

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:<https://orca.cardiff.ac.uk/id/eprint/140913/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Pedziwiatr, Marek A., Kümmerer, Matthias, Wallis, Thomas S.A., Bethge, Matthias and Teufel, Christoph 2021. There is no evidence that meaning maps capture semantic information relevant to gaze guidance: Reply to Henderson, Hayes, Peacock, and Rehrig (2021). *Cognition* 214 , 104741. 10.1016/j.cognition.2021.104741

Publishers page: <http://dx.doi.org/10.1016/j.cognition.2021.104741>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



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 Kümmerer³, Thomas S.A. Wallis⁴, Matthias Bethge³, Christoph Teufel¹

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

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

³University of Tübingen, Tübingen, Germany

⁴Technical 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.

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

39

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

52

53 The first part of our study used the same logic but extended it to multiple saliency models. This
54 approach was motivated by the finding that low-level features, on which classic saliency models
55 (such as GBVS) rely, provide a poor explanation for gaze guidance in free viewing of natural
56 scenes (Kümmerer et al., 2015; Kümmerer, Wallis, Gatys, Bethge, et al., 2017; Kümmerer et al.,
57 2020). It is therefore important to benchmark new methods against a range of saliency models,
58 including state-of-the-art models. We replicated Henderson and Hayes' (2017) key finding: MMs
59 outperform GBVS in predicting human eye-movements, and explain more unique variance.
60 However, MMs did not consistently outperform other models that also use exclusively low-level
61 features (AWS and ICF; Garcia-Diaz et al., 2012; Kümmerer, Wallis, Gatys, & Bethge, 2017).
62 Moreover, DeepGaze II (DGII; Kümmerer et al., 2016; Kümmerer, Wallis, Gatys, & Bethge, 2017),
63 a modern saliency model based on high-level features, generated better predictions – a finding
64 our paper replicates in two separate data sets – and explained more unique variance than MMs.
65 Based on the reasoning outlined above, these results would imply that the image features and/or
66 computational mechanisms underpinning DGII's predictions provide better explanations for the
67 guidance of eye movements in free-viewing of natural (non-contrived) scenes than those
68 measured by MMs. Strong evidence supporting the usefulness of MMs in understanding
69 oculomotor control, and of their utility for gaze prediction over and above alternative features
70 or modelling frameworks, would require MMs to outperform these models and, ideally, explain
71 more unique variance.

72

73 Our findings are directly relevant to an evaluation of MMs as a tool to measure semantic
74 information. Predictions by DGII are based on an image-computable, high-level feature space
75 (Kümmerer et al., 2016; Kümmerer, Wallis, Gatys, & Bethge, 2017). We argue that these features
76 can be carriers of meaning but, in and of themselves, do not amount to meaning. HHPR have an
77 even stronger interpretation of DGII's semantic emptiness, stating that it is not clear whether
78 *“deep learning models like DG2 can ever in principle capture object-scene semantic features, or
79 indeed any type of semantic feature”*. Based on (i) the assumption that MMs measure the
80 distribution of semantic information and (ii) the logic of the original MMs study, the result that
81 DGII outperforms MMs, and explains more unique variance, would therefore lead to the
82 conclusion that (semantically-neutral) high-level features rather than ‘meaning’ guide eye-
83 movements. Note that this conclusion applies to any type of meaning that MMs might measure,
84 including the concept of *“context-free semantic density for local scene regions”*. Critically

85 however, due to the findings of the second part of our study, we do not subscribe to this view.
86 Rather, we question whether MMs index unique semantic information relevant for gaze
87 guidance over and above semantically-neutral, high-level features.

88

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

98

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

115

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

129

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

138

139 References:

140

- 141 Garcia-Diaz, A., Fdez-Vidal, X. R., Pardo, X. M., & Dosil, R. (2012). Saliency from hierarchical
142 adaptation through decorrelation and variance normalization. *Image and Vision Computing*,
143 30(1), 51–64. <https://doi.org/10.1016/j.imavis.2011.11.007>
- 144 Harel, J., Koch, C., & Perona, P. (2006). Graph-Based Visual Saliency. *Advances in Neural*
145 *Information Processing Systems*. <https://doi.org/10.1.1.70.2254>

- 146 Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as
147 revealed by meaning maps. *Nature Human Behaviour*, 1(October).
148 <https://doi.org/10.1038/s41562-017-0208-0>
- 149 Henderson, J. M., & Hayes, T. R. (2018). Meaning guides attention in real-world scene images:
150 Evidence from eye movements and meaning maps. *Journal of Vision*, 18(6), 10.
151 <https://doi.org/10.1167/18.6.10>
- 152 Henderson, J. M., Hayes, T. R., Rehrig, G., & Ferreira, F. (2018). Meaning Guides Attention
153 during Real-World Scene Description. *Scientific Reports*, 8(1), 13504.
154 <https://doi.org/10.1038/s41598-018-31894-5>
- 155 Henderson, J. M., Weeks, P. A., & Hollingworth, A. (1999). The effects of semantic consistency
156 on eye movements during complex scene viewing. *Journal of Experimental Psychology:*
157 *Human Perception and Performance*, 25(1), 210–228. [https://doi.org/10.1037/0096-](https://doi.org/10.1037/0096-1523.25.1.210)
158 1523.25.1.210
- 159 Kümmerer, M., Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., & Torralba, A. (2020).
160 MIT/Tübingen Saliency Benchmark. <https://saliency.tuebingen.ai/>
- 161 Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2015). Information-theoretic model comparison
162 unifies saliency metrics. *Proceedings of the National Academy of Sciences*, 112(52), 16054–
163 16059. <https://doi.org/10.1073/pnas.1510393112>
- 164 Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2016). DeepGaze II: Reading fixations from deep
165 features trained on object recognition. 1–16. <http://arxiv.org/abs/1610.01563>
- 166 Kümmerer, M., Wallis, T. S. A., Gatys, L. A., & Bethge, M. (2017). Understanding Low- and High-
167 Level Contributions to Fixation Prediction. *Proceedings of the IEEE International Conference*
168 *on Computer Vision, 2017-Octob*, 4799–4808. <https://doi.org/10.1109/ICCV.2017.513>
- 169 Kümmerer, M., Wallis, T. S. A., Gatys, L. A., Bethge, M., Kummerer, M., Wallis, T. S. A., Gatys, L.
170 A., Bethge, M., Kümmerer, M., Wallis, T. S. A., Gatys, L. A., Bethge, M., Kummerer, M.,
171 Wallis, T. S. A., Gatys, L. A., Bethge, M., Kümmerer, M., Wallis, T. S. A., Gatys, L. A., &
172 Bethge, M. (2017). Understanding Low- and High-Level Contributions to Fixation
173 Prediction. *Proceedings of the IEEE International Conference on Computer Vision, 2017-Octob*,
174 4799–4808. <https://doi.org/10.1109/ICCV.2017.513>
- 175 Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during
176 picture viewing. *Journal of Experimental Psychology: Human Perception and Performance*,
177 4(4), 565–572. <https://doi.org/10.1037/0096-1523.4.4.565>
- 178 Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2019). The role of meaning in attentional
179 guidance during free viewing of real-world scenes. *Acta Psychologica*, 198(December 2018),
180 102889. <https://doi.org/10.1016/j.actpsy.2019.102889>
- 181 Pedziwiatr, M. A., Kümmerer, M., Wallis, T. S. A., Bethge, M., & Teufel, C. (2021). Meaning maps
182 and saliency models based on deep convolutional neural networks are insensitive to image

183 meaning when predicting human fixations. *Cognition*, 206(10), 104465.
184 <https://doi.org/10.1016/j.cognition.2020.104465>

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

188 Wu, C. C., Wick, F. A., & Pomplun, M. (2014). Guidance of visual attention by semantic
189 information in real-world scenes. In *Frontiers in Psychology* (Vol. 5, Issue FEB, pp. 1–13).
190 <https://doi.org/10.3389/fpsyg.2014.00054>
191