THE OFFICIAL:
International Journal of Contemporary Humanities,
Vol. 7 December 2023
ISSN 2207-2837

AI ART AND GENDER: A METAMODERNIST EXPLORATION OF BIASES AND POSSIBILITIES

J. Rosenbaum, RMIT University

This artistic exploration delves into the intricate interplay between artificial intelligence (AI), gender representation, and the ethos of metamodernism. Rooted in critical AI discourse and artistic creation, this research seeks to transcend the binary constraints of traditional gender categorisation within AI systems. Drawing inspiration from metamodernist philosophy, which advocates oscillation between opposing concepts, the artworks discussed not only critique AI's inherent biases and limitations but also employ AI as a tool to illuminate societal challenges. This paper comprises a diverse portfolio of interactive artworks, including *Set in Stone*, *Frankenstein's Telephone*, *Gender Tapestry* and *Self_Portrait*, each a unique commentary on AI biases, creatively presented with a touch of irony and humour. Through these creations, viewers are encouraged to engage with the multifaceted issues of AI ethics and gender identity.

Key words: AI, gender, bias, metamodernism

My artwork makes heavy use of different Artificial Intelligence and Deep Neural Network systems. From the earliest works using style transfer and augmented reality to generative adversarial networks to image classifiers and examinations of existing systems and now, to text to image AI or latent diffusion. I work collaboratively with AI. It is not a replacement for me, the artist, but it is a tool I work with to help me create art. Because it learns from the information that I provide and creates its own work based on my input, I am responsible for how it learns to

create. By the same rationale, I also learn from what it creates, and I learn how to produce better datasets and craft better learning parameters. I am frequently surprised by the results, and often, they inform the final work even when they are unexpected. If metamodernism is, by its nature, the oscillation between the modern and the postmodern (Vermeulen and Van Den Akker, 2010), then my AI work can be seen as metamodernist, oscillating between problematising and conceptualising change. It provides meaningful ways to look forward while remaining in conversation with the problematic nature of Artificial Intelligence. I will examine some of my works and explore how they can be seen through a metamodernist lens and in the process explore how we can understand the relationship between postmodernism and metamodernism.

My artwork, *Set in Stone* (Rosenbaum, 2022) explores whether we can affect image generation systems in real time and debias their results. *Set in Stone* is an ongoing artistic project resulting from my inquiries into training bias out of image generation algorithms. I used three discrete 3D rendered datasets, and deliberately manipulated the data during training to see if it could show a shift in bias. I trained the GAN first on masculine marble faces, then added in feminine marble faces, and finally added faces that break out of the binary mould by exploring colour and diverse hairstyles.[TP1] [JR2] The concept is subverting the static binary stone with colour bursting through, a concept of self-expression taking root over the homogeneity of the binary marble statuary. The works resulting from this so far have included a large-scale mosaic wallpaper and a video installation of the disruption training taking place. The wallpaper contains sample images generated during training rendered into a mosaic of a face.

I explored this concept through the medium of Generative Adversarial Networks or GANs. A GAN is a specific method of image creation that uses two neural networks, a generator and a discriminator (Goodfellow et al., 2014). The creator gives the GAN a series of images referred to as a dataset. The generator creates images based on the dataset while the discriminator judges it against the same dataset and passes some while failing others. As images are approved or rejected the generator learns which images satisfy the discriminator and continues to try to improve its results. Eventually it reaches a point called Mode Collapse (Nibali, 2017) where only a few images are being continually generated and passed over and over in a feedback loop as

only a few images are passed each time and the generator has stopped learning from the discriminator. Each cycle of learning is referred to as an epoch. The number of iterations per epoch is determined based on the number of images in the grid and the number of images in the dataset. While training, the GAN puts out sample images to show its progress. These images from the information held deep in the process of learning are referred to as occurring in the latent space of training (Nguyen et al., 2016).

The development of the work is the area I find most intriguing in training a GAN, not the finished result, but rather what occurs in between. It evolves and develops the images like a Polaroid photograph. There is a similar sense of wonder and interest in that development process, the evolution of the machine's understanding of its creation and the dataset. Through the samples, we can gain an appreciation for how the machine learns and grows and that process is highly interesting to me. I wanted to see how images can be manipulated during training in that latent space by manipulating the dataset of images. Through the data manipulation in this project, I am literally displacing the masculine images and the biassed training of the neural network to make room for further identities. I am creating gender trouble (Butler, 2006), intentionally, through this piece. Problematizing gender bias in AI while showing a potential way forward to debias systems and how it looks while the bias is trained out of it and the system learns new data.

The introduction of new data showed not only the signature aesthetics of the neural network and a debiasing of the results but created a dialogue about how we are the sum of our experiences, not just what you see on the surface. Through this work I learned that debiasing a neural network like a GAN with new data is possible. The work is arranged as a mosaic with a video of the training process. The images within the mosaic are all samples from training and show different stages of learning and bias. As you look at the different faces, some appear to be more resolved than others. As new data was added during training, some of the samples that were generated directly after the addition of new data were strangely disrupted, especially when colour was introduced. This has allowed for some interesting artefacts that clearly represent, to me, the core of machine learning art, the attempt to understand and create. The artefacts generated along the way show the learning and the disruption of knowledge. These are a core element to the aesthetic

of machine learning art and while many AI artists attempt to remove these artefacts, the uniqueness of the art may be removed in the process.

This project taught me a lot about data and a lot of technical issues relating to bias, but it also taught me a lot about how the strength of the imagery affects the results and how the journey of the development is so much of this work. This work challenges the assumption that AI is impartial, revealing how it absorbs and responds to new data. If this were a postmodernist work it would mostly problematize the issues (Baciu et al., 2016), but I choose to view this work through a metamodernist lens, a work that shows that change is possible and that we can work through issues with concerted effort. Showing the existing problems of gender and AI gender bias, with a visual language built on nostalgia and historical marble works (Brunton, 2018), showing joy and colour breaking through, optimism, breaking through the stone and revealing more potential than just being set in stone.

Frankenstein's Telephone (Rosenbaum, 2022) is an AI version of the game 'telephone' where a message is passed from one player to the next by whispering in a neighbour's ear, who then passes on their version of what they heard until the message travels around in a circle. The resulting message is often humorously garbled.[SH3] [JR4] In '*Frankenstein's Telephone*', I am interested in the shifts pertaining to gender and personhood. The sequence of the different AI systems I used reveals aspects of how AI classification works and, potentially, where bias occurs, and which parts of the chain are weakest. The red colour in the segmentation assigned in DeepLab (Chen et al., 2016) is ungendered. It simply stands for "person." I believe that the simplicity of that classification category is what we should aspire to with AI classifications. As demonstrated in these works, however, some of these open-source AI networks have been designed to ascribe gender. The discomfiting images of computer-generated humanity calls to mind the story of Frankenstein's monster (Shelley 1891), a narrative that many queer people, especially transgender people, identify with. The real monsters are the people behind the scenes that subconsciously perpetuate bias. The datasets used in these works are created by humans and are labelled by humans. The captions use gendered language written by humans and interpreted by machines. This is where the bias appears, where the mistakes and misgendering occur. These

machine-generated conceptions of humanity, with their unexpected forms and eyes and mouths could be seen as modern monsters, but no less deserving of their label ‘person’.

I used the first text to image generator ATTNGAN (Xu et al., 2017) and connected it to classification and repainting systems to see how different AI systems interpret the work of each other. By passing it through multiple stages I can see how the neural networks observe gender and create assumptions around gender. This image is first generated by ATTNGAN from a prompt supplied by the user. It is then segmented in DeepLab (Chen et al., 2016) removing all concepts of gender from the image. DeepLab labels the segmented person as “person”. Therefore, in the next steps when it is repainted, the interpretation of the new image by the caption generator was the key detail. After segmentation a new image is generated by SPADE-COCO (Park et al., 2019) filling each segmented zone with content echoing the label. The final stage is the generated caption from Im2Text (Ordonez et al., 2011).

This is an ongoing issue with massive datasets scraped from the internet and crowd sourced data collection. This facelessness, the lack of agency for the people in the photographs, and the lack of identity for people behind the scenes doing this tedious work, is rendered into numbers. I observed some clear gendering of the people in the MSCOCO (Lin et al., 2014) dataset images and captions. As I read more about how mass workers are used to caption dataset images, and the instructions they are given and the conditions they work under, I gained an understanding of the reasons behind the gendering of these barely human appearing images. Background data plays a large part in the gender of the person (Wang et al., 2018), as does the action they are supposedly doing. The biases of the workers and the level of detail in the instructions and oversight also play a role, often perpetuating bias. This work problematizes the large corporate datasets that we are becoming increasingly reliant upon and the use of underpaid mass worker groups (Aguinis et al., 2021), but presents the results in a humorous game of linked AI systems that highlight the absurdity of our reliance on such systems. This oscillation speaks to the core of metamodernism. Pulling away from the fanaticism of AI and swinging towards irony (Vermeulen and Van Den Akker, 2010), highlighting the issues through absurdism.

We often say gender is a spectrum without considering that it is more than a linear construct. *Gender Tapestry* (Rosenbaum, 2022) is a gender classification artwork where instead of a gender, the user receives a custom mixed colour, the results of a custom, multi-label image classifier. I began this project to explore not only the history of facial recognition and gender classification, but also colour perception. Like different computer monitors, Most humans experience colour in different ways, some differences are huge while some are very small, but we all see colour uniquely (Emery and Webster, 2019). The theory that I explored through this work was to consider that gender perception and experience is much the same as colour perception, that everyone has a slightly different experience of gender because it is informed by our life experiences and culture and our own personal perception of our gender.

I wanted to create something that equated colour with gender, so that instead of receiving a gender classification, people would receive their own unique colour mix. *Gender Tapestry* (Rosenbaum, 2022) is a work that explores gender as a 3D colour space. The result was an evolving mosaic using the colour results of the users and a GAN trained on their faces. The mosaic gained in complexity as more people interacted with the system and received their custom gender colours. I particularly enjoy working with multiple AI systems in a complex network of multi modal interaction. Throughout my research I have explored gender classification and its fallibility and lack of inclusiveness. I may have started *Gender Tapestry* with the aim of creating a better way to look at gender classification, but in the process, I think I have explored a better, metamodernist, way to look at gender. *Gender Tapestry* is a community work, it works best the more people interact with it and help it learn and grow. It problematizes gender classifiers but educates through the medium of classification and colour. colour and gender have long been associated together, but to take it beyond the binary, into a full spectrum of colour possibilities is to make gender as a social construct more tangible. The classifier was trained using pronouns and largely artificial faces in a multi label system (Fürnkranz et al., 2008) that acknowledges that people use multiple sets of pronouns. Pronouns are a key element to the understanding of gender, understanding through semantic relationships (Lauscher et al., 2022), and understanding through exploration and discovery of our own identities (Pullen Sansfaçon et al., 2020). A pronoun is a personal label, a way of announcing our identity to the world, it is also

a way to learn about and understand gender and explore our personal relationships with it. Further to this, rather than sanitising the images that went into the classifier, as is typical in most gender classifiers (Golomb et al., 1991), I queered the results by leaning in to the performativity of gender (Butler, 1988). Therefore, while each image will always receive the same colour, different ways of expressing yourself and your gender will receive different results. This work explores a wholesome different way to perceive gender, beyond the confines of a binary, to an optimistic future of acceptance. An opportunity to believe in gender as more than a biology, more than a binary, and to believe that we can get past limited AI perceptions of gender. This belief, this sense of hope and optimism characterises metamodernism over postmodernism (Koford, 2022).

Continuing my explorations of gender and AI, *Self_Portrait* explores gender through different connected AI systems that examines how certain words, like personality terms, can have gendered connotations. It uses a multi-label classifier (Fürnkranz et al., 2008) trained on certain word groups and assigns those to the user, it then passes a prompt using those assigned words to Stable Diffusion (Rombach et al., 2022) using ControlNets (Zhang et al., 2023) to mimic the viewer's position. The AI draws its impression of what a person with those personality terms looks like. However, a truly metamodernist work would not simply accept what an AI classifies a person as, so there is a second stage to the work where the user can choose to accept the assigned terms or change them to better suit themselves. A new portrait is generated and then transplanted onto the viewer's face using augmented reality.

All of this must use AI. There is no way that users will get original works every time without using a system like Stable Diffusion. The AI is intrinsically tied to this artwork, just as gender is, because it is an examination of the assumptions AI makes about us. Different AI systems make snap judgments of us every day, from our browsing habits on Facebook to our style of writing on twitter (Verhoeven et al., 2016). Every action becomes a part of profiling users individually on each of these sites (Flekova and Gurevych, 2013). As it is algorithmically determined, our profile may not always be what we choose. Because of profiling and the biases that get reinforced, more insidious systems may end up causing issues, like job searching and applications (Keinert-Kisin,

2016), the justice system (Richardson et al., 2019) and financial systems (Johnson et al., 2019). *Self_Portrait* can't show the breadth of AI access and overreach, but it can show how we take the labels different AI systems assign us for granted. I consulted for the classifier with ChatGPT (OpenAI, 2021) to produce a list of words associated with genders and generated batches of faces using those terms in Stable Diffusion to help fill in the dataset. Because this work is about what AI perceives with specific words, and by using systems like Stable Diffusion and ChatGPT we can have a direct dialogue with the generation system and see its assumptions and biases writ large which makes it uniquely suitable for interrogating itself and AI. Increasingly, I use AI in multiple stages of development, from ideation and planning to dataset and programming support. This isn't because I believe AI is faultless, but rather because I believe that interrogating AI with AI is the best way to understand the limitations of AI. AI magnifies biases, so what better way to hold a mirror up to those issues and biases than with AI?

Self_Portrait is about labels, and about accepting, or not, the labels that are given us by external forces. I think that, in the past, labels were something that happened to us, that people hung on us. They ostracised us for our labels. Now labels are something we claim for ourselves, they help us find communities and connections. They can still be sources for ostracization, but now they are most often chosen and accepted, embraced. This work is about not just accepting the labels thrust upon us by an unknowing and uncaring machine but choosing our labels for ourselves. I enjoyed watching people use the system and laugh as their labels came up and then take great care choosing their own. That care and interest in what these words mean for us is what this work is truly about. Rather than the acceptance of a bleak postmodernist society where we accept the assignations of AI systems, this is a metamodernist look at how we can work with AI to accept ourselves. This work was created in concert with Melanie Huang for a site-specific work for the Science Gallery Melbourne. The collaboration that built this work from a small concept to a cohesive artwork experienced by 20,000 people is part of the spirit of Metamodernism, working together, using open source technology, creating a large communal mood board together and working towards a metanarrative (Stoev, 2022) that informed the final work.

In exploring the intersections of AI, gender, and perception, my work embodies the spirit of metamodernism. Metamodernism calls for an oscillation between opposing concepts, acknowledging the limitations and biases of AI systems while still using them to shed light on societal issues. Just as metamodernism seeks to transcend the binary nature of modernism and postmodernism, I aim to challenge the binary categorisation of gender within AI systems. My work oscillates between castigating AI for its biases and shortcomings and using AI to highlight and explore those shortcomings in a way that allows viewers to engage with the subject. Balancing that oscillation to ensure that neither the critique nor the engagement is lost is a key part of my artistic approach, just as I must balance the scientific and the artistic.

Metamodernism encourages a reevaluation of grand narratives and a departure from cynicism towards a more sincere and optimistic perspective. In my work, I strive to challenge the existing gender classification systems and question why AI needs to know a person's gender in the first place. By redirecting the focus towards pronouns, a more inclusive and respectful approach is possible. My aim with my work is not only to highlight the possibilities of AI but also why we must approach it with caution. Exploring the biases while showcasing an optimistic view towards creating work that engages with the problems of AI as it works with AI systems itself. My work *Set in Stone* highlights an optimistic view of AI bias through the artistic debiasing of a biassed neural network, showcasing a metamodernist optimism beyond the postmodernist problematization. Through my works *Frankenstein's Telephone* and *Self_Portrait*, I explore the biases inherent in existing AI systems using irony to highlight the issues and challenges, promoting interactive engagement with these serious problems in a lighthearted way. *Gender Tapestry* provides an optimistic way to explore gender classification and further our understanding of gender. Allowing engagement between people and AI to create a collaborative dialogue that fosters understanding of the limitations of gender categorisation, and a joyous use of colour and community engagement.

My multidisciplinary use of AI systems and collaboration with other artists and experts reflect the metamodernist idea of embracing complexity and hybridity. Through the integration of various AI techniques, such as style transfer, GANs, and latent diffusion, I create multi-modal

interactive experiences that engage participants in a dialogue about the assumptions and biases embedded in AI.

By exploring the potential of AI and questioning its limitations, my work embodies the metamodernist spirit of critical engagement, sincere exploration, and a constant reevaluation of existing systems. I believe that through this lens, we can strive for a more nuanced and inclusive understanding of gender and challenge the assumptions made by AI systems.

Gender remains a critical issue in AI systems, from the categorisation and gendering of people to the biases and the continued perception of gender as a binary rather than as a spectrum. Most gender classification systems work on a binary and don't allow for the myriad differences of humanity, and are therefore incorrect much of the time, but, through my research I have largely ended up with the question, why? Why do these systems need to know people's genders at all? Why is so much time and energy taken up with trying to classify something that could be deduced with a simple question? Instead of trying to use AI to determine your gender, all that really needs to be asked is 'what are your pronouns'?

REFERENCES

Aguinis, H., Villamor, I., Ramani, R.S., 2021. MTurk Research: Review and Recommendations. J. Manag. 47, 823–837. https://doi.org/10.1177/0149206320969787

Baciu, C., Opre, D., Riley, S., 2016. A New Way of Thinking in the Era of Virtual Reality and Artificial Intelligence. https://doi.org/10.13140/RG.2.1.3986.6483

Brunton, J., 2018. Whose (Meta)modernism?: Metamodernism, Race, and the Politics of Failure. J. Mod. Lit. 41, 60. https://doi.org/10.2979/jmodelite.41.3.05

Butler, J., 2006. Gender Trouble : Feminism and the Subversion of Identity. Taylor & Francis Group, Florence, UNITED KINGDOM.

Butler, J., 1988. Performative Acts and Gender Constitution An Essay in Phenomenology and Feminist Theory. Theatre J. 40, 519. https://doi.org/10.2307/3207893

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2016. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs.

Emery, K.J., Webster, M.A., 2019. Individual differences and their implications for colour perception. Curr. Opin. Behav. Sci. 30, 28–33. https://doi.org/10.1016/j.cobeha.2019.05.002

Flekova, L., Gurevych, I., 2013. Can We Hide in the Web? Large Scale Simultaneous Age and Gender Author Profiling in Social Media. PAN CLEF.

Fürnkranz, J., Hüllermeier, Eyke, Loza Mencía, Eneldo, Brinker, Klaus, Fawcett Fürnkranz, T.J., Loza Mencía, E, Hüllermeier, E, Brinker, K, 2008. Multilabel classification via calibrated label ranking. Mach Learn 73, 133–153. https://doi.org/10.1007/s10994-008-5064-8

Golomb, B.A., Lawrence, D.T., Sejnowski, T.J., 1991. Sexnet: A neural network identifies sex from human faces. Adv. Neural Inf. Process. Syst. 3 572–7.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative Adversarial Networks.

Johnson, K., Pasquale, F., Chapman, J., 2019. Artificial Intelligence, Machine Learning, and Bias in Finance: Toward Responsible Innovation. FORDHAM LAW Rev. 88.

Keinert-Kisin, C. author, 2016. Corporate social responsibility and discrimination : gender bias in personnel selection. Springer International Publishing : Imprint: Springer, Cham.

Koford, K.L., 2022. A Hug for Humanity: Metamodernism and Masculinity on Television in Ted Lasso.

Lauscher, A., Crowley, A., Hovy, D., 2022. Welcome to the Modern World of Pronouns: Identity-Inclusive Natural Language Processing beyond Gender.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P., 2014. Microsoft COCO: Common Objects in Context. ArXiv E-Prints.

Nguyen, A., Clune, J., Bengio, Y., Dosovitskiy, A., Yosinski, J., 2016. Plug & play generative networks: Conditional iterative generation of images in latent space, Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/CVPR.2017.374

Nibali, A., 2017. Mode collapse in GANs [WWW Document]. AidenNibali.org. URL http://aiden.nibali.org/blog/2017-01-18-mode-collapse-gans/
OpenAI, 2021. ChatGPT.

Ordonez, V., Kulkarni, G., Berg, T.L., 2011. Im2Text: Describing images using 1 million captioned photographs. Adv. Neural Inf. Process. Syst. 24 25th Annu. Conf. Neural Inf. Process. Syst. 2011 NIPS 2011 1–9.

Park, T., Liu, M.-Y., Wang, T.-C., Zhu, J.-Y., 2019. Semantic Image Synthesis with Spatially-Adaptive Normalization. arXiv:1903.07291.

Pullen Sansfaçon, A., Medico, D., Suerich-Gulick, F., Temple Newhook, J., 2020. "I knew that I wasn't cis, I knew that, but I didn't know exactly": Gender identity development, expression and affirmation in youth who access gender affirming medical care. Int. J. Transgender Health 21, 307–320. https://doi.org/10.1080/26895269.2020.1756551

Richardson, R., Schultz, J.M., Crawford, K., 2019. DIRTY DATA, BAD PREDICTIONS: HOW CIVIL RIGHTS VIOLATIONS IMPACT POLICE DATA, PREDICTIVE POLICING SYSTEMS, AND JUSTICE. N. Y. Univ. Law Rev. 94.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-Resolution Image Synthesis with Latent Diffusion Models.
Rosenbaum, J., 2022. AI perceptions of gender.

Stoev, D., 2022. Metamodernism or Metamodernity. Arts 11, 91. https://doi.org/10.3390/arts11050091

Verhoeven, B., Daelemans, W., Plank, B., 2016. TwiSty: A multilingual Twitter stylometry corpus for gender and personality profiling. Proc. 10th Int. Conf. Lang. Resour. Eval. LREC 2016 1632–1637.

Vermeulen, T., Van Den Akker, R., 2010. Notes on metamodernism. J. Aesthet. Cult. 2, 5677. https://doi.org/10.3402/jac.v2i0.5677

Wang, T., Zhao, J., Yatskar, M., Chang, K.-W., Ordonez, V., 2018. Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations 1.

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X., 2017. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. arXiv:1711.10485.

Zhang, L., Rao, A., Agrawala, M., 2023. Adding Conditional Control to Text-to-Image Diffusion Models.

Do you analyse the aesthetic choices of marble - the connotations of classical greco-roman form that you seek to subvert - or anything like that? You could allude to classical, and neo-classical 'white face' marble bodies and trace the leaning-toward-patriarchal/fascistic/authoritative tendencies of that aesthetic as something to be subverted [TP1]

[JR2]Ah! I forgot that this was something I explored in another publication and then never actually wrote up in here!

would it be worth noting the racial connotations of this (Chinese Whispers) and how this aligns to misunderstanding and otherness....? Might be out of scope? [SH3]

[JR4]I keep going back and forth on that one TBH!