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 the 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 judgments of scientists, something that is only possible by the multiplicity of aspects involved in theoretical evaluations. 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 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" [26, 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" [12, p. viii]. As these two examples intend to illustrate, scientists often take into account the beauty of their 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 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") [24].

Despite this ambiguity, one could ask whether our assessments of beauty are ultimately objective or subjective. For Plato, beauty was one of the everlasting forms — non-physical essences that constituted the true reality —, and beautiful things were beautiful by participating of this form [20, 19]. Other philosophers defended that beauty is a property of the objects, given by a certain disposition of the parts, shapes, and colours [25]. Contrary to that view, some philosophers hold a subjective view of beauty — as in the maxim that beauty lies is in the eye of the beholder. Beauty would be reduced to a subjective feeling of pleasure, without reference to the properties of the objects that cause these sensations [23].

Hume [8] and Kant [9] offer a compromise between this two positions. According to them, aesthetic judgments are not universal, but nevertheless posses an *intersubjective* character. Although personal experience varies from person to person, we can discuss about our judgments of beauty and justify our taste. Assuming that view, in which aesthetic judgments can be shared and rationally discussed, even if they are not at the end an objective property of things, 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 "epistemic values", which we can loosely approximate to the most salient aspects of beauty. According to him, scientists would choose theories based on considerations such as precision, simplicity, consistency, scope, and fertility [18].

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 centre 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 [14]. This simpler solution to astronomical problems was a main driver of adhesion to the Copernican paradigm.

Even precision (taken in a broad sense, comprising all metrics related to the predictive power of a model, such as accuracy, R-squared, f-1 score, among others) 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* is due to the perfection of its strokes.

We can see these criteria regularly present in the context of AI too. Precision, and in particular accuracy, is undoubtedly the most desired characteristic in algorithms. 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 [13], neural networks have been adopted in a large number of areas, given their ability to learn highly complex relations.

Although accuracy is always highly regarded, research in AI also takes into account different aspect of algorithms, as their simplicity. Recently, the non-interpretable character of deep learning algorithms has brought to the fore several strategies to produce transparent or, at least, explainable models [22]. Comprehensibility is also of major importance to scientists.

As a third criterion connected to beauty, we can mention considerations of broadness. Many AI models, despite their good accuracy in specific tasks, are restricted to particular applications and, even, to particular datasets. On the contrary, models such as T5 [21] 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 hypothesis in the latter is what gives us confidence to develop models in AI that assume contextual definitions of word meaning.

Finally, fertility is also an important criterion for AI, which can be understood as a theory's capacity of stimulating 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 [28], 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. We can choose theories based on criteria related to "beauty", which, as we saw, involve other things beyond precision.

But, even if scientists consider the beauty of theories when appraising them, what is its epistemic relevance? Ultimately, is not precision all that matters for science? 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 [17, 16, 18] and developed by Kitcher [10, 11] 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 — precision. 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 promising 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 the multiple aesthetic criteria (which sometimes may generate complicated tradeoffs), and the variability they give to scientific research. Kepler's acceptance of the heliocentric paradigm illustrates this situation well. His preference for the heliocentric model was largely motivated by his religious beliefs regarding the divinity of the Sun [1]. His belief on the central position of the Sun was mainly motivated by his desire to elaborate an astronomical theory consistent with theology. Kepler's effort to developed 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, precision 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" — the same evidence may be unable to establish the support of a particular theory. Incompatible theories may be equally grounded on empirical evidence. In this case, aesthetic criteria, such as simplicity and consistency, may be the only resource of scientists for choosing among theories.

Disregarding precision in the short run may be actually positive for precision itself 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 precision, can have an important role in the development of AI programs. 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 [15] — 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 the inconsistencies with established observations and the growing explanatory complexity. It was only after repetitive reformulations that they elaborated the famous double helix model [29].

Finally, there is a detrimental consequence of the exaggerated focus on accuracy, which is connected to the inequality of resources in machine learning. In a society where capital is unequally distributed, an exclusive focus on precision can reward those who do trivial work, but are endowed with abundant computational resources. When we test for precision, the quality of a model is not being tested in isolation, but is dependent on the computational resources available, the dataset size, researching time, among other things. By taking precision as the only relevant criterion, groundbreaking work may be replaced for repetitive work leveraged on abundant resources. That is why other considerations are so relevant: a simple model, with few parameters, tested in a challenging dataset, may reveal more relevant aspects of a problem than a large model trained on a huge dataset.

In sum, aesthetic values may 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 [27, 6]. To be clear, we are not claiming that *every* aesthetic factor 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 they not be reduced to 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 precision 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 criteria 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

Currently, precision is an almost universal criteria in AI. As we intended to show, however, science is generally open to other considerations: scientists consider not only how accurate the predictions of a theory are, but also how beautiful the theory is — its simplicity, consistency, broadness, and fertility. These criteria have a fundamental epistemic significance, in that they allow diversity of judgments, and permit to explore theories that have not shown high degrees of performance yet; some of which may prove better in the long run. For that reason, a broader assessment of AI models, which includes other measures beyond precision, may prove a better approach to this field, compared to the current precision-focused paradigm.

References

[1] P. Barker and B. R. Goldstein. Theological foundations of kepler's astronomy. *Osiris*, 16:88–113, 2001.

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

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

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

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

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

[13] 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.

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

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

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

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

[18] 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.

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

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

[21] 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.

[22] 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.

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

[24] 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.

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

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

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

[28] 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.

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