://doi.org/10.3389/fnhum.2019.00029

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan.* CRC Press.

McNally, R. J. (2016). Can network analysis transform psychopathology? *Behaviour Research and Therapy, 86,* 95-104. https://doi.org/10.1016/j.brat.2016.06.006

Molnar, C. (2019). *Interpretable machine learning. A guide for making black box models explainable.* https://christophm.github.io/interpretable-ml-book/

Mõttus, R., Wood, D., Condon, D. M., Back, M. D., Baumert, A., Costantini, G., ... & Zimmermann, J. (2020). Descriptive, predictive and explanatory personality research:

Different goals, different approaches, but a shared need to move beyond the Big Few traits. *European Journal of Personality, 34*, 1175-1201. https://doi.org/10.1002/per.2311

Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. Oxford University Press.

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science, 1,* 27-42. https://doi.org/10.1177/2515245917745629

Rosenbusch, H., Soldner, F., Evans, A. M., & Zeelenberg, M. (2021). Supervised machine learning methods in psychology: A practical introduction with annotated R code. *Social and Personality Psychology Compass, 15,* e12579. https://doi.org/10.1111/spc3.12579

Rudin, C. (2019). Please stop explaining black box models for high stakes decisions. *ArXiv*, https://arxiv.org/abs/1811.10154v3

Saarimäki, H., Gotsopoulos, A., Jääskeläinen, I. P., Lampinen, J., Vuilleumier, P., Hari, R., ... & Nummenmaa, L. (2016). Discrete neural signatures of basic emotions. *Cerebral Cortex, 26,* 2563-2573. https://doi.org/10.1093/cercor/bhv086

Sanchis-Segura, C., Aguirre, N., Cruz-Gómez, Á. J., Félix, S., & Forn, C. (2021). Beyond "sex prediction": Estimating and interpreting multivariate sex differences and similarities in the brain. *ResearchSquare*, https://doi.org/10.21203/rs.3.rs-741734/v1

Scarantino, A. (2012). How to define emotions scientifically. *Emotion Review, 4,* 358–368. https://doi.org/10.1177/1754073912445810

Schaffer, C. (1993). Overfitting avoidance as bias. *Machine Learning, 10,* 153-178. https://doi.org/10.1023/A:1022653209073

Schmid, M., Wickler, F., Maloney, K. O., Mitchell, R., Fenske, N., & Mayr, A. (2013). Boosted beta regression. *PLoS ONE, 8,* e61623. https://doi.org/10.1371/journal.pone.0061623

Sepehrband, F., Lynch, K. M., Cabeen, R. P., Gonzalez-Zacarias, C., Zhao, L., D'Arcy, M., ... & Clark, K. A. (2018). Neuroanatomical morphometric characterization of sex differences in youth using statistical learning. *Neuroimage, 172,* 217-227. https://doi.org/10.1016/j.neuroimage.2018.01.065

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*, 289-310. https://doi.org/10.1214/10-STS330

Yang, Z., Yu, Y., You, C., Steinhardt, J., & Ma, Y. (2020). Rethinking bias-variance trade-off for generalization of neural networks. In H. Daumé III & A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning* (pp. 10767-10777). PLMR. https://proceedings.mlr.press/v119/yang20j.html

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*, 1100-1122. https://doi.org/10.1177/1745691617693393