Bias and Creativity

Róisín Loughran
Regulated Software Research Centre
Dundalk Institute of Technology
Dundalk
Ireland
Roisin.loughran@dkit.ie

Abstract

This paper proposes a discussion on bias and its place in Computational Creativity research. Recent developments in Artificial Intelligence research have become more cognizant of the dangers and pitfalls in not recognising and addressing unseen biases within algorithmic systems. As many such methods are used for creative tasks, we propose that, as a community, we must consider bias possibilities and the implications they could have on the outputs and outcomes of research from this community.

Introduction

Despite many writings, experiments and discussions on the topic, Creativity is still a poorly defined concept. Trying to compute a poorly defined concept is immediately fraught with difficulties. Despite this persistent ambiguity, the field of Computational Creativity (CC) has been examining this problem for many years. Some of the main arsenal for this task have been methods and tools based in Machine Learning (ML) and Artificial Intelligence (AI). Such tools have been shown in recent years to be susceptible to various ethical issues, including, but not limited to detrimental bias. So we ask: Is bias within CC inevitable? And if it is, is this a bad thing?

This may be a complex question, but it is one worth considering. Despite academic and policy approaches to address Bias in AI as detailed in the following section, Big Tech have not always taken such matters as seriously as they should. While Google set up an Ethical AI Team in 2018, the controversial firing of Timnit Gebru in December 2019, and subsequent firing of Margaret Mitchell, who both co-led this Ethical AI Team, for refusing to withdraw a paper that criticised the use of large language models, demonstrates that this is a controversial topic that will not be easily solved (The Irish Times, 2021). Despite public outcry and an open letter of support for Gebru (Medium, 2020), Google did not reverse their decision, nor did they offer support in response to the harassment that subsequentially erupted towards the researchers on social media (Schiffer, 2021).

Bias, fairness and ethics are vitally important considerations for all applications of AI, and CC research is not exempt. In this short discussion paper, we propose a number of areas from which to consider bias in CC. First we consider bias, in its multiple forms and how it has been treated in AI. Then we look at the various types of algorithms that have been typically used in CC and consider if they are all as susceptible as each other. Finally we consider Creativity itself and how it may be rooted in biased decision making.

What is Bias?

As humans we all have inherent biases; when presented with a choice we have tendencies to lean towards one outcome, whether that is based on preference, exposure, belief or something intangible. Biases can be either conscious or subconscious. But, while the notion of *bias* invokes a very negative context, our innate biases are not inherently bad.

When it comes to trying to define bias, there appears to be no one standard clear definition. Dictionary definitions can include reference to prejudices: 'Inclination or prejudice for or against one person or group, especially in a way considered to be unfair.' (Lexico, 2022) or distortion: 'Systematic distortion of results or findings from the true state of affairs, or any of several varieties of processes leading to systematic distortion.' (*A Dictionary of Public Health*, 2007). Yet one of the seminal papers on bias in judgment refers to it as '..decisions based on beliefs concerning the likelihood of uncertain events' (Tversky and Kahneman, 1974). Thus, a bias is simply a decision, one that is informed, either correctly or incorrectly, by some *a priori* belief or understanding we already possess. While the dictionary definitions define bias in terms of unfairness and distortion, the truth is that every day we use heuristics to make sense of the world around us. If we had no biases, our opinions of the world would be akin to white noise.

Detrimental Bias

Biases help us make decisions and form part of our personalities; it is when we encounter discriminatory bias that such judgments can be unfair, illegal or dangerous towards some in our society. As humans, we have inherent biases, and there is a strong potential for us to bring these biases into any algorithmic system we may create or deploy. The potential for algorithms to mirror human biases in decision making has been identified as one of the most straight forward ethical challenges in implementing AI in healthcare (Char, Shah and Magnus, 2018). For this reason there has been much academic research in to the types of biases that may be found or introduced to algorithmic systems in recent years (Mehrabi *et al.*, 2021) along with methods aimed to mitigate these effects (Bellamy *et al.*, 2019).

The problem of detrimental bias within AI systems is also increasingly being identified by regulatory authorities. NIST have recently published a standard on identifying and managing bias in AI (National Institute of Standards and Technology- US Department of Commerce, 2022) and IEEE plan to release the P7003 standard on Algorithmic Bias Considerations later this year (Koene, Dowthwaite and Seth, 2018). Bias, fairness and trustworthiness all contribute to the ethical implementation of AI. Ethics is an even larger consideration than that of bias, and many guidelines have been proposed to ensure ethical implementation of AI such as those proposed by the European Commission on the 'Ethics Guidelines for Trustworthy AI' (European Commission, 2019), although those proposed by the European Commission on are critical of these guidelines (Gille, Jobin and Ienca, 2020).

Fairness

If we remove all discriminatory biases from an algorithmic system we should be able to consider it fair. But, similar to bias, fairness is concept that is colloquially understood but difficult to universally define (Gajane and Pechenizkiy, 2017). Nevertheless many strides have been made to address fairness in AI including the development of fairness, accountability and transparency machine learning (FATML) (Veale and Binns, 2017). This study proposed three methods for addressing this: trusted third parties could be selective with data, online collaborative platforms with diverse organisations could promote fairness and unsupervised learning techniques could allow a fairness hypothesis be built for selective testing. Chen et al. noted that many ML models focus on balancing fairness and accuracy, but they argued that fairness should be evaluated in context of the given data and through data collection and study, rather than through constraint of the model (Chen, Johansson and Sontag, 2018). Binns further considered the nature of fairness and what it means for a ML algorithm to be fair by considering existing works on moral and political philosophy (Binns, 2018). This study questioned should fairness equate to equal opportunity for everyone or focus on minimising harm to the most marginalised. Such studies note that while many approaches to fairness in ML focus on data preparation, model-learning and use of the system, there is still much to be learned about the nature of fairness and discrimination before we can understand how applied ML can address this.

Algorithmic Bias

A variety of ML and AI techniques have been used to emulate Creativity over the years. Is any one more or less prone to bias than the other?

The Data-driven

The explosion of deep learning and in particular Convolutional Neural Nets (CNN) has been largely fueled by the creation of and accessibility to large image datasets. Such methods are commercially very favourable, but due to unbalanced, badly labelled datasets these are some of the most problematic systems in relation to detrimental bias. Birhane et al. discuss several dangers from ill-considered data curation practices including justice, consent and ethical transgressions (Birhane and Prabhu, 2021). Many detrimental biases are found to be discriminatory in relation to sensitive or protected characteristics such as race, gender etc. For this reason these characteristics are often not made available, although simply removing such characteristics from datasets has been shown to exacerbate rather than solve the issue, as latent relationships between other, non-sensitive attributes, can cause proxies that lead to the same biases (Chen *et al.*, 2019). Some methods have been proposed to use these proxies as a way to identify and mitigate against biasing against these characteristics (Lahoti *et al.*, 2020).

When considering bias in an AI system, the data does seem like the primary culprit as the source of bias; a system can only learn and reproduce the data and patterns it is given. But there are more aspects to consider. A recent systematic review found that Data-driven innovation (DDI) suffers from three major sources of bias: data bias, method bias and societal bias (Akter *et al.*, 2021). Thus, even in systems that are driven by the data, we should consider other internal design mechanisms and external influencing factors that can lead to detrimental bias.

The Evolutionists

Evolutionary Computation (EC) comprises a family of heuristic search methods based on Darwin's theory of survival of the fittest. A population of random solutions to a given problem is created and then iteratively improved ('evolved') over a series of generations. This improvement is driven by a fitness function – an evaluation measure of each individual derived by the creator of the algorithm. Such EC methods have been widely used in creative systems such as music, art and design (EvoSTAR, 2022).

EC systems may also work with large datasets, but there are further design decisions within their architectures that could lead to bias. Most notably it is the choice of fitness function that will dictate which individuals are deemed more fit and are hence given a better change of surviving to the next generation. This creates a statistical bias in favour of individuals that conform to the fitness defined. For objective, measurable tasks, this may be what is expected, but for subjective creative tasks, might this be creating an unwanted, or unexpected bias within the system?

Objective Search

Many other systems such as Generative Adversarial Networks (Elgammal *et al.*, 2017) among others have been used in the generation and study of creative artefacts and procedures. While they may differ in their architecture and style, one commonality among AI systems is that they each aim towards a specific objective. That may be to reduce an error, reach a goal or solve a problem, but a system must have an objective to train and aim towards.

The problem with such methods for creative tasks is that the best objective is not always easy to define. How would one pre-define the best melody, sketch or poem? A better search method may be to search for novelty rather than a pre-specified objective. Novelty search proposes that the optimal solution to a problem can be found when looking for a different solution or when looking for no particular solution at all (Lehman and Stanley, 2010). If you are searching for novelty, rather than an objective, it may be less likely that your search will be biased.

Creative Bias

The above may lead us to believe it is the AI, ML and computational tools we use that cause bias within a system. But what of the aesthetic, ever elusive, *Creativity* that we chase? Is Creativity itself susceptible to, or even dependent on, bias?

Like bias and fairness, Creativity is a concept that is understandable by most, yet hard to define in a generalized context. So, in effect, we are trying to ascertain if an ill-defined concept is susceptible to an undefined phenomenon. But, as noted above, we do have an innate understanding of what bias is and how it affects our judgments. In a similar manner we do have ways of measuring Creativity. It has been proposed that to identify Creativity the system must be able to display novelty and value (Boden, 1998).

Novelty At its most absolute meaning, novelty is an unbiased concept; either something is new or it is not. However, often what is meant by novelty is that it is new to the creator. An individual does not have to create something new to the world to have displayed creativity. Personal or P-Creativity is as valid as Historical H-Creativity. In this sense, the novelty of P-Creativity could be biased to the individual.

Value The value of a creative artefact is surely a biased measurement. The monetary value of an aesthetic artefact is measured by what the highest bidder is willing to pay for it. Such a measure is surely influenced by styles, fashion, popularity and a wealth of other immeasurable external biases, along with the internal biases of the buyers. Of course, the monetary value is only one, very superficial, measure of an artefact's worth. A generated piece may have artistic, academic, historical, personal or many other forms of value. But it is likewise difficult to imagine how such a measured value could be determined without any biases.

CC Evaluation

It has been noted numerous times, that evaluation does not take enough precedence in CC experiments (Jordanous, 2012). This is likely due to the complexity of defining what creativity is; how can one measure what you cannot define? Nevertheless, evaluation methods for creativity in computational systems have been proposed. However, many such methods center on human evaluations and judgments which are costly and may lead to limitations (Loughran and O'Neill, 2016). Using human evaluators is costly in both time and money. Furthermore, if we acknowledge that our human biases are subjective to our preferences then we must accept that any human evaluator will evaluate towards their own personal preferences. If someone is adjudicating the creativity of a music generation system, it is difficult to confirm they are judging the system on its creativity and not merely how much they like the melody it produces. When judging creative artefacts, humans tend to mistake what they subjectively 'like' for what it objectively 'good'.

For accurate human based evaluation, you must ascertain their expert knowledge in the given domain. For those judging music, for instance, you should determine how many years of formal music training they have had. Such data may help group certain subjects together, but you must acknowledge this training may not remove a bias but simply introduce new ones. Classically trained musicians may expect, and then favour, outputs of a high musical quality, or music technology students may expect high production value. Even the most experienced adjudicator is still subject to their own learned opinions and biases.

Crowdsourcing

With online resources, it is now quite simple and cost-effective to evaluate on a large cohort of people as Crowdsourcing platforms are increasingly being used for creative tasks (Oppenlaender *et al.*, 2020). However, using large, unregulated crowds to evaluate a creative artefact will surely introduce bias. If you are not sure what demographic your audience is from or what bias profile they have, how can you use their personal preferences as any evaluation of merit? If, instead of paying for a platform, you merely share an online evaluation survey yourself, you are introducing this into a personal circle of people who are, most likely, highly interested or trained in the specific field that you are interested in. In other words, if you created an online survey to evaluate your generated music, how would a random set of people around the globe judge this music in comparison to those on your Twitter feed?

Discussion

As noted earlier, bias is not an intrinsically bad word, or concept; our biases are simply based on heuristics that we need to make decisions. If we consider how we approach the development of a CC system, we must make a number of decisions before we even start development such as:

- The domain(s) within which we will develop and/or test the system;
- The representation used;
- The algorithm(s) employed;
- The validation method(s).

Each of these decision will be influenced by the developers education, experience, personal background and preferences. And many of these choices will require further, more intricate choices along the way – what genre of music will your system compose? What architecture will you use for implementation? Many of these choices are subjective and have no definitive best answer; we do not know the exact number of neurons an ANN must have to make a picture 'creative'. The fact is that we require the freedom to make these choices in order to have the scope to even investigate what it means to be creative. Our learned tendencies, preferences or *biases* may be necessary for us to find creativity in all the mundaneness out there.

In saying that, we know that AI will mimic human behaviors, even the worst of them. Therefore, it is still vitally important that we consider any harmful biases or discriminations that may be emulated by our systems. It is such detrimental biases that we must identify, evaluate and mitigate against.

Detrimental Bias in Creative Systems

We have considered biases in relation to CC in this paper, but where might the most detrimental biases be found in our community?

Demographic As a computer science field, we must acknowledge the lack of women represented in the CC community. Likewise, we must be aware of underrepresentation of other ethnical and minority groups. Such a homogenous demographic is missing out on significant potential contributions to our field. This is not an easy problem to tackle, nor is it unique to CC. However, active and meaningful steps aimed at increasing the diversity within CC research could only benefit the quality and range of our outputs. We would encourage the CC community to actively discuss what measures could be taken to address this.

Training Data Historically, artists have been predominantly male. Hence the training databases, in art, music etc., will have already been curated from a male-generated perspective. If a system is learning from data that has been created predominantly by men, then the female perspective within the training data is missing. It would be difficult to ascertain to what extent this may bias a system, but it is worth consideration. For example, in visual art, there is a strong bias towards the female nude form as opposed to the male form. While acceptable, typical or even encouraged in its day, this is certainly a bias in subject matter. In a similar manner, many training artefacts would be assumed to be biased towards Western style – unless the given study explicitly states otherwise.

Domain CC research can be undertaken in almost any problem domain, as many problems require critical, creative thinking. Despite the fact that much early research in creativity was illustrated using logical tasks, it has been noted that there has been a lack of studies on scientific and logical problems in more recent years (Loughran and O'Neill, 2017). If creativity is not dependent on the application domain, we must acknowledge that an over-representation in one domain over another may introduce a bias within the field in general. The consideration of new application fields may attract new researchers into the field and develop creativity research into new areas.

Complexity Systems that have more complex representation or require and utilise a lot of domain-specific information may appear more impressive and hence be judged to be more creative. We must ensure not to be biased towards more complex systems or become overly impressed by flashy displays.

Bias Types Mehrabi et al. identify 22 types of bias that can be found in ML systems (Mehrabi *et al.*, 2021). While there are many other works discussing types of bias that may be possible within such systems and, arguably, no such list could ever be exhaustive, this is an excellent resource to consider the types of biases your system may be susceptible to. When developing your creative system, it is worth reviewing each bias type to determine if your proposed system may be detrimentally susceptible to these, or other, biases.

Conclusions

As a field within AI, CC researchers should be aware of the possibilities and dangers that bias could pose to their work. This short paper is only intended to start the discussion around biases within CC systems and how we must be vigilant to recognise, acknowledge and, if necessary, mitigate against such biases. We recognise that, as humans, our biases form part of our personalities – our likes and dislikes lead us to make creative choices. We must assume that these biases can, and in some cases should, be passed on to the systems that we develop. These systems generate creative artefacts through the targets, fitness, datasets or benchmarks that we use in their development. We must be aware that the preferences and biases we have learned or inherently own, can be integrated, either consciously, or unconsciously, into our developed systems.

As scientists, we all wish for the most comprehensive, fair and accurate conclusions to our own undertakings. We can only achieve this if we ensure we question the decisions and assumptions we make, at each step of our own processes.

Author Contributions

R.L. ideated and wrote this paper alone.

Acknowledgements

This work has been carried out with the support of Dundalk Institute of Technology.

References

A Dictionary of Public Health (2007) *A Dictionary of Public Health*. Oxford University Press. doi: 10.1093/acref/9780195160901.001.0001.

Akter, S. *et al.* (2021) 'Algorithmic bias in data-driven innovation in the age of AI', *International Journal of Information Management*, 60, p. 102387. doi: 10.1016/J.IJINFOMGT.2021.102387.

Bellamy, R. K. E. *et al.* (2019) 'AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias', *IBM Journal of Research and Development*, 63(4–5). doi: 10.1147/JRD.2019.2942287.

Binns, R. (2018) 'Fairness in Machine Learning: Lessons from Political Philosophy', *Proceedings of Machine Learning Research*, 81, pp. 1–11.

Birhane, A. and Prabhu, V. U. (2021) 'Large image datasets: A pyrrhic win for computer vision?', *Proceedings - 2021 IEEE Winter Conference on Applications of Computer Vision, WACV 2021*, pp. 1536–1546. doi: 10.1109/WACV48630.2021.00158.

Boden, M. A. (1998) 'Artificial Intelligence Creativity and artificial intelligence', *Artificial Intelligence*, 103.

Char, D. S., Shah, N. H. and Magnus, D. (2018) 'Implementing Machine Learning in Health Care — Addressing Ethical Challenges', *New England Journal of Medicine*, 378(11), pp. 981–983. doi: 10.1056/nejmp1714229.

Chen, I. Y., Johansson, F. D. and Sontag, D. (2018) 'Why is my classifier discriminatory?', in *Advances in Neural Information Processing Systems*, pp. 3539–3550.

Chen, J. *et al.* (2019) 'Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved ACM Reference Format'. doi: 10.1145/3287560.3287594.

Elgammal, A. *et al.* (2017) 'CAN: Creative adversarial networks generating "Art" by learning about styles and deviating from style norms', in *Proceedings of the 8th International Conference on Computational Creativity, ICCC 2017*.

European Commission (2019) *Ethics guidelines for trustworthy AI - Publications Office of the EU*. Available at: https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1 (Accessed: 11 April 2022).

EvoSTAR (2022) *EvoMUSART – EvoStar 2022*. Available at: http://www.evostar.org/2022/evomusart/ (Accessed: 12 April 2022).

Gajane, P. and Pechenizkiy, M. (2017) 'On Formalizing Fairness in Prediction with Machine Learning'. doi: 10.1145/3306618.

Gille, F., Jobin, A. and Ienca, M. (2020) 'What we talk about when we talk about trust: Theory of trust for AI in healthcare', *Intelligence-Based Medicine*, 1–2(June), p. 100001. doi: 10.1016/j.ibmed.2020.100001.

Jordanous, A. (2012) 'A Standardised Procedure for Evaluating Creative Systems: Computational Creativity Evaluation Based on What it is to be Creative', *Cognitive Computation 2012 4:3*, 4(3), pp. 246–279. doi: 10.1007/S12559-012-9156-1.

Koene, A., Dowthwaite, L. and Seth, S. (2018) 'IEEE P7003 standard for algorithmic bias considerations', *Proceedings - International Conference on Software Engineering*, (May), pp. 38–41. doi: 10.1145/3194770.3194773.

Lahoti, P. *et al.* (2020) 'Fairness without demographics through adversarially reweighted learning', *Advances in Neural Information Processing Systems*, 2020-Decem.

Lehman, J. and Stanley, K. O. (2010) 'Efficiently evolving programs through the search for novelty', in *Proceedings of the 12th Annual Genetic and Evolutionary Computation Conference, GECCO '10*, pp. 837–844. doi: 10.1145/1830483.1830638.

Lexico (2022) *BIAS | Meaning & Definition for UK English | Lexico.com*. Available at: https://www.lexico.com/definition/bias (Accessed: 31 March 2022).

Loughran, R. and O'Neill, M. (2016) 'Generative Music Evaluation: Why do We Limit to 'Human' ?', in *Conference on Computer Simulation of Musical Creativity*. Huddersfield. Available at: https://www.researchgate.net/publication/304284746 (Accessed: 12 April 2022).

Loughran, R. and O'Neill, M. (2017) 'Application domains considered in computational creativity', in *Proceedings of the 8th International Conference on Computational Creativity, ICCC 2017*.

Medium (2020) 'Standing with Dr. Timnit Gebru - Google Walkout for Real Change', *Medium*, 4 December, pp. 1–4. Available at: https://googlewalkout.medium.com/standing-with-dr-timnit-gebru-isupporttimnit-believeblackwomen-6dadc300d382 (Accessed: 11 April 2022).

Mehrabi, N. *et al.* (2021) 'A Survey on Bias and Fairness in Machine Learning', *ACM Computing Surveys*. Association for Computing Machinery. doi: 10.1145/3457607.

National Institute of Standards and Technology- US Department of Commerce (2022) 'Towards a Standard for Identifying and Managing Bias in Artificial Intelligence', *Natl. Inst. Stand. Technol. Spec. Publ*, 1270, p. 86. doi: 10.6028/NIST.SP.1270.

Oppenlaender, J. *et al.* (2020) 'Creativity on Paid Crowdsourcing Platforms', in *Conference on Human Factors in Computing Systems - Proceedings*. doi: 10.1145/3313831.3376677.

Schiffer, Z. (2021) 'Timnit Gebru was fired from Google — then the harassers arrived', *The Verge*, pp. 1–11. Available at: https://www.theverge.com/22309962/timnit-gebru-google-harassment-campaign-jeff-dean (Accessed: 11 April 2022).

The Irish Times (2021) 'Google Fires Second AI Ethics Leader as Dispute Over Research, Diversity Grows', *18 news*, p. 2. Available at: https://www.irishtimes.com/business/technology/google-fires-second-ai-ethics-leader-as-dispute-over-research-diversity-grows-1.4490768 (Accessed: 11 April 2022).

Tversky, A. and Kahneman, D. (1974) 'Judgment under Uncertainty: Heuristics and Biases', *Science*, 185(4157), pp. 1124–1131. Available at: http://www.jstor.org/stable/1738360.

Veale, M. and Binns, R. (2017) 'Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data', *Big Data and Society*, 4(2), pp. 1–17. doi: 10.1177/2053951717743530.