The Beauty Everywhere: How do Aesthetic Criteria Participate in the Development of AI

Abstract

"Beauty" is a highly disputed word in philosophy and art. It also appears frequently in scientific debates. But what is the role of beauty in science, and how can it be useful to AI? In this paper, we argue that scientific progress depends on the diversity of the judgment of scientists, something that is only possible because multiple aspects are involved in the evaluation of theories. Particularly important within these criteria are those related to aesthetic considerations, such as simplicity, consistency, broadness, and fertility. We claim that AI should be less focused on accuracy and related metrics, and instead should aim at integrating epistemic measures related to these aesthetic concepts.

1 Introduction

Contrary to common sense impression, aesthetic evaluations, comprising judgements of "beauty" and related notions, are a regular concern for scientists. The famous mathematician Hermann Weyl, for instance, is quoted as saying: "My work has always tried to unite the true with the beautiful and when I had to choose one or the other, I usually chose the beautiful" [27, p. 278]. In the same vein, computer scientist Donald E. Knuth writes in the preface to his *The art of computer programming*: "I have tried to include all of the known ideas about sequential computer programming that are both beautiful and easy to state" [13, p. viii]. As these two examples intend to illustrate, scientists often take into account the beauty of theories and models, using beauty as a criterion for evaluating the adequacy of theories. But what exactly is responsible for making theories *beautiful*? What aspects of theories give them their aesthetic character? And, especially, does beauty have any epistemic relevance for the development of artificial intelligence (AI)?

2 The Nature of Aesthetic Judgments

There is no universal agreement in philosophy, art, or science, as to which properties of things make them beautiful. Part of the reason for this lack of definition comes from the fact that "beauty" is a polysemous word. As the philosopher Wittgenstein used to say, we find distinct things beautiful and in different ways, such as a beautiful landscape, a contemporary work of art, a person, or a contested soccer game. The notion of "beauty" can be used in multiple registers of language (or, as Wittgenstein called it, "language games") [25].

Putting aside this ambiguity, one could ask whether our assessments of beauty, broadly conceived, are ultimately objective or subjective. In Plato's view, beauty is one of the everlasting forms — non-physical essences that constitute the true reality —, and beautiful things are beautiful by participating of this form [21, 20]. For other philosophers, beauty is a property inherent to objects, given by a certain disposition of its parts, shapes, and colours [26]. Contrary to that view, some philosophers hold a subjective view of beauty — as in the maxim that beauty lies in the eye of the beholder. Beauty is, thus, conceived as a subjective feeling of pleasure, without reference to the properties of the objects that would cause such sensations [24].

Hume [9] and Kant [10] offer a compromise between this two positions. According to them, aesthetic judgments, although not universal, possess an *intersubjective* character. We can expose our judgments of beauty and justify our taste, even though our personal experiences are not the same. Assuming that view, in which aesthetic judgments can be shared and rationally discussed, we can consider what aspects are generally raised in aesthetic contexts. Moreover, among these attributes that make something to be beautiful, we can try to identify some of the aspects of beauty that are important to science and AI.

3 Beauty in Science

Thomas Kuhn famously named five criteria regularly used in the evaluation of theories. According to him, scientists choose theories based on considerations such as precision, simplicity, consistency, scope, and fertility [19].

What kind of criteria are these? Mainly, they are *epistemic* criteria. Although the precise definition of "epistemic" is contested in philosophy, such considerations are epistemic in nature in the sense that they deal with the knowledge provided by theories. Those are criteria are related to things such as: how well a theory is able to explain empirical phenomena, how it improves our understanding or if it offers a true description of the world. This is particularly true as respect to accuracy.

In addition to their epistemic nature, these criteria have a pragmatic character too. Even when they do not epistemically contribute to assess a theory, they may be useful in weighting the practical advantages of adopting a certain theory. For example, all things fixed, we prefer a simpler theory, because it is easier to manipulate. Simplicity, consistency, and scope are all typically used for pragmatic purposes.

Nonetheless, there seems to be another aspect related to these criteria, besides their pragmatic or epistemic nature. These values, in particular simplicity and consistency, are similar to what, in other fields, we would consider as *aesthetic* values.

Let us look at simplicity. We praise works that avoid overwhelming complexity and express only what is essential. Mondrian's paintings and Brancusi's sculptures are paradigmatic of this minimalist approach in art. Simplicity is also valued in scientific matters, as a way to decide between alternative scientific theories. Classical astronomy, for instance, made extensive use of epicycles (circular trajectories whose center is over another circular trajectories) to explain certain motions. Although Copernicus' heliocentric model did not bring immediate gains in accuracy, it mostly dispensed the use of epicycles, providing a conceptually simpler model [15]. This simpler solution to astronomical problems was a main driver of adhesion to the Copernican paradigm.

Even accuracy[1] can be taken as a component of aesthetic judgments. When evaluating a still life drawing or a scientific theory, we pay attention to the exactness of the details and the high degree of technical performance that engender a particular composition. The cult of Da Vinci's *Mona Lisa*, for example, is largely due to the perfection of its strokes.

We can see these multiple criteria regularly present in the context of AI too [1]. Accuracy is undoubtedly the most valued characteristic when evaluating algorithms. The reason is that computer programs are mostly used to perform specific tasks, such as question answering or image classification. Given their task-oriented goal, its natural that algorithms be measured according to how well they perform in these concrete tasks. The attention given to deep learning since the early 2010s provides a neat example of the importance in AI of accuracy considerations. Since their groundbreaking implementations in visual recognition [14], neural networks have been adopted in a large number of areas, given their ability to learn highly complex relations.

Although accuracy is generally seen as the single most valued criterion in AI, research in this field takes into account other aspects of algorithms and models as well. Simplicity is one of them. The preoccupation with elaborating models which are easier to understand, and not just accurate, is a frequent concern for computer scientists. Recently, the non-interpretable character of deep learning algorithms has brought to the fore several strategies to produce transparent or, at least, explainable models [23].

As a third criterion used in AI, which seems more clearly connected to beauty, we can mention considerations of broadness. Many AI models, despite their great performance in specific tasks, are restricted to particular applications or datasets. On the contrary, models such as T5 [22] and the GPT family [2] are valued precisely for their amplitude of scope: they can be used to answer questions, generate dialogues, or summarize texts, without many adjustments. Their capacity of dealing with a large number of tasks is what makes these models so interesting.

As to consistency, we could mention the contextual notion of meaning, which approximates many natural language processing models to investigations in linguistics. The theoretical foundation of this

[1] We take this notion in a broad sense, comprising the family of metrics that measure the success of a theory in predicting or classifying data, including precision, R-squared, f-1 score, among others.

hypothesis is what gives us confidence to develop models in AI that assume contextual definitions of word meaning.

Finally, fertility is also an important criterion in AI; which can be understood as a theory's ability to foster new ideas and discoveries. An example of that would be the rise of attention mechanisms in the last years. Since their first implementations in natural language processing [30], researchers have tried to incorporate attention to deal with other problems occasioned by the large size of inputs in encoder-decoder architectures. Attention has been a fertile idea by giving rise to a variety of tools, such as self-attention, multi-head attention, and a whole family of attention-like concepts. Attentions' ability to inspire new approaches, together with its empirical plausibility (which makes it consistent with what we know about human cognition), is what explains part of its appeal.

4 The Epistemic Significance of Beauty

When Dirac [7] claimed that he would still believe in the relativity theory even if it presented mismatches with evidence, he was implying that precision was not the sole criterion he adopted for assessing theories. Theories may be chosen because they are thought to be beautiful — something that goes beyond the capacity of producing outcomes close to the true values.

But even if scientists consider the beauty of theories when appraising them, what is the epistemic relevance of beauty-related criteria? Ultimately, shouldn't a theory be evaluated only as regards its accuracy? In this section, we argue that the plurality of aesthetic criteria is fundamental to the development of science. Not only is "beauty" used as a criterion employed by scientists in practice (*descriptive thesis*), but it is also a necessary ingredient for science itself (*normative thesis*).

We would like to expose here an argument sketched by Kuhn [18, 17, 19] and developed by Kitcher [11, 12] and D'Agostino [4, 5], which is know as the "Risk-spreading argument". Suppose that a group of rational scientists share the same evidence, and that they have a single goal in sight — accuracy. Given a set of alternative theories, every scientist will produce the same evaluation. If we add the assumption that scientists work with the theory they consider the best, the consequence is that, in this community, every scientist will work with the same theory.

The problem with this situation of absolute consensus is that the best theory at time t may not be the best theory at time $t+1$. A theory that is not very promising at a certain time may be developed by scientists and turn up to be better than its alternatives in the future. This is only possible if there is some level of disagreement among scientists; something that depends on having multiple criteria to evaluate theories.

On the contrary, if all scientists stick with the same theory, for the reason that it is the most effective at a particular time, the community as a whole will not be able to explore alternatives that may prove more fruitful in the long run. An analogy from machine learning can be useful. Conducting a greedy search in order to generate a sequential output, in which we always choose the most likely outcome at a specific point, may result in a sub-optimal choice: sampling the most likely output at each time step does not necessarily lead to the most likely sequence. To avoid this situation, we can conduct a beam search, which explores several paths simultaneously, and finds better sequences overall. Better results can be achieved if more paths are explored.

Science is full of examples that demonstrate the importance of the interaction of several evaluating criteria, including those connected to beauty (which sometimes may generate complicated tradeoffs). This variability is essential to the success of scientific research. Kepler's acceptance of the heliocentric paradigm illustrates this situation well. His belief on the central position of the Sun was largely motivated by his desire to elaborate an astronomical theory consistent with theology. In the end, Kepler's effort to develop the heliocentric theory — using ellipses to describe the orbit of planets — significantly improved the accuracy of astronomical predictions, leading to the full acceptance of this model among astronomers.

In fact, accuracy is often not able by itself to provide a clear decision criterion. This is due to is what is called in philosophy the "underdetermination of theories by evidence" [16]. The underdetermination thesis states that the same evidence may be unable to establish the support of a particular theory: incompatible theories may be equally grounded on empirical evidence. As a consequence of underdetermination, aesthetic judgments, concerning simplicity and consistency for example, may be the only resource of scientists for choosing between theories.

Disregarding accuracy in the short run may be actually fundamental for generating accurate theories in the long run. Scientists need time to connect theoretical postulates to experience, develop adequate instruments of measurement, isolate additional hypotheses, and understand empirical phenomena, before theories can bring precise results. Paying too much attention to a theory's performance from the start may end up killing it before it is able to show its true potential.

In this sense, aesthetic values, in addition to accuracy, can have an important role in the development of models in AI. Beauty can have a specially useful role in the case of negative results. When put to work on specific tasks, models often do not work as expected. If judged only based on their current performance, most of them would be promptly rejected. Employing other criteria may help to diversify scientists' choice, allowing the community to explore multiple paths; some of them perhaps achieving better results in the long run.

More importantly, anomalies [16] — results that conflict with what would be expected to achieve with an accepted theory — can put into question our theoretical presuppositions and highlight the shortcomings of a theory. This tends to be a major factor of innovation in science. As an example from genetics, James Watson and Francis Crick, the discoverers of the DNA structure, changed their theory multiple times, in response to anomalies found in their research, such as its inconsistencies with established observations and the increasing explanatory complexity. It was only after several attempts that they finally elaborated the famous double helix model [31].

There is also a detrimental consequence of the exaggerated focus on accuracy, caused by the inequality of resources in machine learning. When a model's accuracy is measured, the model is not being tested in isolation, but is dependent on the computational resources available, the dataset size, research time, among other things. By taking accuracy as the only relevant criterion, groundbreaking work may be replaced for repetitive work leveraged on abundant resources. In a society where capital is unevenly distributed, an exclusive focus on precision can reward trivial work supported by abundant computational resources; and innovative ideas may remain untested due to the lack of access to adequate resources. That is why other considerations are so relevant: a simple model, with few parameters, tested in a challenging dataset, may bring more insight into a problem than a large model trained on a huge dataset.

Finally, we must consider that accuracy (as in fact any of these criteria) interacts with other types of (primarily) non-scientific values. The almost exclusive value given to a model's performance, for instance, reinforces a certain type of approach, based on data-hunger and large-scale models. A preference for this type of research has theoretical (what kind of research is being conducted) and political consequences (who is favored with the unequal access to resources) [8, 1]. The continuous growth in the size of models has also moral and practical implications as well. The training of large models, for instance, consumes substantial amounts of energy [29]. So, favoring this type of approach means favoring a particular resource allocation in society.

In sum, aesthetic values have a chief epistemological relevance: in attending them, we may produce theories that are better in explaining phenomena and giving us an understanding of the world [28, 6]. To be clear, we are not claiming that *every* aesthetic judgments has an epistemic role. Whereas some assessments of beauty can be related to epistemic aspects (e.g., simplicity), other features may be totally unrelated to the quality of theories (e.g., choosing a theory because of the font in which a paper is written). Not every value is epistemically relevant. But some aesthetic values are, and it is important to scientific development that evaluations not be grounded on a single criterion.

5 A Philosophical Critique of Modern Measures in AI

When we think of evaluating an AI model, accuracy and related metrics are what comes to our mind. More than anything, a theory should predict phenomena well. This is truer in AI than in other areas, given its natural focus on practical applications.

As we have argued, however, other criteria can, and should, also be used in evaluating theories. If this is true, then AI should perhaps be less focused on accuracy and attempt to encompass other forms of evaluation. Not only should we measure the performance of a model, but also how simple, consistent, broad, and fertile it is.

Accuracy and similar metrics have a clear advantage to other criteria, in the sense that they offer straightforward measures of performance. But measures for the other criteria can be considered as

well. Take simplicity, for example. To judge how simple a theory is, we can measure the number of parameters, the interpretability of the features, the number of computations performed, the number of different components in an ensemble model, among other things. Structural Risk Minimization, Akaike Information Criterion, and Bayesian Information Criterion are examples of measures that attempt to provide a trade-off between accuracy and complexity.

As regards to broadness, we could look for measures that take into account the performance of a model in multiple tasks, instead of considering its performance in a single task. For consistency, evaluations could be more theoretically grounded; works that bring evidence for approaches in AI should be specially valuable [3].

Among all criteria, fertility is the one that is less obviously associated with concrete measures. It can serve, nonetheless, as a good indicator of how to direct research efforts. Being able to solve new tasks or solving them in a different way is a strong criterion for preferring a theory. Fertility is also where AI can become more scientifically-minded. It is important not only to solve tasks well, but also to be able to explain things; to obtain understanding of problem situations; to find new associations; and to uncover new and unexpected phenomena.

6 Conclusion

Accuracy receives a special treatment in AI. This is due to the task-oriented nature of this field and the relative ease with which this criterion can be measured. As we intended to show, however, researchers should be open to other considerations as well. In particular, we highlighted the importance of aesthetic judgments, which involve assessments of simplicity, consistency, broadness, and fertility. These criteria have a fundamental epistemic significance: they allow a diversity in evaluations, leading scientists to explore several different theories. By exploring multiple paths at the same time, including theories that have not shown a high accuracy yet, the community reduces the risk of getting stuck with worse theories and may find better theories in the long run. For that reason, a broader assessment of AI models, which includes other measures beyond accuracy, may prove a better approach to this field, compared to the current accuracy-focused paradigm.

References

[1] A. Birhane, P. Kalluri, D. Card, W. Agnew, R. Dotan, and M. Bao. The values encoded in machine learning research. *arXiv preprint arXiv:2106.15590*, 2021.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[3] C. Caucheteux, A. Gramfort, and J.-R. King. Disentangling syntax and semantics in the brain with deep networks. In *International Conference on Machine Learning*, pages 1336–1348. PMLR, 2021.

[4] F. D'Agostino. Kuhn's risk-spreading argument and the organization of scientific communities. *Episteme*, 1(3):201–209, 2005.

[5] F. D'agostino. *Naturalizing Epistemology: Thomas Kuhn and the 'Essential Tension'*. Palgrave Macmillan, 2010.

[6] H. W. De Regt. *Understanding scientific understanding*. Oxford University Press, 2017.

[7] P. Dirac. The excellence of einstein's theory of gravitation. In M. Goldsmith, A. Mackay, and J. Woudhuysen, editors, *Einstein: the First Hundred Years*, pages 41–46. Pergamon, 1980.

[8] R. Dotan and S. Milli. Value-laden disciplinary shifts in machine learning. *arXiv preprint arXiv:1912.01172*, 2019.

[9] D. Hume. *Of the standard of taste*. De Gruyter, 2019.

[10] I. Kant. *Critique of the power of judgment (The Cambridge Edition of the Works of Immanuel Kant)*. Cambridge University Press, 2000.

[11] P. Kitcher. The division of cognitive labor. *Journal of Philosophy*, 87(1):5–22, 1990.

[12] P. Kitcher. *The Advancement of Science*. Oxford University Press, 1993.

[13] D. Knuth. *The art of computer programming*. Addison Wesley, 1997.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[15] T. Kuhn. *The Copernican revolution: Planetary astronomy in the development of western thought*, volume 16. Harvard University Press, 1985.

[16] T. S. Kuhn. *The Structure of Scientific Revolutions. 2. Ed*. University of Chicago Press, 2012, 1962.

[17] T. S. Kuhn. Postscript. In *The Structure of Scientific Revolutions. 2. Ed*. University of Chicago Press, 2012, 1970.

[18] T. S. Kuhn. Reflections on my critics. In *The Road Science Structure*. University of Chicago Press, 2000, 1970.

[19] T. S. Kuhn. Objectivity, value judgment, and theory choice. In *The Essential Tension: Selected Studies in Scientific Tradition and Change*. University of Chicago Press, 1977, 1977.

[20] Plato. Hippias major. In J. M. Cooper and D. S. Hutchinson, editors, *Plato: Complete Works*. Hackett Publishing Co., 1977.

[21] Plato. Symposium. In J. M. Cooper and D. S. Hutchinson, editors, *Plato: Complete Works*. Hackett Publishing Co., 1977.

[22] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[23] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[24] G. Santayana. *The sense of beauty*. Routledge, 2019.

[25] S. Schroeder et al. The emergence of wittgenstein's views on aesthetics in the 1933 lectures. *Estetika: The European Journal of Aesthetics*, 57(1):5–14, 2020.

[26] R. Scruton. *Beauty*. Oxford University Press, 2009.

[27] I. Stewart. *Why Beauty is Truth*. Basic Books, 2007.

[28] M. Strevens. *Depth: An account of scientific explanation*. Harvard University Press, 2011.

[29] E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[31] J. Watson. *The double helix*. Hachette UK, 2012.