There is no evidence that Meaning Maps capture semantic information relevant to gaze guidance: reply to Henderson, Hayes, Peacock, and Rehrig (2021)

Marek A. Pedziwiatr[1,2], Matthias Kümmerer[3], Thomas S.A. Wallis[4], Matthias Bethge[3], Christoph Teufel[1]

[1]*Cardiff University, Cardiff University Brain Research Imaging Centre (CUBRIC), School of Psychology, Cardiff, United Kingdom*

[2] *Queen Mary University of London, Department of Biological and Experimental Psychology, London, United Kingdom*

[3]*University of Tübingen, Tübingen, Germany*

[4] *Technical University Darmstadt, Institute for Psychology and Centre for Cognitive Science, Darmstadt, Germany*

Abstract

The concerns raised by Henderson, Hayes, Peacock, and Rehrig (2021) are based on misconceptions of our work. We show that Meaning Maps (MMs) do not predict gaze guidance better than a state-of-the-art saliency model that is based on semantically-neutral, high-level features. We argue that there is therefore no evidence to date that MMs index anything beyond these features. Furthermore, we show that although alterations in meaning cause changes in gaze guidance, MMs fail to capture these alterations. We agree that semantic information is important in the guidance of eye-movements, but the contribution of MMs for understanding its role remains elusive.

We welcome the opportunity to clarify the rationale, results, and conclusions of our paper on Meaning Maps (MMs; Pedziwiatr et al., 2021) in response to the points raised by Henderson, Hayes, Peacock, and Rehrig (henceforth, HHPR; [reference to the commentary to be inserted]). HHPR's core criticism of our paper is based on three misconceptions. They argue that (i) we are "*denying the influence of semantic content*" on eye-movements, that (ii) we claim that "*because Meaning Maps do not capture object-scene semantic consistency, they do not capture any aspects of semantic content*", and that (iii) we argue that because MMs do not outperform DGII in predicting human gaze the two "*must reduce to the same type of non-semantic underlying representation*". While we concede that the title of our paper could have been more nuanced, we made none of these claims. We agree with HHPR that semantic information is important in guiding eye-movements – in fact, our paper (Pedziwiatr et al., 2021) corroborates previous studies demonstrating this importance. A method that quantifies the spatial distribution of semantic information in images would therefore be a useful research tool. However, our findings suggest that it is unclear whether MMs as currently formulated can serve this purpose. Here, we will summarise the rationale of the MMs approach, describe how the logic of our own study directly builds on this rationale, and finally, detail key conclusions that can be derived from our findings.

The paper that introduced the concept of MMs (Henderson & Hayes, 2017) contrasted two sets of predictions regarding where people look in images: one derived from MMs and the other from a saliency model called GBVS (Harel et al., 2006). The logic of this approach is simple: to the extent that one predictor outperforms the other, the winning predictor's image features and/or computational mechanisms better capture the factors that guide eye movements. MMs are generated by using crowdsourced ratings of the 'meaningfulness' of image patches, and then spatially smoothing the ratings to create a map over the whole image. Henderson and Hayes showed that MMs that are created in this way outperform GBVS in predicting human fixation locations, and that they explain more unique variance in the eye-movements data. Based on the assumption that MMs measure semantic information, they concluded that "*both previous and current results are consistent with a theory in which meaning is the dominant force guiding attention through scenes*" (Henderson & Hayes, 2017).

The first part of our study used the same logic but extended it to multiple saliency models. This approach was motivated by the finding that low-level features, on which classic saliency models (such as GBVS) rely, provide a poor explanation for gaze guidance in free viewing of natural scenes (Kümmerer et al., 2015; Kümmerer, Wallis, Gatys, Bethge, et al., 2017; Kümmerer et al., 2020). It is therefore important to benchmark new methods against a range of saliency models, including state-of-the-art models. We replicated Henderson and Hayes' (2017) key finding: MMs outperform GBVS in predicting human eye-movements, and explain more unique variance. However, MMs did not consistently outperform other models that also use exclusively low-level features (AWS and ICF; Garcia-Diaz et al., 2012; Kümmerer, Wallis, Gatys, & Bethge, 2017). Moreover, DeepGaze II (DGII; Kümmerer et al., 2016; Kümmerer, Wallis, Gatys, & Bethge, 2017), a modern saliency model based on high-level features, generated better predictions – a finding our paper replicates in two separate data sets – and explained more unique variance than MMs. Based on the reasoning outlined above, these results would imply that the image features and/or computational mechanisms underpinning DGII's predictions provide better explanations for the guidance of eye movements in free-viewing of natural (non-contrived) scenes than those measured by MMs. Strong evidence supporting the usefulness of MMs in understanding oculomotor control, and of their utility for gaze prediction over and above alternative features or modelling frameworks, would require MMs to outperform these models and, ideally, explain more unique variance.

Our findings are directly relevant to an evaluation of MMs as a tool to measure semantic information. Predictions by DGII are based on an image-computable, high-level feature space (Kümmerer et al., 2016; Kümmerer, Wallis, Gatys, & Bethge, 2017). We argue that these features can be carriers of meaning but, in and of themselves, do not amount to meaning. HHPR have an even stronger interpretation of DGII's semantic emptiness, stating that it is not clear whether "*deep learning models like DG2 can ever in principle capture object-scene semantic features, or indeed any type of semantic feature*". Based on (i) the assumption that MMs measure the distribution of semantic information and (ii) the logic of the original MMs study, the result that DGII outperforms MMs, and explains more unique variance, would therefore lead to the conclusion that (semantically-neutral) high-level features rather than 'meaning' guide eye-movements. Note that this conclusion applies to any type of meaning that MMs might measure, including the concept of "*context-free semantic density for local scene regions*". Critically however, due to the findings of the second part of our study, we do not subscribe to this view. Rather, we question whether MMs index unique semantic information relevant for gaze guidance over and above semantically-neutral, high-level features.

In the second part of our study, we sought to determine how MMs (and DGII) predict fixations when meaning is dissociated from the presence of complex visual features. Specifically, we assessed the extent to which MMs (and DGII) capture semantic information related to object-scene (in)consistencies. It is widely acknowledged in the literature that this type of meaning is important for eye-movements (Williams & Castelhano, 2019; Wu et al., 2014). In line with previous work (Henderson et al., 1999; Loftus & Mackworth, 1978), we found that people fixate more on objects that are semantically inconsistent with the scene than those that are consistent. This shows that semantic information changed gaze guidance. However, neither DGII nor MMs indexed this change.

HHPR argue that MMs as originally proposed were never intended to be able to measure meaning associated with object-context relationships. This intention was unclear to us from the original paper, given that HHPR only define the limits of "*context-free meaning*" in later papers (Henderson et al., 2018; Henderson & Hayes, 2018) and their current commentary. Incidentally, because the coarse and fine patches seen by raters are fixed in size, the extent to which they are actually 'context free' depends on the size of objects in the image. It may be possible for a rater to recognize a (semantically inconsistent) shoe on a bathroom sink, if those objects are small enough in the image. In any case, our study provides empirical support for HHPR's claim that 'context-free' MMs do not index the semantic information contained in object-scene relationships. In our target paper, we explicitly acknowledge the possibility that MMs might capture other types of semantic information but highlight that their insensitivity to meaning related to object-scene relationships limits their usefulness. Moreover, this limitation may be difficult to fix: data from a (forthcoming) follow-up study suggest that "*contextualised MMs*" (Peacock et al., 2019) also fail to capture semantic information linked to object-scene relationships, despite the fact that they have been designed to be sensitive to this type of meaning.

Our study shows that MMs provide a worse explanation for oculomotor control than a saliency model that is based on semantically-neutral features, and that MMs fail to capture changes in gaze guidance in response to experimental manipulations of meaning. Taken together, these findings led us to favour the explanation that MMs do not measure unique semantic information that is relevant for gaze guidance over and above semantically-neutral, high-level features. What, then, do Meaning Maps measure? To construct MMs, raters are asked to "*assess the meaningfulness of each patch based on how informative or recognizable they thought it was*" (Henderson & Hayes, 2017). The ambiguity of key terms in these instructions allows raters to make up their own minds about how to approach the task. It seems entirely plausible that raters base their assessment on high-level image features similar to those used by DGII. We disagree with HHPR that this interpretation "*requires assuming that raters ignore the instructions they are given*", since those features may often be "*informative*" as to the presence of "*recognizable*" objects.

In summary, we argue that high-level features can often be the carriers of meaning but, in and of themselves, do not amount to meaning. We see no empirical evidence to suggest that MMs (and cMMs) index anything more than these high-level features (though that does not mean they *must*). When such features are experimentally dissociated from one specific but important type of meaning, current MMs (and contextualised MMs) do not capture changes in meaning relevant for human eye-movements. While our work highlights limitations of the current MMs approach, we hope that this debate will contribute to the development of a tool to index the distribution of meaning across an image, either within the MMs approach, or beyond.

References:

Garcia-Diaz, A., Fdez-Vidal, X. R., Pardo, X. M., & Dosil, R. (2012). Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing*, *30*(1), 51–64. https://doi.org/10.1016/j.imavis.2011.11.007

Harel, J., Koch, C., & Perona, P. (2006). Graph-Based Visual Saliency. *Advances in Neural Information Processing Systems.* https://doi.org/10.1.1.70.2254

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour*, *1*(October). https://doi.org/10.1038/s41562-017-0208-0

Henderson, J. M., & Hayes, T. R. (2018). Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps. *Journal of Vision*, *18*(6), 10. https://doi.org/10.1167/18.6.10

Henderson, J. M., Hayes, T. R., Rehrig, G., & Ferreira, F. (2018). Meaning Guides Attention during Real-World Scene Description. *Scientific Reports*, *8*(1), 13504. https://doi.org/10.1038/s41598-018-31894-5

Henderson, J. M., Weeks, P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(1), 210–228. https://doi.org/10.1037/0096-1523.25.1.210

Kümmerer, M., Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., & Torrabla, A. (2020). *MIT/Tübingen Saliency Benchmark*. https://saliency.tuebingen.ai/

Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2015). Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, *112*(52), 16054–16059. https://doi.org/10.1073/pnas.1510393112

Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2016). *DeepGaze II: Reading fixations from deep features trained on object recognition*. 1–16. http://arxiv.org/abs/1610.01563

Kümmerer, M., Wallis, T. S. A., Gatys, L. A., & Bethge, M. (2017). Understanding Low- and High-Level Contributions to Fixation Prediction. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-Octob*, 4799–4808. https://doi.org/10.1109/ICCV.2017.513

Kümmerer, M., Wallis, T. S. A., Gatys, L. A., Bethge, M., Kummerer, M., Wallis, T. S. A., Gatys, L. A., Bethge, M., Kümmerer, M., Wallis, T. S. A., Gatys, L. A., Bethge, M., Kummerer, M., Wallis, T. S. A., Gatys, L. A., Bethge, M., Kümmerer, M., Wallis, T. S. A., Gatys, L. A., & Bethge, M. (2017). Understanding Low- and High-Level Contributions to Fixation Prediction. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-Octob*, 4799–4808. https://doi.org/10.1109/ICCV.2017.513

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(4), 565–572. https://doi.org/10.1037//0096-1523.4.4.565

Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2019). The role of meaning in attentional guidance during free viewing of real-world scenes. *Acta Psychologica*, *198*(December 2018), 102889. https://doi.org/10.1016/j.actpsy.2019.102889

Pedziwiatr, M. A., Kümmerer, M., Wallis, T. S. A., Bethge, M., & Teufel, C. (2021). Meaning maps and saliency models based on deep convolutional neural networks are insensitive to image meaning when predicting human fixations. *Cognition*, *206*(10), 104465. https://doi.org/10.1016/j.cognition.2020.104465

Williams, C. C., & Castelhano, M. S. (2019). The changing landscape: High-level influences on eye movement guidance in scenes. *Vision (Switzerland)*, *3*(3), 1–20. https://doi.org/10.3390/vision3030033

Wu, C. C., Wick, F. A., & Pomplun, M. (2014). Guidance of visual attention by semantic information in real-world scenes. In *Frontiers in Psychology* (Vol. 5, Issue FEB, pp. 1–13). https://doi.org/10.3389/fpsyg.2014.00054