

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

TESLA, INC.,

Petitioner,

v.

PERCEPTIVE AUTOMATA LLC,

Patent Owner.

Case No. IPR2025-01575

Patent No. 11,753,046

DECLARATION OF DR. ELI SABER, PH.D.

TABLE OF CONTENTS

I.	INTRODUCTION	1
II.	BACKGROUND AND QUALIFICATIONS	2
III.	MATERIALS CONSIDERED AND TIME OF INVENTION	8
IV.	LEGAL STANDARDS	11
A.	Overview	11
B.	Claims.....	11
C.	Burden of Proof for Invalidity.....	13
D.	Claim Construction	13
E.	Person of Ordinary Skill in the Art	14
F.	Prior Art.....	14
G.	Anticipation	15
H.	Obviousness.....	16
V.	PERSON OF ORDINARY SKILL IN THE ART	19
VI.	THE '046 PATENT AND THE CHALLENGED CLAIMS	19
A.	Overview of the '046 Patent.....	19
B.	Claim Overview	21
VII.	CLAIM CONSTRUCTION	23
VIII.	GROUND 1	26
A.	Cox	26
B.	Cox Does Not Teach “user”/“users”	31
C.	Cox Does Not Teach Elements 1[b-iii], 8[b-iii], or 15[b-iii] (“generating summary statistics describing the state of mind of the	

	user in the image based on the received responses from the plurality of human observers”).....	32
D.	Cox Does Not Teach Elements 1[b-iv], 8[b-iv], or 15[b-iv] (“storing the summary statistics in association with the image as part of the training data”)	47
E.	Cox Does Not Teach Elements 1[c], 8[c], or 15[c] (“training a model using the training data, the model configured to receive an input image showing a user and predict summary statistics describing a state of mind of the user in the input image”)	50
F.	Both the Petition and Dr. Janét’s Declaration Fail to Analyze the Factors Required for a Proper Obviousness Analysis.....	53
IX.	Ground 2	54
A.	Cox and Ross.....	54
B.	Cox and Ross Do Not Teach Elements 1[b-iii], 8[b-iii], or 15[b-iii] (“generating summary statistics describing the state of mind of the user in the image based on the received responses from the plurality of human observers”)	56
C.	Cox and Ross Do Not Teach Elements 1[b-iv], 8[b-iv], or 15[b-iv] (“storing the summary statistics in association with the image as part of the training data”)	57
D.	Cox and Ross Do Not Teach Elements 1[c], 8[c], or 15[c] (“the model configured to receive an input image showing a user and predict summary statistics describing a state of mind of the user in the input image”).....	58
E.	A POSITA Would Not Have Been Motivated to Combine Cox and Ross in the Manner Tesla and Dr. Janét Allege	61
X.	GROUND 3	68
XI.	GROUND 4	69
A.	Cox and Munro.....	69

B.	Cox and Munro Do Not Teach Elements 1[b-iii], 8[b-iii], or 15[b-iii] (“generating summary statistics describing the state of mind of the user in the image based on the received responses from the plurality of human observers”).....	71
C.	Cox and Munro Do Not Teach Elements 1[b-iv], 8[b-iv], or 15[b-iv] (“storing the summary statistics in association with the image as part of the training data”)	78
D.	Cox and Munro Do Not Teach Elements 1[c], 8[c], or 15[c] (“the model configured to receive an input image showing a user and predict summary statistics describing a state of mind of the user in the input image”).....	80
E.	A POSITA Would Not Have Been Motivated to Combine Cox and Munro in the Manner Tesla and Dr. Janét Allege	83
XII.	GROUND 5	83
A.	Ellenbogen and Munro	83
B.	Ellenbogen and Munro Do Not Teach Elements 1[b-iii], 8[b-iii], or 15[b-iii] (“generating summary statistics describing the state of mind of the user in the image based on the received responses from the plurality of human observers”).....	85
C.	Ellenbogen and Munro Do Not Teach Elements 1[b-iv], 8[b-iv], or 15[b-iv] (“storing the summary statistics in association with the image as part of the training data”)	87
D.	Ellenbogen and Munro Do Not Teach Elements 1[c], 8[c], or 15[c] (“training a model using the training data, the model configured to receive an input image showing a user and predict summary statistics describing a state of mind of the user in the input image”)	89
E.	A POSITA Would Not Have Been Motivated to Combine Ellenbogen and Munro in the Manner Tesla and Dr. Janét Allege.....	92
XIII.	CONCLUSION.....	93

I, Eli Saber, Ph.D., hereby declare as follows:

I. INTRODUCTION

1. I have been retained as an independent expert consultant on behalf of Perceptive Automata, LLC (“Patent Owner” or “Perceptive”) in this *Inter Partes* Review proceeding (“IPR”) before the Patent Trial and Appeal Board (“Board” or “PTAB”) of the United States Patent and Trademark Office (“USPTO”) regarding U.S. Patent No. 11,753,046 (the “’046 patent”) (Ex. 1001).

2. I am being compensated for the time I spend on this matter, in addition to expenses. My compensation in this case is in no way contingent on the nature of my findings, the presentation of my findings in testimony, or the outcome of this or any other proceeding. I have no other interest in this proceeding.

3. I understand that Tesla, Inc. (“Petitioner” or “Tesla”) has challenged the patentability of claims 1-19 of the ’046 patent in this proceeding based on certain alleged prior art references identified and discussed in Tesla’s Petition (Paper 1).

4. I have been asked to consider, and provide my opinions regarding, whether the alleged prior art references, as presented in Grounds 1-5 of the Petition, render obvious claims 1-19 of the ’046 patent. In my opinion, the alleged prior art references, as presented in Grounds 1-5 of the Petition, do not render obvious any of the challenged claims.

5. This declaration is based on the information currently available to me. To the extent that additional information becomes available, I reserve the right to continue my investigation and study, which may include review of documents and information that may be produced, as well as testimony from depositions that have not yet taken place. My opinions are set forth below.

II. BACKGROUND AND QUALIFICATIONS

6. I have summarized in this section my educational background, career history, publications, and other relevant qualifications. I understand my full *curriculum vitae* is submitted as Exhibit 2040.

7. I am a Professor in the Department of Electrical and Microelectronic Engineering (EME) at the Chester F. Carlson Center for Imaging Science at the Rochester Institute of Technology (RIT) in Rochester, New York. I also serve as the Director of the Image, Video and Computer Vision Laboratory. I joined RIT's full-time faculty in 2004. I have over thirty-four (34) years of industry and academic experience, thirty-two (32) of which are in the fields of Computer Vision, Digital Image/Video Processing, Pattern Recognition and Machine Learning.

8. I received a Ph.D. in Electrical Engineering from the University of Rochester in 1996. My concentration was in Signal/Image/Video Processing, Pattern Recognition, and Computer Vision. I received a Master of Science Degree in Electrical Engineering from the University of Rochester in 1992 in the same

concentration as listed above, and I received a Bachelor of Science Degree in Electrical and Computer Engineering, Summa Cum Laude, from the State University of New York at Buffalo in 1988.

9. Before becoming a full-time professor, I worked for Xerox Corporation from 1988 to 2004. During my 16 years at Xerox, I was – generally speaking – responsible for: a) delivering color management solutions, image processing innovations, architectures, and algorithms for high-end color printing systems, b) xerographic sub-systems for a variety of color products, and c) control systems for toner production facilities. Among other roles, as an Advanced Development Scientist and Manager, I established the Advanced Design Laboratory – an Imaging/Xerographics lab – and provided technical and managerial leadership for the Electrical, Imaging and Xerographics Department. As a Product Development Scientist and Manager, I led the research and development of color management algorithms and image quality metrics for various product platforms. I also led the Image Science, Analysis, and Evaluation area, with twelve to fifteen direct reports and a budget of approximately \$2 million.

10. From 1997 until 2004, I was an adjunct faculty member at the Electrical Engineering Department of the Rochester Institute of Technology and at the Electrical & Computer Engineering Department at the University of Rochester. I was responsible for teaching undergraduate and graduate coursework in signal,

image, video processing, pattern recognition, and communications and performing research in multimedia applications, pattern recognition, image understanding and color engineering.

11. Since joining RIT full time, I have been responsible for teaching several undergraduate and graduate courses in Digital Signal Processing, Digital Image Processing, Digital Video Processing, Advanced Neural Networks and Deep Learning, Pattern Recognition, Engineering Analysis, Random Signal & Noise, Advanced Engineering Mathematics, Matrix Methods, Pattern Recognition, Communications, Modern Control Theory, and Linear Systems.

12. I proposed, developed and introduced the “Digital Video Processing” and the “Advanced Neural Network and Deep Learning: Part 1” courses; and reestablished the “Pattern Recognition” course at the Electrical Engineering Department at RIT. These graduate courses were instrumental in providing MS/PhD students with the proper foundation and knowledge in all aspects of the subject areas in order to perform in-depth research in the field and continue to advance the state of the art.

13. The Digital Video Processing course is designed to provide a thorough and comprehensive coverage of the field. Topics covered include the mathematical representation of digital video including motion and time-varying image formation models, spatio-temporal sampling/reconstruction, sampling structure conversions,

motion estimation using optical flow, block-based, pel-recursive, Bayesian and point correspondence type methodologies, segmentation, motion tracking, filtering, restoration, super-resolution and compression.

14. The Pattern Recognition course covers pertinent concepts in Bayes decision theory, linear and nonlinear classifiers, artificial neural networks, feature selection/generation/extraction and various types of clustering algorithms and techniques.

15. The Advanced Neural Networks and Deep Learning: Theory, Mathematics and Algorithms course is intended to provide an in-depth comprehensive study of the theory, mathematics and algorithms for advanced neural networks and their usage in deep learning. It is primarily targeted for upper graduate students that are planning to pursue advanced studies/research in the field. The course begins with an in-depth study of the mathematical methodologies for neural networks and optimization (zero-, first- and second-order) techniques. It then progresses to provide a thorough study of shallow and deep neural networks architectures and algorithms, activation functions, regression, two- and multi-class classification, unsupervised learning, feature engineering/selection, backpropagation along with applications to a variety of engineering problems.

16. I am currently working to develop a follow-on course (Part 2) to the Advanced Neural Networks and Deep Learning course described above. This course

is intended to cover convolutional and recurrent style neural networks, principles of feature engineering, kernel methods, tree-based learners, attention mechanisms/transformers, reinforcement learning and generative adversarial networks. This course is targeted to debut in Fall 2027.

17. I am also currently working to revamp the graduate course titled “Image and Video Compression”. The updated version of this course will start by providing a brief history of video coding and video coding standards and then progress to discussing video basics, formats and fundamentals of video compression. It will then progress to cover JPEG, JPEG 2000, H.261, MPEG-1, MPEG-2, H.263, MPEG-4 visual, H.264 and H.265 and if time permits, AV-1 and H.266. The course is intended to debut in Fall 2028.

18. My current research focuses on machine learning for object tracking/remote sensing applications; image/video processing for multimedia, military & biomedical applications; and computer vision and three-dimensional scene reconstruction. As a principal investigator (PI) or co-principal investigator, I have acquired almost \$6 million in research funding since joining RIT and have managed multiple government grants from the Department of Defense as well as several corporate grants from Hewlett-Packard, Lenel, and Data Physics. I am currently managing (as PI) a multiyear government grant from the Department of Defense titled “Target Detection/Tracking and Activity Recognition from

Multimodal Data.” In this project, we propose to develop operational target detection/tracking techniques and activity recognition/understanding methods that can leverage multimodal data in a fusion framework using deep learning algorithms while providing accurate and near real-time performance.

19. In 2012, I was awarded the Prestigious Trustees Scholarship, the highest award at RIT with regard to research recognition.

20. I am the author or co-author of forty (40) peer-reviewed journal publications.

21. I have also authored or co-authored 103 conference and workshop publications and a book entitled Advanced Linear Algebra for Engineers with MATLAB, published by CRC Press in February 2009, and I am a named inventor on multiple U.S. and foreign patents.

22. I am a senior member of the Institute of Electrical and Electronic Engineers and a member of the IEEE Signal Processing Society, the Electrical Engineering Honor Society, and Eta Kappa Nu.

23. With respect to the technologies relevant to the '046 Patent, I have extensive experience in developing computer vision and digital image/video processing algorithms and techniques for motion estimation, segmentation, registration, classification, identification, pattern recognition, shape recognition/matching, stereo imaging, facial recognition, biomedical image

processing, compression, surveillance systems, 3d scene reconstruction, machine learning, deep learning for a variety of applications including object recognition and tracking, video analysis, video surveillance, change detection, biomedical, color engineering, grayscale and color document processing.

24. A complete statement of my industrial and academic records including a listing of the above publications and patents is included with my *curriculum vitae*.

III. MATERIALS CONSIDERED AND TIME OF INVENTION

25. The opinions contained in this declaration are based on the documents that I reviewed, my professional judgment, as well as my education, experience, and knowledge regarding technologies related to the '046 patent.

26. In forming my opinions expressed in this declaration, I considered the following materials:

<u>Document</u>	<u>Exhibit/Paper #</u>
'046 patent	1001
File history for the '046 patent	1002
Declaration of Jason Janét	1003
U.S. Patent Application Publication No. 2017/0099200 ("Ellenbogen")	1004
U.S. Patent No. 10,496,091 ("Ross")	1005

U.S. Patent Application Publication No. 2016/0162456 (“Munro”)	1006
International Publication WO 2014/210334 (“Cox”)	1007
U.S. Patent No. 10,733,506 (“Ogale”)	1012
Petition	1
Patent Owner’s Preliminary Response	9
Lord, F.M. (1951), A THEORY OF TEST SCORES AND THEIR RELATION TO THE TRAIT MEASURED. ETS Research Bulletin Series, 1951: i-126. https://doi.org/10.1002/j.2333- 8504.1951.tb00922.x	2041
Kooij JFP, Schneider N, Flohr F, Gavrilă DM (2014) Context-based pedestrian path prediction. In: Proceedings of the European Conference on Computer Vision (ECCV). Berlin: Springer, pp. 618–633.	2042

A. Zyner, S. Worrall, J. Ward and E. Nebot, "Long short term memory for driver intent prediction," 2017 IEEE Intelligent Vehicles Symposium (IV), Los Angeles, CA, USA, 2017, pp. 1484-1489, doi: 10.1109/IVS.2017.7995919	2043
C. Aggarwal, "Neural Networks and Deep Learning", Springer ISBN 978-3-319-94462-3	2044

27. I have been asked to consider the time of the inventions of the challenged claims to be around December 4, 2017, which I understand is the filing date of the earliest non-provisional patent application to which the '046 patent claims priority. Ex. 1001, Cover. My opinions reflect how a person of ordinary skill in the art (as I define in Section V below) would have understood the claims and the specification of the '046 patent, the prior art to the patent, and the state of the art, at the time of the alleged invention.

28. I understand that the '046 patent also claims priority to U.S. Provisional Patent Application No. 62/528,771, which was filed on July 5, 2017. My opinions herein do not differ whether the time of the inventions of the challenged claims is around July 5, 2017 or December 4, 2017. In addition, for the sake of clarity, I have

been asked to assume that all of the references relied upon by Tesla in support of its IPR petition are prior art to the challenged claims of the '046 patent.

29. As I discuss in detail below, it is my opinion that claims 1-19 of the '046 patent are not rendered obvious by the alleged prior art references, as presented in Grounds 1-5 of the Petition.

IV. LEGAL STANDARDS

A. Overview

30. I am not an attorney or a legal expert. The legal standards I applied in reaching the conclusions expressed throughout this declaration were provided to me by counsel. In preparing this declaration, I have endeavored to faithfully apply these legal standards. I do not intend to offer opinions on the law. Below, I describe the law I applied in considering whether the claims of the '046 patent are patentable.

31. I have been informed that an issued patent is presumed valid and that the challenger of the patent's validity bears the burden of proving invalidity of the patent.

B. Claims

32. I have been informed that patent claims describe the inventions made by the inventors and describe what the patent owner owns and what the owner may prevent others from doing.

33. I have been informed that claims may describe products, such as machines or chemical compounds, or processes for making or using a product. I have been informed that claims are usually divided into parts or steps called limitations.

34. I have been informed that there are two types of claims: independent claims and dependent claims. I have been informed that an independent claim sets forth all the requirements that must be met in order to be covered by that claim. I have been informed that it is not necessary to look at any other claim to determine what an independent claim covers. I have been informed that a dependent claim does not itself recite all of the requirements of the claim, but refers to another claim for some of its requirements. I have been informed that, in this way, the claim depends on another claim. I have been informed that the law considers a dependent claim to incorporate all of the requirements of the claims to which it refers.

35. I have been informed that the dependent claim then adds its own additional requirements. I have been informed that to determine what a dependent claim covers, it is necessary to look at both the dependent claim and any other claims to which it refers.

36. I have been informed that the dependent claims incorporate the limitations of the independent claim upon which they depend and, if a reference fails to disclose any of the limitations of an independent claim, the claims which depend on that claim cannot be made invalid by that reference. I have been informed that,

as a result, any of the references at issue in this proceeding that fail to disclose all of the limitations of an independent claim necessarily also fail to disclose all of the limitations of claims which depend on that claim.

C. Burden of Proof for Invalidity

37. I understand that the party challenging the validity of an issued U.S. patent in an IPR bears the burden of proving invalidity through facts supported by a preponderance of the evidence. I understand that a “preponderance of the evidence” means evidence sufficient to prove that a fact is more likely true than it is not.

D. Claim Construction

38. I have been informed that the first step in determining validity is to properly construe the claims. I have been informed that, in this proceeding, the Board must determine the scope of the claims by using the same claim construction standard that United States district courts use. I have been informed that, under this standard, words of a claim are given their ordinary and customary meaning as understood by a person of skill in the art at the time of the invention. I have been informed that such a person looks to the claims, specification, and file history in understanding the meaning of the claim terms. I have applied this understanding in this declaration.

E. Person of Ordinary Skill in the Art

39. I have been informed that a person having ordinary skill in the art may be defined by factors such as: (a) the levels of education and experience of the inventor and other persons actively involved in the field; (b) the types of problems encountered in the field; (c) prior art solutions to those problems; (d) rapidity with which innovations are made; and (e) the sophistication of the technology.

F. Prior Art

40. I have been informed that a patent cannot be invalidated based on one person's testimony alone without corroborating evidence, particularly documentary evidence. Moreover, I have been informed that corroboration is required of any witness whose testimony alone is asserted to invalidate a patent, regardless of his or her level of interest in relation to that patent. I have been informed that, accordingly, even if the testifying individual has no interest in the litigation, corroboration is still required.

41. In this declaration, I may use the phrase "prior art" when describing specific references. However, my use of this phrase should not be interpreted as agreement that the item(s) described qualify as prior art under the patent laws. I merely use this phrase to address the items labeled as "prior art" by Tesla and its expert. As I mention above, I have been asked to assume that the references Tesla

relies upon in support of this IPR petition are prior art to the challenged claims of the '046 patent.

G. Anticipation

42. I have been informed that a patent claim can be found invalid as anticipated if each and every element of the claim, as properly construed, is found either explicitly or inherently in a single prior art reference. I have been informed that, if the single prior art reference does not disclose any one claim limitation, then that claim is not anticipated. I have been informed that the single prior art reference cannot merely disclose each element; rather, it must disclose all of the elements as arranged in the claim. I have been informed that for a prior art reference to “inherently” disclose something, the reference must necessarily function in accordance with, or include the claimed limitations, to anticipate. I have been informed that the missing descriptive matter cannot merely probably or possibly include the claimed limitations. I have been informed that, with respect to an anticipation analysis, it is improper to combine separate, distinct embodiments of a single prior art reference.

43. I have been informed that for a claim to be anticipated, each claim limitation must be disclosed in a single prior art reference, and the claimed arrangement or combination of those limitations must also be disclosed in that same prior art reference. I have been informed that for a reference to be anticipatory, the

reference must contain an enabling disclosure. I have been informed that merely naming or describing the subject matter is insufficient if it cannot be produced without undue experimentation.

H. Obviousness

44. I have been informed that a patent claim can be found invalid as obvious if the differences between the claim and the prior art references would have been obvious to a person of ordinary skill in the art at the time of the invention, taking into account (1) the scope and content of the prior art, (2) the differences between the prior art and the claims, (3) the level of ordinary skill in the art, and (4) any objective indicia of non-obviousness. I further understand that a claim can be found to be obvious in light of either a single prior art reference or multiple prior art references. I understand that references qualify as prior art for an obviousness determination only when they are analogous to the claimed invention. I understand that two separate tests define the scope of analogous prior art: (1) whether the art is from the same field of endeavor, regardless of the problem addressed; and (2) if the reference is not within the field of the inventor's endeavor, whether the reference still is reasonably pertinent to the particular problem with which the inventor is involved.

45. I understand that to be obvious in light of a single prior art reference or multiple prior art references, there must be a reason either to modify either the single

prior art reference or to combine the two or more prior art references to achieve the claimed invention. I understand that this reason may come from a teaching, suggestion, or motivation to combine or may come from the reference or references themselves, the knowledge or “common sense” of one skilled in the art, or from the nature of the problem to be solved, and may be explicit or implicit from the prior art as a whole.

46. The combination of familiar elements according to known methods is likely to be obvious when it does no more than yield predictable results.

47. I further understand that there is no suggestion to combine, however, if a reference teaches away from its combination with another source, and that a reference may be said to teach away when a person of ordinary skill, upon reading the reference, would be discouraged from following the path set out in the reference, or would be led in a direction divergent from the path that was taken by the applicant, or if it suggests that the line of development flowing from the reference’s disclosure is unlikely to be productive of the result sought by the applicant. If when combined, the references would produce a seemingly inoperative device, then they teach away from their combination. I understand that it is improper to combine references where the references teach away from their combination.

48. I understand that the degree of teaching away will depend on the particular facts; in general, a reference will teach away if it suggests that the line of

development flowing from the reference's disclosure is unlikely to be productive of the result sought by the applicant. I understand that a reference teaches away, for example, if (1) the combination would produce a seemingly inoperative device, or (2) the references leave the impression that the product would not have the property sought by the applicant.

49. I understand that if the teachings of a prior art reference would lead one skilled in the art to make a modification that would render another prior art device inoperable, then such a modification would generally not be obvious. I understand that if a proposed modification would render the prior art device unsatisfactory for its intended purpose, then the proposed modification would generally not be obvious.

50. I understand that it is improper to use hindsight in making the obviousness determination.

51. I understand that some articulated reasoning with some rational underpinning to support the legal conclusion of obviousness must be made. For example, I understand that there may be no rational underpinning to support the legal conclusion of obviousness where systems are already complete and the party challenging validity seeks to mix and match components, which each exist in the separate systems, from these complete systems.

V. PERSON OF ORDINARY SKILL IN THE ART

52. I understand that Tesla's expert (Dr. Janét) has opined that:

A person having ordinary skill in the art ("POSITA") relevant to the '046 Patent as of July 5, 2017 would have had at least: (1) a bachelor's degree in electrical engineering, computer engineering, computer science, or equivalent course work, with three years of work experience in computer vision, autonomous vehicles, and/or machine learning; or (2) a master's degree in electrical engineering, computer engineering, computer science, or equivalent course work, with a focus in computer vision, autonomous vehicles, and/or machine learning.

Ex. 1003, ¶ 63.

53. For purposes of providing my opinions in connection with this declaration, I do not dispute Dr. Janét's declaration, and I have applied his definition in rendering my opinions herein. I believe that my education and experience qualify me as a person of ordinary skill in the art relevant to the '046 Patent as of July 5, 2017 under either prong of Dr. Janét's definition. *See* Section II above and my CV.

VI. THE '046 PATENT AND THE CHALLENGED CLAIMS

A. Overview of the '046 Patent

54. The inventions disclosed and claimed in the '046 patent generally relate to systems, computer-implemented methods, and computer readable media associated with controlling autonomous vehicles by "judging the behavior of people near a vehicle in a way that more closely resembles how human drivers would judge

the behavior.” Ex. 1001 at 3:58-60. The ’046 patent discloses an implementation for achieving this result by obtaining responses to images displaying road scenes to determine a road user’s “state of mind—e.g. that a pedestrian is likely to cross the street or that an oncoming vehicle is unlikely to be willing to yield to the vehicle containing the sensor if the vehicle containing the sensor needs to turn.” *Id.* at 10:34-53. The images of road users (e.g., motorists, cyclists, pedestrians) in a road scene can be captured by the autonomous vehicle and sent to a remote location for obtaining responses and training a model. *Id.* at 7:56-8:14. Training dataset(s) comprising summary statistics of the user responses are used to train a learning based model so that the learning based model can predict whether a road user in a new image has a given “state of mind.” *Id.* at 10:34-53. “The trained model or algorithm makes a prediction of what a pedestrian or other person shown in the ‘live data’ would do based on the summary statistics and/or training labels” *Id.* at 7:32-35. “The predictions produced by the trained model comprise a set of predictions of the state of mind of road users that can then be used to improve the performance of autonomous vehicles, robots, virtual agents, trucks, bicycles, or other systems that operate on roadways by allowing them to make judgments about the future behavior of road users based on their state of mind.” *Id.* at 7:32-35.

B. Claim Overview

55. Independent claims 1, 8, and 15 are reproduced below. In this table, I have included the claim element numbering used in Tesla's IPR petition and Dr. Janét's declaration for the sake of continuity.

Claim 1	Limitations
1[pre]	A computer-implemented method comprising:
1[a]	storing a plurality of images, each image displaying one or more users;
1[b]	generating training data from the plurality of images, the generating comprising, for each image:
1[b-i]	sending the image to a plurality of human observers, each human observer presented with a request to answer a question about a state of mind of a user in the image,
1[b-ii]	receiving, from each of the plurality of human observers, a response representing a judgment by the human observer of the state of mind of the user in the image,
1[b-iii]	generating summary statistics describing the state of mind of the user in the image based on the received responses from the plurality of human observers, and
1[b-iv]	storing the summary statistics in association with the image as part of the training data;
1[c]	training a model using the training data, the model configured to receive an input image showing a user and predict summary statistics describing a state of mind of the user in the input image; and
1[d]	executing the trained model to predict a state of mind of a user in a new image.

Claim 8	Limitations
8[pre]	A non-transitory computer readable storage medium storing instructions that when executed by one or more processors, cause the one or more processors to perform steps comprising:
8[a]	storing a plurality of images, each image displaying one or more users;
8[b]	generating training data from the plurality of images, the generating comprising, for each image:
8[b-i]	sending the image to a plurality of human observers, each human observer presented with a request to answer a question about a state of mind of a user in the image,
8[b-ii]	receiving, from each of the plurality of human observers, a response representing a judgment by the human observer of the state of mind of the user in the image,
8[b-iii]	generating summary statistics describing the state of mind of the user in the image based on the received responses from the plurality of human observers, and
8[b-iv]	storing the summary statistics in association with the image as part of the training data;
8[c]	training a model using the training data, the model configured to receive an input image showing a user and predict summary statistics describing a state of mind of the user in the input image; and
8[d]	executing the trained model to predict a state of mind of a user in a new image.
Claim 15	Limitations
15[pre]	A computing system comprising: one or more processors; and

	a non-transitory computer readable storage medium, storing instructions that when executed by the one or more processors, cause the one or more processors to perform steps comprising:
15[a]	storing a plurality of images, each image displaying one or more users;
15[b]	generating training data from the plurality of images, the generating comprising, for each image:
15[b-i]	sending the image to a plurality of human observers, each human observer presented with a request to answer a question about a state of mind of a user in the image,
15[b-ii]	receiving, from each of the plurality of human observers, a response representing a judgment by the human observer of the state of mind of the user in the image,
15[b-iii]	generating summary statistics describing the state of mind of the user in the image based on the received responses from the plurality of human observers, and
15[b-iv]	storing the summary statistics in association with the image as part of the training data;
15[c]	training a model using the training data, the model configured to receive an input image showing a user and predict summary statistics describing a state of mind of the user in the input image; and
15[d]	executing the trained model to predict a state of mind of a user in a new image.

VII. CLAIM CONSTRUCTION

56. I understand that Perceptive proposes to construe the term “user” in the challenged claims to mean “road user (e.g., motorist, bicyclist, pedestrian).” I understand that, to date, Tesla has not proposed a construction for this or any other

term in the challenged claims. *See* Pet. at 13. Based on my review of the '046 patent claims, specification, and file history, I agree with Perceptive's proposed construction.

57. The '046 patent claims recite a number of limitations involving the term "user." For example, claim 1 recites "each image displaying one or more users," "a state of mind of a user in the image," "training a model using the training data, the model configured to receive an input image showing a user and predict summary statistics describing a state of mind of the user in the input image," and "executing the trained model to predict a state of mind of a user in a new image." Each of these limitations is consistent with and supports construing "user" as "road user (e.g., motorist, bicyclist, pedestrian)."

58. The '046 patent specification also makes clear that the claimed "user" in the context of images and a state of mind is a "road user (e.g., motorist, bicyclist, pedestrian)." *See, e.g.*, Abstract ("A computing device receives an image and a video segment of a road scene . . . response data is received, which includes at least one of an action and a likelihood of the action corresponding to another participant in the road scene."), 3:58-60 ("Systems and methods are described for judging the behavior of people near a vehicle in a way that more closely resembles how human drivers would judge the behavior."), 3:61-65 ("human observers[] view sample images of people (such as pedestrians) near streets and/or vehicles and indicate or are measured

for their understanding of how they believe the people will behave”), 6:41-47 (“the statistics stored in the database 110 cover a large set of images of people on or near roads . . . such as pedestrian, driver, motorcyclist, bicyclist, scooter driver, self-balancing scooter rider, unicyclist, motorized wheelchair user, skateboarder, or others”), 7:18-21 (“the prediction engine 114 uses the trained model from the model training system 112 to predict the actual, ‘real-world’ or ‘live data’ behavior of people on or near a road”), 7:44-50 (“The predictions produced by the trained model comprise a set of predictions of the state of mind of road users that can then be used to improve the performance of autonomous vehicles, robots, virtual agents, trucks, bicycles, or other systems that operate on roadways by allowing them to make judgments about the future behavior of road users based on their state of mind.”), 9:48-50 (“The judgment that the human observer makes is an evaluation of the state of mind of a road user depicted in the derived stimulus.”), 10:37-41 (“the pictured human road user has a given state of mind—e.g. that a pedestrian is likely to cross the street or that an oncoming vehicle is unlikely to be willing to yield to the vehicle containing the sensor if the vehicle containing the sensor needs to turn.”), 12:36-38 (“FIG. 9. is a flowchart showing a process of predicting the state of mind of road users using a trained learning algorithm.”), 13:50-54 (“The models described above can be implemented as a real-time module that makes predictions of road user behavior based on input from cameras or other sensors installed on a car 1000.”).

59. Based on my review of the '046 patent specification, there is no disclosure in the specification about the state of mind of a person that is not a road user. Even the title of the '046 patent (“System and Method of Predicting Human Interaction with Vehicles”) refers to a predicting the state of mind of a road user.

60. I have reviewed the file history of the '046 patent, and nothing in the file history is inconsistent with construing “user” to mean “road user (e.g., motorist, bicyclist, pedestrian).”

61. In my opinion, based on my review of the '046 patent claims, specification, and file history, a POSITA reading the challenged claims in light of the specification of the '046 patent would clearly understand the claimed “user” to mean a “road user (e.g., motorist, bicyclist, pedestrian).”

VIII. GROUND 1

62. I understand that Tesla alleges (and Dr. Janét opines) that claims 1-19 of the '046 patent are obvious in view of Cox alone. *See* Pet. at 13-34; Ex. 1003, ¶¶ 66-120. For the reasons I discuss below, I disagree. In my opinion, none of claims 1-19 are obvious in view of Cox alone.

A. Cox

63. In arriving at my opinions concerning Tesla and Dr. Janét’s assertions regarding claims 1-19 of the '046 patent and Cox, I have considered Cox’s disclosure

from the perspective of a POSITA. Below I discuss some aspects of Cox that are relevant to my opinions.

64. As Cox indicates, its “invention relates to machine learning, in particular to machine learning augmented by measurements acquired from human populations.” Ex. 1007, [0001].

65. Cox goes on to explain that “[t]raditional machine-learning algorithms typically incorporate human knowledge (at least implicitly) as an input.” *Id.*, [0002]. Cox gives the specific example of “a machine-learning system trained to decide whether a face is present in an image will be given set of human-generated ‘ground truth’ labels that signal whether a human judged the image to contain a face or not.” *Id.* “In this case, the job of the machine-learning system is to emit analogous labels (for example, ‘face present’ and ‘face not present’) in response to new, previously unseen images.” *Id.* I note that Cox uses this diametric “face present”/“face not present” example throughout its disclosure. *Id.* at [0002], [0008], [0010], [0012], [0013], [0016], [0017], [0026], [0029], [0033], [0039]-[0048], [0050], [0052]-[0056], Figs. 5, 6, 7, 9, 10, 11, 12 (Cox’s disclosures and experiments all describe methodologies related to face detection). In contrast, the challenged claims of the ’046 patent recite determining “a state of mind of a user” in images and are concerned with training and executing a model to predict what road users will do in a given scene (i.e., image or video). While Cox contains a couple passing references

to an annotator being asked to identify an emotion that corresponds to the image of a face (*see* Ex. 1007, [0016], [0049]), nothing in Cox describes or suggests using such emotions to train a model to predict what road users will do in such a scene.

66. As Cox also explains, one limitation of machine learning at the time was “the unavailability of labeled data; it is difficult to provide an algorithm with enough labeled training data to achieve optimal performance.” Ex. 1007, [0003]. As Cox explains, “[w]ithout sufficiently large datasets, machine-learning algorithms tend to ‘overfit’ the data, adapting to spurious structure present in the training set that is not representative of the larger distribution of all examples in the real world.” *Id.* Cox goes on to state that “[m]achine-learning systems typically combat the effects of overfitting by a process called “regularization,” in which penalties are placed on solutions that are thought to be more likely to be the result of overfitting, typically because they are more complex or because they exhibit less stable behavior under injected noise.”, *Id.*

67. To address the above, Cox proposes imposing the use of additional weightings in the objective function of the machine-learning algorithm as a form of regularization to produce more “human-like” solutions via the introduction and use of penalties for misclassification tasks on the part of the machine-learning system for examples that are easily classified by humans. Ex. 1007, [0005]. As a result, Cox proposes the acquisition and use of “[p]sychometric data characterizing the

annotation of the training objects by the annotators (i.e. human annotators)”. Ex. 1007, [0007]. To this effect, Cox describes using “psychometric data” as an additional input in the machine learning system, in addition to the human annotations of the image. *See id.* at [0005]-[0010]. As described by Cox, “[t]he psychometric data may include or consist essentially of (i) response time for classifying one or more features, (ii) accuracy of feature classification, and/or (iii) presentation time of one or more training objects.” Ex. 1007, [0008], [0010] where “[e]ach of the training objects may include or consist essentially of a digital image, and one or more features for classification may include or consist essentially of human faces.” Ex. 1007, [0008], [0010]. As Cox describes, this information can be used to determine the degree of “penalization” for misclassifying the image (i.e., indicating that the image contains a face when it does not, or vice versa), because “greater penalties may be introduced for misclassification . . . of examples that are easily classified by humans” and “lesser penalties may be introduced for misclassification of examples that are relatively more difficult for humans to classify.” *See id.*, [0005]. These penalties are used in a “human-weighted loss function” to train the model, with the aim of improving its accuracy. *See id.*, [0005]-[0010]. As Cox explains, “[t]he penalization module is executable by the computer processor and is configured to derive a human-weighted loss function based at in part on the classification data in the database and the psychometric data of at least some of the training objects in the database. The

loss function includes penalties for misclassification, and the magnitudes of the penalties increase with increasing deviation from the classification data. The classification module is executable by the computer processor and is configured to classify features of query objects based at least in part on the human-weighted loss function.” Ex. 1007, [0009].

68. Cox goes on to explain that “[o]ne component of a human-regularized maximum-margin kernel machine is the addressing of the issue of the loss function.” Ex. 1007, [0035] and further notes that “[a] prediction during training may be calculated as the output of the classifier for a particular training sample multiplied by its label: $z = yf(x)$. In accordance with various embodiments of the invention, a loss function that applies a linearly increasing cost for misclassifications (one-sided error) is desirable. This is embodied by the hinge loss function, which is defined as: $\phi(z) = \max(0, 1 - z)$ ” Ex. 1007, [0035]

69. Cox further explains that given “the non-linear nature of psychometric curves for visual recognition tasks suggests a model that is much different than linear loss growth when $z < 1$. Therefore, embodiments of the invention assume an alternative model where $wt(x)$ is a weighting function conditioned on a psychometric model for a recognition task t . The weights are determined by the classification data collected from the annotators 160, including the individual and combined statistics of accuracy, response time and presentation time, over a sampling of images across

varying degrees of difficulty. Human-weighted loss is thus defined as: $\phi(z) = \max(0, w_t(z))$ ”, Ex. 1007, [0036]. It should be noted that both the hinge loss function and the proposed human-weighted loss function are designed for binary classifications as further disclosed in Cox (see [0047]-[0049] and [0052]-[0055]) and illustrated in figures 4A/4B, 5, 6A/6B, 7A/7B, 8A/8B, 9-12.

70. Finally, Cox explains that “[e]mbodiments of the present invention incorporate a substantial pool of detection data (e.g., more than ten thousand subjects with one hundred trials per subject across all conditions, tested with approximately 7,500 face images) for occluded faces in very large, noisy, perpendicularly obstructed, and Simoncelli textured configurations. These scenarios represent an important set of difficult circumstances for detection, enabling the location of more faces through a better understanding of difficult examples during classifier training.” Ex. 1007, [0047] and provides results on face classification in figures 9, 10 and 11 with comparisons between the use of the hinge loss function and the proposed alternative Human-weighted loss function. Clearly, from the above disclosures, the focus of Cox (training, testing, use of loss functions, etc.) is primarily on face classification.

B. Cox Does Not Teach “user”/“users”

71. In my opinion, Cox does not disclose a “user” or “users,” as required by elements [a], [b-i], [b-ii], [b-iii], [c], and [d] of each of claims 1, 8, and 15.

72. As discussed above, in my opinion, the term “user” should be construed to mean “road user (e.g., motorist, bicyclist, pedestrian).” *See* Section VII.

73. In my opinion, nothing in Cox teaches the claimed “user” under this construction. Instead, the images described in Cox are of faces, and nowhere does Cox teach that those faces are those of a road user (e.g., motorist, bicyclist, pedestrian). *See, e.g.*, Ex. 1007, [0007]-[0010]. Nowhere does Tesla or Dr. Janét allege that any of the images in Cox are of road users. *See* Pet. at 13-34; Ex. 1003, ¶¶ 66-120.

74. Therefore, in my opinion, Cox does teach the claimed “user”/“users.”

C. Cox Does Not Teach Elements 1[b-iii], 8[b-iii], or 15[b-iii] (“generating summary statistics describing the state of mind of the user in the image based on the received responses from the plurality of human observers”)

75. In my opinion, Cox does not disclose elements 1[b-iii], 8[b-iii], or 15[b-iii], which each require “generating summary statistics describing the state of mind of the user in the image based on the received responses from the plurality of human observers.”

76. Tesla and Dr. Janét allege that Cox teaches this element “in multiple ways.” Pet. at 20; Ex. 1003, ¶ 77. I disagree. I discuss each of these “ways” below and the reasons they are incorrect.

Psychometric Data

77. Tesla and Dr. Janét first allege that Cox’s psychometric data constitutes the claimed “summary statistics describing the state of mind of the user in the image.” Pet. at 20-21; Ex. 1003, ¶ 77. I disagree. Cox’s psychometric data does not describe the state of mind of any user in the image that was presented to any annotator. Instead, as Cox makes clear, its psychometric data “may include or consist essentially of (i) response time for classifying one or more features, (ii) accuracy of feature classification, and/or (iii) presentation time of one or more training objects.” Ex. 1007 at [0010]; *see also id.*, [0008], [0018], [0029]. The annotator’s response time for classifying features in the image does not describe the state of mind of any user in the image, but is instead merely a measure of how long it took the annotator to review the image and classify it. Likewise, the accuracy of the feature classification by the annotator does not describe the state of mind of any user in the image, but is instead a measure of the correctness of the annotator’s classification. And finally, the presentation time of the training object(s) does not describe the state of mind of any user in the image, but is merely a measure of how long the image was presented to the annotator(s).

78. I have considered the paragraphs of Cox that the Petition and Dr. Janét’s declaration cite in support of their argument, and in my opinion, none of them support the allegation that somehow Cox’s psychometric data can constitute

“summary statistics describing the state of mind of the user in the image.” *See* Pet. at 21; Ex. 1003, ¶ 77 (both citing Ex. 1007, “¶¶0002-0008, 0010, 0014, 0018, 0025-0032, 0035, 0049, 0063, Fig. 3”). Instead, those paragraphs (to the extent they discuss psychometric data, which not all of them do) describe psychometric data in the manner I describe in the previous paragraph. Both the Petition and Dr. Janét’s declaration also cross-reference their respective discussions of elements 1[b-i] and 1[b-ii]. *See* Pet. at 21; Ex. 1003, ¶ 77. I have reviewed those sections, as well as the portions of Cox cited in them, and they too only (to the extent they discuss psychometric data, which not all of them do) discuss psychometric data in the manner I describe in the previous paragraph.

79. Tesla and Dr. Janét also allege that “Cox’s description of “psychometric data” is consistent with the ’046 Patent’s explanation.” Pet. at 21; Ex. 1003, ¶ 77. I have reproduced the portion of the ’046 patent below that Tesla and Dr. Janét refer to (with the surrounding language for context):

In step 210, summary statistics are generated based on the user responses. **For example, the statistics may characterize the aggregate responses of multiple human observers to a particular derived stimulus. For instance, if the derived stimulus shows a pedestrian walking on a sidewalk towards an intersection, the response can be categorized in terms of how many human observers believe that the pedestrian will stop upon reaching the intersection, continue walking straight across the intersection, turn**

a corner and continue walking along the sidewalk without crossing the intersection, etc. *These summary statistics can characterize the human observer responses in terms of certain parameters associated with the statistics, such as a content of a response, a time associated with entering a response, and a position of an eye of a human observer associated with the response.* The parameters can also be associated with a (1) central tendency, variance, skew, kurtosis, scale, or histogram. For example, the amount of time users took to input their responses can be characterized in terms of central tendency, variance, skew, kurtosis, scale, histogram. Also, the statistics can include a parameter that additionally or alternatively characterizes the movement of the human observers' eyes relative to a display when making the judgments in terms of central tendency, variance, skew, kurtosis, scale, histogram or two-dimensional distribution. In one embodiment, the statistics are stored in the user response database 110 with an index that identifies the raw video or sensor data from which the derived stimulus was generated. In a further embodiment, the statistics stored in the database 110 cover a large set of images of people on or near roads and are categorized in a number of different categories, such as pedestrian, driver, motorcyclist, bicyclist, scooter driver, self-balancing scooter rider, unicyclist, motorized wheelchair user, skateboarder, or others. Moreover, the statistics are respectively stored along with, or linked to, the images of the derived stimuli corresponding to the statistics.

'046 patent, 6:14-49. (emphasis and italicization added)

80. In the excerpt above, the italicized sentence is the only portion that Tesla and Dr. Janét identify. The preceding sentences (which I have bolded above)

make clear that the summary statistics “characterize the aggregate responses of multiple human observers to a particular derived stimulus,” and the ’046 patent mentions the specific example of “if the derived stimulus shows a pedestrian walking on a sidewalk towards an intersection, the response can be categorized in terms of how many human observers believe that the pedestrian will stop upon reaching the intersection, continue walking straight across the intersection, turn a corner and continue walking along the sidewalk without crossing the intersection.” ’046 patent, 6:15-24. Each of these types of responses “describe[es] the state of mind of the user in the image” as required by the challenged claims. The mere fact that the italicized sentence indicates that “[t]hese summary statistics can characterize the human observer responses in terms of certain parameters associated with the statistics, such as . . . a time associated with entering a response” (’046 patent, 6:24-27) does not mean that such a time describes the state of mind of the user in the image. Instead, such a time relates to one or more other factors, such as the annotator’s skill/speed and/or the degree of simplicity in discerning the relevant features in the image. At best, this language in the specification of the ’046 patent suggests that “summary statistics” can include such response times, but nowhere does the ’046 patent suggest that such response times themselves can or do constitute “summary statistics describing the state of mind of the user in the image,” as required by the challenged claims.

Item-Response Curves

81. Tesla and Dr. Janét next allege that Cox’s “item-response curves” can constitute the claimed “summary statistics describing the state of mind of the user in the image.” Pet. at 22; Ex. 1003, ¶ 78. I disagree.

82. The portion of Cox that Tesla and Dr. Janét point to does not describe generating item-response curves as part of the human annotation of images and/or the training process for a machine learning model. Pet. at 22; Ex. 1003, ¶ 78 (both citing Ex. 1007, [0006]). Instead, that portion of Cox merely mentions item-response curves as one of several “well-known techniques . . . for measuring human performance” and indicates that such techniques “may be used as additional streams of input in accordance with embodiments of the present invention.” Ex. 1007, [0006]. Nothing in this disclosure in Cox suggests that these item-response curves “describ[e] the state of mind of the user in the image” as required by the challenged claims. Tesla and Dr. Janét allege that “[b]ecause these response curves present an analysis of collected numerical measurements, a POSITA would have recognized that the curves form statistics regarding the annotations.” Pet. at 22; Ex. 1003, ¶ 78. Even if true, that does not mean that item-response curves “describe[e] the state of mind of the user in the image.”

83. Cox’s only other mention of item-response curves (which neither the Petition nor Dr. Janét’s declaration mention) is an indication that “[t]he

psychometric data may be displayed to analyze, e.g., . . . an item response curve 370 for each of the training objects (i.e., the presented image).” *Id.*, [0029]; *see also* Fig. 3. The displaying of psychometric data to analyze an item-response curve for a training object does not teach or suggest that the item-response curve “describ[es] the state of mind of the user in the image” as required by the challenged claims.

84. As Figure 3 shows, an item-response curve plots an annotator’s ability versus their accuracy in correctly annotating an image relative to ground truth. Such item-response curves (sometimes referred to as item characteristic curves or “ICCs”) have been used in research and in educational testing situations for many decades. For example, early research connected to the Educational Testing Service shows ICCs like those shown in Figure 3 of Cox. *See* Ex. 2041 at Fig. 1. Nothing about these item-response curves or ICCs “describ[es] the state of mind of the user in the image” as required by the challenged claims.

85. I have considered the additional portions of Cox that Tesla and Dr. Janét identify in connection with this discussion. *See* Pet. at 22; Ex. 1003, ¶ 78 (both citing Ex. 1007, “¶¶ 0002-0008, 0010, 0014, 0018, 0025-0032, 0035, 0049, 0063, Fig. 3”). None of those cited portions contain any disclosure that teaches or suggests that item-response curves “describ[e] the state of mind of the user in the image” as required by the challenged claims. Likewise, I have considered the sections of the Petition and Dr. Janét’s declaration that they each cross-reference (*see* Pet. at 22; Ex.

1003, ¶ 78), but none of those sections point to any disclosure in Cox that teaches or suggests that item-response curves “describ[e] the state of mind of the user in the image” as required by the challenged claims.

86. As I discuss more below, while Tesla and Dr. Janét allege that Cox’s item-response curves meet the claimed summary statistics for elements 1[b-iii], 8[b-iii], and 15[b-iii], Tesla and Dr. Janét do not allege that Cox’s item-response curves meet the claimed “the summary statistics” for elements 1[b-iv], 8[b-iv], and 15[b-iv]. *See* Pet. at 23-24; Ex. 1003, ¶ 80. I understand that the fact that elements 1[b-iv], 8[b-iv], and 15[b-iv] recite “*the* summary statistics” means that these elements refer back to the summary statistics recited in elements 1[b-iii], 8[b-iii], and 15[b-iii], respectively. Because Tesla and Dr. Janét have not demonstrated how the item-response curves mentioned in Cox meet all of the claims’ requirements of “summary statistics,” this is an additional reason that Tesla and Dr. Janét have not shown how Cox’s item-response curves can constitute the claimed “summary statistics.”

Human-Weighted Loss Function

87. Tesla and Dr. Janét next allege that Cox’s “human-weighted loss function” can constitute the claimed “summary statistics describing the state of mind of the user in the image.” Pet. at 22; Ex. 1003, ¶ 79. I disagree.

88. As Cox repeatedly describes, its human-weighted loss function is “computationally derived” “based at least in part on the classification data and the

psychometric data.” Ex. 1007, [0007]; *see also id.*, [0008], [0009], [0019], [0036].

While the human-weighted loss function that Cox describes can be derived based in part on the classification data, the human-weighted loss function itself does not “describ[e] the state of mind of the user in the image.” Instead, the human-weighted loss function is a computational result derived from the classification data and the psychometric data (as disclosed by Cox on multiple occasions including [0009], [0032], [0034]-[0036]), and it does not itself describe the user in the image or their state of mind.

89. Generally speaking, a loss function in machine learning applications is a mathematical measure that quantifies the difference between a model’s predictions and the actual target values. It provides a numerical estimate of prediction error, which learning algorithms use to evaluate model performance and guide parameter updates (e.g. network weights, hyperparameters) during training. The objective of the training process is to minimize the loss function (i.e. minimize the difference between the predicted and the actual target value), enabling the model to make more accurate predictions as the weights/parameters of the network are updated using algorithms such as backpropagation.

90. Different machine learning tasks use different loss functions. For example, Mean Squared Error (MSE) is commonly used for regression problems, while Cross-Entropy Loss is widely used for classification tasks. By minimizing the

loss, the model learns patterns in the data and improves its generalization to unseen examples thereby minimizing the effects of overfitting. As Cox itself explains, an additional way “[m]achine learning systems typically combat the effects of overfitting [is] by a process called ‘regularization’ in which penalties are placed on solutions that are thought to be more likely to be the result of overfitting, typically because they are more complex or because they exhibit less stable learning behavior under injected noise.” Ex. 1007, [0003].

91. Cox explains that its “human-weighted loss function may include or consist essentially of (i) computationally classifying, by the computer processor, at least some of the training objects based at least in part on an initial loss function, thereby generating training data, (ii) comparing the training data to the classification data to identify, within the training data, features misclassified in comparison to the classification data, (iii) assigning the penalties for misclassification to the misclassified features in the training data, and (iv) incorporating the penalties for misclassification within the initial loss function to generate the human-weighted loss function.” Ex. 1007, [0008].

92. Cox explains that “[o]ne component of a human-regularized maximum-margin kernel machine is the addressing of the issue of the loss function” and that “a [hinge] loss function that applies a linearly increasing cost for misclassifications (one-sided error) is desirable.” Ex. 1007, [0035].

93. Cox also indicates that “the non-linear nature of psychometric curves for visual recognition tasks suggests a model that is much different than linear loss growth when $z < 1$.” Ex. 1007, [0036]. Therefore, Cox proposes the use of an “alternative model where $w_t(x)$ is a weighting function conditioned on the psychometric model for a recognition task t . The weights are determined by the classification data collected from the annotators 160, including the individual and combined statistics of accuracy, response time and presentation time, over a sampling of images across varying degrees of difficulty.” Cox refers to this as its human-weighted loss, which it defines as: $\phi(z) = \max(0, w_t(z))$. Ex. 1007, [0036].

94. Nowhere in the above disclosures or descriptions does Cox teach or suggest that the human-weighted loss function generates summary statistics describing the state of mind of the user in the image as claimed in elements 1[b-iii], 8[b-iii], and 15[b-iii] or serve to predict summary statistics describing a state of mind of the user as claimed in elements 1[c], 8[c], and 15[c]. To the contrary, the human-weighted loss function of Cox is proposed as an improvement over a hinge loss function to incorporate the nonlinear nature of psychometric curves in the training of the classification algorithm of Cox.

95. I have considered the additional portions of Cox that Tesla and Dr. Janét identify in connection with this discussion. *See* Pet. at 22; Ex. 1003, ¶ 79 (both citing

Ex. 1007, “0002-0008, 0010, 0014, 0018, 0025-0032, 0035, 0048-0049, 0063, Fig. 3”). None of those cited portions contains any disclosure that teaches or suggests that a human-weighted loss function “describ[es] the state of mind of the user in the image” as required by the challenged claims. Likewise, I have considered the sections of the Petition and Dr. Janét’s declaration that they each cross-reference (*see* Pet. at 22; Ex. 1003, ¶ 79), but none of those sections point to any disclosure in Cox that teaches or suggests that a human-weighted loss function “describ[es] the state of mind of the user in the image” as required by the challenged claims.

96. As I discuss more below, while Tesla and Dr. Janét allege that Cox’s human-weighted loss function meets the claimed summary statistics for elements 1[b-iii], 8[b-iii], and 15[b-iii], Tesla and Dr. Janét do not allege that Cox’s human-weighted loss function meets the claimed “the summary statistics” for elements 1[b-iv], 8[b-iv], and 15[b-iv]. *See* Pet. at 23-24; Ex. 1003, ¶ 80. I understand that the fact that elements 1[b-iv], 8[b-iv], and 15[b-iv] recite “*the* summary statistics” means that these elements refer back to the summary statistics recited in elements 1[b-iii], 8[b-iii], and 15[b-iii], respectively. Because Tesla and Dr. Janét have not demonstrated how the human-weighted loss function in Cox meets all of the claims’ requirements of “summary statistics,” this is an additional reason that Tesla and Dr. Janét have not shown how Cox’s human-weighted loss function can constitute the claimed “summary statistics.”

End-of-Test Statistics

97. Tesla and Dr. Janét next allege that Cox’s “end-of-test statistics” can constitute the claimed “summary statistics describing the state of mind of the user in the image.” Pet. at 22-23; Ex. 1003, ¶ 79. I disagree.

98. Cox’s only disclosure of end-of-test statistics is in connection with Figure 8, as to which Cox states the following:

For example, annotators may be asked to press a specific key if they observe a face in an image 810 and another key if they do not observe a face in the image. At the end of a series of such queries, end-of-test statistics 820 may be displayed to the annotators 160 via display module 167.

Ex. 1007, [0048].

99. As the above disclosure makes clear, these end-of-test statistics are displayed to the annotators on a display module when they have completed reviewing the images. Cox does not offer any disclosure as to the contents of these end-of-test statistics, and certainly does not teach or suggest that they “describ[e] the state of mind of the user in the image” as required by the challenged claims.

100. Dr. Janét opines that “a POSITA would have understood from Cox that this data ‘may be used to inform machine learning in accordance with embodiments of the invention.’” Ex. 1003, ¶ 79. I disagree with Dr. Janét’s conclusion, as nothing in paragraph [0048] of Cox indicates that the “data” that “may be used to inform machine learning” is referring to the end-of-test statistics mentioned several

sentences earlier. In my opinion, because Cox describes the end-of-test statistics as something that is presented to the annotator at the end of their review of the images, such statistics would not be of the type that would be used to inform the machine learning algorithms, but would instead be statistics such as the annotator's accuracy and/or speed.

101. Even setting aside that issue and assuming that the end-of-test statistics described in this one sentence in Cox are “used to inform machine learning,” that does not at all teach or suggest that such end-of-test statistics “describ[e] the state of mind of the user in the image” as required by the challenged claims.

102. I have considered the additional portions of Cox that Tesla and Dr. Janét identify in connection with this discussion. *See* Pet. at 23; Ex. 1003, ¶ 79 (both citing Ex. 1007, “0002-0008, 0010, 0014, 0018, 0025-0032, 0035, 0048-0049, 0063, Fig. 3”). None of those cited portions contains any disclosure that teaches or suggests that these end-of-test statistics “describ[e] the state of mind of the user in the image” as required by the challenged claims. Likewise, I have considered the sections of the Petition and Dr. Janét's declaration that they each cross-reference (*see* Pet. at 23; Ex. 1003, ¶ 79), but none of those sections point to any disclosure in Cox that teaches or suggests that these end-of-test statistics “describ[e] the state of mind of the user in the image” as required by the challenged claims.

103. The portions of the Petition and Dr. Janét’s declaration discussed above relate to claim 1 of the ’046 patent. Both the Petition and Dr. Janét’s declaration merely refer back to these discussions when discussing element 8[b-iii] and 15[b-iii]. *See* Pet. at 32-33; Ex. 1003, ¶¶ 98, 113. In my opinion, Cox fails to teach or suggest elements 8[b-iii] and 15[b-iii] for the same reasons as discussed above.

104. As I discuss more below, while Tesla and Dr. Janét allege that Cox’s end-of-test statistics meet the claimed summary statistics for elements 1[b-iii], 8[b-iii], and 15[b-iii], Tesla and Dr. Janét do not allege that Cox’s end-of-test statistics meet the claimed “the summary statistics” for elements 1[b-iv], 8[b-iv], and 15[b-iv]. *See* Pet. at 23-24; Ex. 1003, ¶ 80. I understand that the fact that elements 1[b-iv], 8[b-iv], and 15[b-iv] recite “*the* summary statistics” means that these elements refer back to the summary statistics recited in elements 1[b-iii], 8[b-iii], and 15[b-iii], respectively. Because Tesla and Dr. Janét have not demonstrated how the end-of-test statistics in Cox meet all of the claims’ requirements of “summary statistics,” this is an additional reason that Tesla and Dr. Janét have not shown how Cox’s end-of-test statistics can constitute the claimed “summary statistics.”

105. Therefore, in my opinion, Cox does not teach or suggest elements 1[b-iii], 8[b-iii], or 15[b-iii] (“generating summary statistics describing the state of mind of the user in the image based on the received responses from the plurality of human observers”).

D. Cox Does Not Teach Elements 1[b-iv], 8[b-iv], or 15[b-iv] (“storing the summary statistics in association with the image as part of the training data”)

106. In my opinion, Cox does not disclose elements 1[b-iv], 8[b-iv], or 15[b-iv], which each require “storing the summary statistics in association with the image as part of the training data.”

107. Tesla and Dr. Janét allege that Cox teaches this element through its disclosure of a “database of training objects” that is “populated with stored computer records specifying, for each of the plurality of objects, (i) classification data comprising annotations received from a plurality of human annotators’ and ‘(ii) psychometric data characterizing the annotation of the training object by the plurality of human annotators.”” Pet. at 23; Ex. 1003, ¶ 80 (both citing Ex. 1007, [0009]). I disagree. In making this allegation, Tesla and Dr. Janét again allege that Cox’s psychometric data constitutes the claimed summary statistics. However, for all of the reasons I discuss above with respect to elements 1[b-iii], 8[b-iii], and 15[b-iii], Cox’s psychometric data cannot constitute the claimed summary statistics, as that data does not describe the state of mind of the user in the image. For those same reasons, Cox’s psychometric data cannot constitute the claimed summary statistics for elements 1[b-iv], 8[b-iv], or 15[b-iv].

108. I have considered the additional portions of Cox that Tesla and Dr. Janét identify with respect to this claim element. *See* Pet. at 23-24; Ex. 1003, ¶ 80. None

of them alter my opinions expressed above that Cox's psychometric data does not constitute the claimed summary statistics. Instead, those portions of Cox merely reinforce the idea that Cox's psychometric data merely includes information such as response time and presentation time, which do not describe the state of mind of a user in the image. *See, e.g.*, Ex. 1007, [0018], [0029].

109. I note that, for elements 1[b-iii], 8[b-iii], and 15[b-iii] (and as I discuss above with respect to those elements), Tesla and Dr. Janét allege that other aspects of Cox's disclosure (i.e., item-response curves, the human-weighted loss function, or the end-of-test statistics) can constitute the claimed summary statistics. Neither the Petition nor Dr. Janét's declaration make any effort to demonstrate how any of those other items mentioned in Cox could meet the claimed summary statistics in elements 1[b-iv], 8[b-iv], or 15[b-iv]. *See* Pet. at 23-24; Ex. 1003, ¶ 80. I understand that the fact that elements 1[b-iv], 8[b-iv], and 15[b-iv] recite "*the* summary statistics" means that these elements refer back to the summary statistics recited in elements 1[b-iii], 8[b-iii], and 15[b-iii], respectively. Because Tesla and Dr. Janét have not demonstrated how these other items mentioned in Cox meet all of the claims' requirements of "summary statistics," this is an additional reason that Tesla and Dr. Janét have not shown how these other items can constitute the claimed "summary statistics."

110. Even setting this failure aside, nothing in Cox suggests storing the item-response curves, the human-weighted loss function, or the end-of-test statistics in any database, or storing them “in association with the image as part of the training data,” as required by the challenged claims. Cox merely mentions item-response curves as one of several “well-known techniques . . . for measuring human performance” and indicates that such techniques “may be used as additional streams of input in accordance with embodiments of the present invention.” Ex. 1007, [0006]. As for the human-weighted loss function, Cox repeatedly describes the human-weighted loss function as “computationally derived” “based at least in part on the classification data and the psychometric data.” *Id.*, [0007]; *see also id.*, [0008], [0009], [0019], [0036]. But nowhere does Cox describe storing it anywhere, much less in association with any image. And, Cox merely indicates that “end-of-test statistics 820 may be displayed to the annotators 160 via display module 167” (*id.*, [0048]) but says nothing about storing it anywhere, much less in association with any image.

111. The portions of the Petition and Dr. Janét’s declaration discussed above relate to claim 1 of the ’046 patent. Both the Petition and Dr. Janét’s declaration merely refer back to these discussions when discussing element 8[b-iv] and 15[b-iv]. *See* Pet. at 32-34; Ex. 1003, ¶¶ 99, 114. In my opinion, Cox fails to teach or suggest elements 8[b-iii] and 15[b-iii] for the same reasons as discussed above.

112. Therefore, in my opinion, Cox does not teach or suggest elements 1[b-iv], 8[b-iv], or 15[b-iv] (“storing the summary statistics in association with the image as part of the training data”).

E. Cox Does Not Teach Elements 1[c], 8[c], or 15[c] (“training a model using the training data, the model configured to receive an input image showing a user and predict summary statistics describing a state of mind of the user in the input image”)

113. In my opinion, Cox does not disclose elements 1[c], 8[c], or 15[c], which each require “training a model using the training data, the model configured to receive an input image showing a user and predict summary statistics describing a state of mind of the user in the input image.”

114. Tesla and Dr. Janét allege that this element is met through Cox’s disclosure that “its model is trained by (1) receiving annotated training images showing emotional states, (2) predicting classifications for the training images using the model, and (3) optimizing the model such that differences between the predicted classifications and annotated training images are minimized.” Pet. at 24-25; Ex. 1003, ¶ 81. Thus, as best I understand Tesla and Dr. Janét’s assertion, they appear to allege that “predicting classifications for the training images using the model” in Cox meets the claimed requirement that “the model [be] configured to . . . predict summary statistics describing a state of mind of the user in the input image,” as that is the only thing that Tesla and Dr. Janét allege that Cox predicts. I note that this assertion is different than for elements 1[b-iii], 8[b-iii], and 15[b-iii] discussed above

(where Tesla and Dr. Janét allege that Cox’s psychometric data, item-response curves, human-weighted loss function, or end-of-test statistics constitute the claimed summary statistics), and also different than for elements 1[b-iv], 8[b-iv], and 15[b-iv] (where Tesla and Dr. Janét only allege that Cox’s psychometric data constitutes the claimed summary statistics). Tesla and Dr. Janét have not provided any explanation for why a POSITA would understand the challenged claims to encompass generating one type of summary statistics in those other claim elements, yet have a model configured to predict a different type of summary statistics in elements 1[c], 8[c], and 15[c].

115. Even setting aside this shortcoming in the Petition and Dr. Janét’s declaration, I disagree that Cox’s “classifications for the training images” can constitute the claimed summary statistics. In the context of Cox’s disclosure, classification is the decision (whether by a human annotator or a model) of how to categorize the object in the image. The cited disclosures of Cox are generally focused on classifying an image as either containing or not containing a face. Pet. at 24-25; Ex. 1003, ¶ 81 (both citing Ex. 1007, “¶¶0016-0020, 0029-0037, 0049, 0057, 0060; Fig. 4B (430)”). Even the limited discussion of classifications beyond the “face present”/“face not present” classification do not teach predicting any summary statistics. *See, e.g.*, Ex. 1007, [0016] (describing examples of “a participant [being] be shown . . . an image 110 of a person and asked that person’s gender” or “an image

120 of all or a portion of a person's face and asked to select an emotion that corresponds to the image, e.g., jealous, panicked, arrogant, or hateful"). This disclosure is also in the context of human annotation, but even setting that aside, none of these binary or categorical classifications are any kind of prediction of summary statistics.

116. Also, even if Tesla and Dr. Janét had alleged that Cox's psychometric data, item-response curves, human-weighted loss function, or end-of-test statistics meet the "summary statistics" in these element, I disagree that any of those items meet these elements because nothing in Cox describes its model being configured to predict any of those items.

117. I have considered the additional portions of Cox that Tesla and Dr. Janét identify with respect to this claim element. *See* Pet. at 24-25; Ex. 1003, ¶ 81. None of them alter my opinions expressed above that Cox's predicting classifications for the training images does not constitute predicting summary statistics. Instead, those portions of Cox merely reinforce the idea that Cox's predicting classifications amount to predicting how to categorize the image (i.e., face/no face, or a given gender or emotion), none of which constitute predicting summary statistics.

118. The portions of the Petition and Dr. Janét's declaration discussed above relate to claim 1 of the '046 patent. Both the Petition and Dr. Janét's declaration merely refer back to these discussions when discussing element 8[c] and 15[c]. *See*

Pet. at 32, 34; Ex. 1003, ¶¶ 100, 115. In my opinion, Cox fails to teach or suggest elements 8[c] and 15[c] for the same reasons as discussed above.

119. Therefore, in my opinion, Cox does not teach or suggest elements 1[c], 8[c], or 15[c] (“training a model using the training data, the model configured to receive an input image showing a user and predict summary statistics describing a state of mind of the user in the input image”).

F. Both the Petition and Dr. Janét’s Declaration Fail to Analyze the Factors Required for a Proper Obviousness Analysis

120. As I discuss in Section IV above, one of the factors I understand that is required to be considered in an obviousness analysis is the differences between the prior art and the claims. While I note that Dr. Janét mentions this requirement in his declaration as well (*see* Ex. 1003, ¶ 40), nowhere in his discussion of Ground 1 does Dr. Janét discuss any differences between Cox and the challenged claims of the ’046 patent.

121. As I discussed above, Cox’s disclosure is focused on binary face detection processes, and many of its features are specifically geared toward that objective, including its use of a human-weighted loss function. In contrast, the challenged claims of the ’046 patent relate to training and executing a machine learning model to predict summary statistics that describe the state of mind of a road user in an image. Dr. Janét does not account for these key differences between Cox and the challenged claims. As I discuss above, these differences are not miniscule or

insignificant because they are rooted in the specific (but differing) applications that Cox and the challenged claims are each aimed to address, and they are a key factor in the opinions I express above concerning the elements of the challenged claims that Cox does not teach or suggest.

IX. GROUND 2

122. I understand that Tesla alleges (and Dr. Janét opines) that claims 1-19 of the '046 patent are obvious in view of Cox and Ross. *See* Pet. at 34-45; Ex. 1003, ¶¶ 121-150. For the reasons I discuss below, I disagree. In my opinion, none of claims 1-19 are obvious in view of Cox and Ross.

A. Cox and Ross

123. In arriving at my opinions concerning Tesla and Dr. Janét's assertions regarding claims 1-19 of the '046 patent and Cox and Ross, I have considered both Cox and Ross's disclosures from the perspective of a POSITA. In my discussion of Ground 1 above, I discussed aspects of Cox that are relevant to my opinions. Below I discuss some aspects of Ross that are relevant to my opinions.

124. Ross describes "[b]ehavior and intent estimations of road users for autonomous vehicles" (Ex. 1005, Title) which represents an entirely different application than Cox face/no face detection system. Ross explains the following:

One aspect of the disclosure provides a method of maneuvering a vehicle. The method includes receiving, by one or more processors, data from a perception system of the vehicle identifying an object

corresponding to a road user, the information also identifying characteristics of the road user as well as contextual information about an environment in which the vehicle is currently driving; generating, by the one or more processors, a prediction of the intent of the object using the data, wherein the prediction of the intent of the object includes a next action to be initiated by the road user and a point in time for initiation of the next action, wherein the next action is included in a predetermined list of actions; generating, by the one or more processors, a prediction of the behavior of the road user for a predetermined period of time into the future using the data, the prediction of the behavior indicating a likelihood that the road user will initiate the next action during the predetermined period of time, the predetermined period of time including at least a period of time during which the vehicle is expected to have passed the object; and when the prediction of the behavior indicates that the road user is not likely to initiate the next action during the predetermined period of time, maneuvering, by the one or more processors, the vehicle according to the prediction of the intent prior to the vehicle passing the object.

Ex. 1005, 1:30-54. Figure 10 of Ex. 1005 provides an example flow diagram in accordance with aspects of the proposed disclosure.

125. Ross does not disclose or mention any details with regards to summary statistics, probability distributions of users, user responses associated with the object and/or annotations by human observers.

B. Cox and Ross Do Not Teach Elements 1[b-iii], 8[b-iii], or 15[b-iii] (“generating summary statistics describing the state of mind of the user in the image based on the received responses from the plurality of human observers”)

126. In my opinion, the combination of Cox and Ross (as Tesla and Dr. Janét propose) does not teach elements 1[b-iii], 8[b-iii], or 15[b-iii], which each require “generating summary statistics describing the state of mind of the user in the image based on the received responses from the plurality of human observers.”

127. Tesla and Dr. Janét merely refer back to their discussion of Cox in Ground 1 for these elements. *See* Pet. at 42; Ex. 1003, ¶ 137. In other words, Tesla and Dr. Janét do not rely on any disclosure from Ross as supposedly meeting this element, but instead rely on Cox alone.

128. As I discussed above with respect to Ground 1, in my opinion, Cox does not teach elements 1[b-iii], 8[b-iii], or 15[b-iii] because, among other reasons, the items Tesla and Dr. Janét identify from Cox (i.e., psychometric data, item-response curves, the human-weighted loss function, and the end-of-test statistics) are not “summary statistics describing the state of mind of the user in the image.” Tesla and Dr. Janét’s proposed combination of Ross with Cox does not alter my opinions in that regard.

129. The portions of the Petition and Dr. Janét’s declaration discussed above relate to claim 1 of the ’046 patent. Both the Petition and Dr. Janét’s declaration merely refer back to these discussions when discussing element 8[c] and 15[c]. *See*

Pet. at 45; Ex. 1003, ¶¶ 144, 147. In my opinion, Cox fails to teach or suggest elements 8[c] and 15[c] for the same reasons as discussed above.

130. Therefore, in my opinion, Cox and Ross do not teach or suggest elements 1[b-iii], 8[b-iii], or 15[b-iii] (“generating summary statistics describing the state of mind of the user in the image based on the received responses from the plurality of human observers”).

C. Cox and Ross Do Not Teach Elements 1[b-iv], 8[b-iv], or 15[b-iv] (“storing the summary statistics in association with the image as part of the training data”)

131. In my opinion, the combination of Cox and Ross (as Tesla and Dr. Janét propose) does not teach elements 1[b-iv], 8[b-iv], or 15[b-iv], which each require “storing the summary statistics in association with the image as part of the training data.”

132. Tesla and Dr. Janét merely refer back to their discussion of Cox in Ground 1 for these elements. *See* Pet. at 42; Ex. 1003, ¶ 137. In other words, Tesla and Dr. Janét do not rely on any disclosure from Ross as supposedly meeting this element, but instead rely on Cox alone.

133. As I discussed above with respect to Ground 1, in my opinion, Cox does not teach elements 1[b-iv], 8[b-iv], or 15[b-iv] because, among other reasons, the psychometric data Tesla and Dr. Janét identifies from Cox is not “summary statistics describing the state of mind of the user in the image,” and Tesla and Dr. Janét do not

offer any explanation for how the other items identified for elements 1[b-iii], 8[b-iii], and 15[b-iii] (i.e., the item-response curves, the human-weighted loss function, and the end-of-test statistics) supposedly meet this element. Tesla and Dr. Janét's proposed combination of Ross with Cox does not alter my opinions in that regard.

134. The portions of the Petition and Dr. Janét's declaration discussed above relate to claim 1 of the '046 patent. Both the Petition and Dr. Janét's declaration merely refer back to these discussions when discussing element 8[c] and 15[c]. *See* Pet. at 45; Ex. 1003, ¶¶ 144, 147. In my opinion, Cox fails to teach or suggest elements 8[c] and 15[c] for the same reasons as discussed above.

135. Therefore, in my opinion, Cox and Ross do not teach or suggest elements 1[b-iv], 8[b-iv], or 15[b-iv] ("storing the summary statistics in association with the image as part of the training data").

D. Cox and Ross Do Not Teach Elements 1[c], 8[c], or 15[c] ("the model configured to receive an input image showing a user and predict summary statistics describing a state of mind of the user in the input image")

136. In my opinion, the combination of Cox and Ross (as Tesla and Dr. Janét propose) does not teach elements 1[c], 8[c], or 15[c], which each require "the model configured to receive an input image showing a user and predict summary statistics describing a state of mind of the user in the input image."

137. In discussing this element, Tesla and Dr. Janét allege the following:

As combined, the training of Cox’s model would have been performed as discussed in Section VI.A.2.h (element 1[c] – Ground 1) except the training data includes training images and annotations with respect to actions by a person in images reflecting possible intents of the person, and Cox’s model is trained on such training data as discussed in Section VI.A.2.h (element 1[c] – Ground 1). See Sections VI.B.2, VI.B.3.b, VI.B.3.c, VI.A.2.h.

Pet. at 42-43; Ex. 1003, ¶ 138.

138. Nothing in this discussion mentions any disclosure of Ross as supposedly teaching a model that is configured to “predict summary statistics describing a state of mind of the user in the input image.” In fact, no portion of Ross is mentioned at all in this discussion.

139. Tesla and Dr. Janét refer back to their discussion of Cox in Ground 1 for these elements, as well as to their discussions of alleged motivation to combine Cox and Ross and to elements 1[b-i]-[b-iv]/8[b-i]-[b-iv]/15[b-i]-[b-iv]. *See* Pet. at 42-43; Ex. 1003, ¶ 138. Tesla and Dr. Janét’s discussion of Ground 1 concerns Cox alone, and while nothing in their discussion of elements 1[c], 8[c], or 15[c] discusses or references any portion of Ross, some of these cross-referenced sections do.

140. Tesla and Dr. Janét’s discussion of alleged motivation to combine Cox and Ross (Pet. at 36-40; Ex. 1003, ¶¶ 125-130) does not discuss anything about predicting summary statistics and instead only includes one passing reference to “Ross’s disclosure of observed road user actions demonstrating intents and

prediction model predicting road user intents.” Pet. at 39; Ex. 1003, ¶ 130. This lone passing reference does not explain whether Tesla and Dr. Janét are alleging that Ross (or the combination of Cox and Ross) teaches these elements by allegedly “predicting road user intents.” Setting aside that lack of clarity, I disagree that Ross’s teaching of predicting road user intents amounts to a teaching of a “model configured to receive an input image showing a user and predict summary statistics describing a state of mind of the user in the input image,” because the mere predicting of road user intents is not a prediction of any summary statistics. Tesla and Dr. Janét offer no explanation here for how Ross’s teaching of predicting such intents in any way involves or includes predicting any summary statistics.

141. Tesla and Dr. Janét’s discussion of elements 1[b-i]/8[b-i]/15[b-i] likewise discusses Ross’s prediction of road user intents in passing. *See* Pet. 40-42; Ex. 1003, ¶¶ 133-135. However, nothing in that discussion is alleged to be a summary statistic of any kind that is allegedly predicted by Ross’s model (or the alleged model of the combination of Cox and Ross).

142. Tesla and Dr. Janét’s discussion of elements 1[b-ii]-[b-iv]/8[b-ii]-[b-iv]/15[b-ii]-[b-iv] contains no discussion of Ross, much less any discussion about predicting summary statistics. *See* Pet. 42; Ex. 1003, ¶¶ 136-137.

143. Tesla and Dr. Janét also refer back to their discussion of element 1[c] from Ground 1. Pet. at 42-43; Ex. 1003, ¶ 138 (both referring to Section VI.A.2.h).

That section (like all sections related to Ground 1) only concerns Cox, not Ross. Additionally, as I discussed above with respect to Ground 1 and elements 1[c]/8[c]/15[c], Cox's prediction of classifications does not teach any prediction of summary statistics. Nothing about Tesla and Dr. Janét's combination of Cox with Ross alters my opinions in that regard.

144. The portions of the Petition and Dr. Janét's declaration discussed above relate to claim 1 of the '046 patent. Both the Petition and Dr. Janét's declaration merely refer back to these discussions when discussing element 8[c] and 15[c]. *See* Pet. at 45; Ex. 1003, ¶¶ 144, 147. In my opinion, Cox fails to teach or suggest elements 8[c] and 15[c] for the same reasons as discussed above.

145. Therefore, in my opinion, Cox and Ross do not teach or suggest elements 1[c], 8[c], or 15[c] ("the model configured to receive an input image showing a user and predict summary statistics describing a state of mind of the user in the input image").

E. A POSITA Would Not Have Been Motivated to Combine Cox and Ross in the Manner Tesla and Dr. Janét Allege

146. In my opinion, a POSITA would not have been motivated to combine Cox and Ross in the manner Tesla and Dr. Janét suggest.

147. Tesla and Dr. Janét allege the following:

A POSITA would have been motivated to combine Cox and Ross such that

- Cox’s system transmits a training image to multiple human annotators for annotation with a prompt to answer about an action by a person or road user in the image reflecting a possible intent of the person, as taught by Ross; and
- Cox’s system is trained to analyze an action by a person or road user, reflecting a possible intent in images, as taught by Ross.

Pet. at 36; Ex. 1003, ¶ 125.

148. Nowhere do Tesla or Dr. Janét describe how a POSITA would have accomplished either of these proposed modifications of Cox in view of Ross. Dr. Janét’s “analysis” of the alleged motivation to combine Cox and Ross spans only a few conclusory paragraphs with no supporting analysis.

149. In my opinion, a POSITA would not have been motivated to combine Cox and Ross. Ross relates to autonomous vehicle intent prediction but uses a fundamentally different approach than Cox. Ross does not use crowd-sourced human annotations to train its models like Cox does. Instead, Ross’s training data is based on observations of actual road user behavior over time. *See, e.g.*, Ex. 1005, 4:25-32, 4:51-65. In my opinion, this fundamentally different approach means that a POSITA would not have been motivated to combine Cox and Ross in the manner Tesla and Dr. Janét suggest.

150. Tesla and Dr. Janét’s alleged motivation to combine is essentially that both Cox and Ross relate to “human-like” predictions, but this is too broad. Nearly every machine learning system (and especially those that are intended for use in

autonomous vehicles) including supervised machine learning systems as disclosed in Cox aspire to “human-like” performance. *See* Ex. 2044 at p. VII (“Neural networks were developed to simulate the human nervous system for machine learning tasks by treating the computational units in a learning model in a manner similar to human neurons. The grand vision of neural networks is to create artificial intelligence by building machines whose architecture simulates the computations in the human nervous system.”). In essence, supervised machine learning – as described and primarily utilized in Cox for face classification - is a type of machine learning in which a model is trained using specifically labeled data, where each input is paired with a known and appropriately labeled output or target value. The model learns the relationship between the training inputs and outputs by minimizing prediction errors during training via the use of properly designed loss functions, enabling it to generalize and make accurate predictions on new, unseen data for classification purposes as disclosed in Cox. Hence, the pairing of the training data (input-output pair) as well as the design of the network architecture (i.e. the number of inputs and the input layer, the number of hidden layers, the number of nodes per layer, the type of activation function, and the type of output produced by the network, etc.), the regularization methodology and the selection of the loss function are generally unique to the application at hand.

151. Hence, the specific combination Tesla and Dr. Janét propose, i.e., using Cox’s crowd-sourced psychometric annotation framework with Ross’s autonomous vehicle intent prediction, requires hindsight reconstruction of the claimed invention. Cox is generally concerned with improving the ability of a machine learning model to predict certain classifications regarding human faces—primarily whether a face is or is not present, but also tangentially with other questions such as the gender of the subject in the image or an emotion conveyed by a face in the image. Ex. 1007, [0016], [0048], [0049]. In contrast, Ross is generally concerned with observing the behavior of road users over time (including past the relevant point in time to make a decision about what the road user will actually do) and using that information to train a machine learning model to predict what such road users will do. *See, e.g.*, Ex. 1005, 4:25-32, 4:51-65. Hence, Ross’s system is not concerned or designed for classification purposes as described in Cox but rather for predicting intents and behaviors of road users – a different application altogether that requires a different design of a machine learning system with different training/validation/test datasets.

152. These two technological issues are fundamentally different, and in my opinion, a POSITA looking at Cox’s disclosure would not look to Ross’s disclosure to address either issue Tesla and Dr. Janét identify.

153. On the first of the two issues (i.e., “transmit[ting] a training image to multiple human annotators for annotation with a prompt to answer about an action

by a person or road user in the image reflecting a possible intent of the person”), because Ross is not directed to transmitting images to crowdsourced human annotators, a POSITA would not have been motivated to combine Cox with Ross to arrive at this modification.¹ Nothing in the Petition or in Dr. Janét’s discussion of the alleged motivation to combine points to any teaching in Ross that suggests that the combination of Cox with Ross would address this. Also, based on my review of Ross (independent of any reasoning in the Petition or in Dr. Janét’s declaration), nothing in Ross would motivate a POSITA to modify Cox in view of Ross in this manner.

154. On the second issue (i.e., training Cox’s system “to analyze an action by a person reflecting a possible intent in images, as taught by Ross”), here too, a POSITA would not have been motivated to combine Cox with Ross to arrive at this

¹ I note that Ross does describe a process where “the next action included in the predetermined list of actions that the road user ends up performing is marked as the correct intent.” Ex. 1005, 4:56-58. It is unclear whether this marking is performed by human annotators or by the system. Either way, it does not equate to crowdsourced human annotation of images (such as in Cox), where humans are presented with images and asked to annotate the image with their own belief or judgment as to what is shown.

modification. As discussed above, Cox is generally concerned with improving the ability of a machine learning model to predict certain classifications regarding human faces—primarily whether a face is or is not present, but also tangentially other questions such as the gender of the subject in the image or an emotion conveyed by a face in the image. Ex. 1007, [0016], [0048], [0049]. In contrast, Ross is generally concerned with observing the behavior of road users over time (including past the relevant point in time to make a decision about what the road user will actually do) and using that information to train a machine learning model to predict what such road users will do. *See, e.g.*, Ex. 1005, 4:25-32, 4:51-65. Given these fundamental differences, a POSITA would not have been motivated to modify Cox’s human-annotation driven face-recognition/classification system to incorporate Ross’s technologically distinct mechanism of using road user behavior over time to train a model to predict road user behavior.

155. In my opinion, Cox and Ross are also directed to different endeavors within the field of machine learning. Cox is generally concerned with improving machine learning models for analyzing the presence/non-presence (and characteristics of) a potential face in an image. Ex. 1007, [0002], [0013], [0014], [0016], [0017], [0033], [0039]-[0050]. In contrast, Ross is generally concerned with improving machine learning models for predicting the intent of a road user. Ex. 1005 at, e.g., Figure 10, 1:30-3:17. In my opinion, given the particular challenges of these

problems within the field of machine learning, a POSITA would not have been motivated to look to the disclosures of Ross to combine them with those of Cox.

156. Research in the field of intelligent and autonomous vehicle systems also confirms the unique nature of the problems faced by attempting to analyze image, video, or other sensor data in real time driving situations. For example, in 2014 Kooij et al. noted that “pedestrian path prediction is a challenging problem, due to the highly dynamic nature of pedestrian motion, and systems need to react with limited computation time.” Ex. 2042 at 1. And, in 2017 Zyner et al. noted that “the ability to infer the future intentions of drivers to predict the likelihood of potential collisions” “is a challenging task, particularly in intersections where complex traffic scenarios result in a proportionally high number of accidents.” Ex. 2043 at 1. These examples of other researchers in the field of intelligent and autonomous vehicle systems noting the unique challenges of autonomous vehicle operating environments only further reflects that a POSITA would not have been motivated to look to disclosures in that field to combine them with Cox’s crowdsourced annotation system for improving face detection, which does not present the same unique challenges.

157. Therefore, in my opinion, a POSITA would not have been motivated to combine Cox and Ross in the manner Tesla and Dr. Janét suggest.

X. GROUND 3

158. I understand that Tesla alleges (and Dr. Janét opines) that dependent claims 5, 12, and 19 of the '046 patent are obvious in view of Cox and Ellenbogen. *See* Pet. 46-49; Ex. 1003, ¶¶ 151-159. For the reasons I discuss below, I disagree. In my opinion, none of claims 5, 12, and 19 are obvious in view of Cox and Ellenbogen.

159. I understand from the Petition and Dr. Janét's declaration that Ground 3 challenges dependent claims 5, 12, and 19 as being obvious over Cox in view of Ellenbogen. Neither Tesla nor Dr. Janét allege in this ground that Ellenbogen teaches any elements of the independent claims. *See* Pet. 46-49; Ex. 1003, ¶¶ 151-159.

160. As I discuss above with respect to Ground 1, Cox does not teach several elements of the challenged claims.

161. I have considered Tesla and Dr. Janét's allegations with respect to Ground 3, and none of those allegations alter my opinions expressed above with respect to the elements of the challenged claims.

162. In my opinion, therefore, dependent claims 5, 12, and 19 of the '046 patent are not obvious in view of Cox and Ellenbogen because Cox does not teach all of the elements of independent claims 1, 8, and 15 (from which each of those claims respectively depend) for the same reasons as discussed above for Ground 1.

XI. GROUND 4

163. I understand that Tesla alleges (and Dr. Janét opines) that claims 1-19 of the '046 patent are obvious in view of Cox and Munro. *See* Pet. 49-54; Ex. 1003, ¶¶ 160-176. For the reasons I discuss below, I disagree. In my opinion, none of claims 1-19 are obvious in view of Cox and Munro.

A. Cox and Munro

164. In arriving at my opinions concerning Tesla and Dr. Janét's assertions regarding claims 1-19 of the '046 patent and Cox and Munro, I have considered both Cox and Munro's disclosures from the perspective of a POSITA. In my discussion of Ground 1 above, I discussed aspects of Cox that are relevant to my opinions. Below I discuss some aspects of Munro that are relevant to my opinions.

165. In marked contrast to Cox and the '046 patent, Munro is generally directed to natural language processing. Thus, instead of being concerned with analyzing the contents of an image for a given purpose (i.e., a face present/face not present determination as described in Cox, or determining the intent of a road user in an image as described in the '046 patent), Munro is directed to improving processes for analyzing the content of documents. *Compare, e.g.*, Ex. 1007, [0007]-[0010] *and* Ex. 1001, 1:27-61 *with, e.g.*, Ex. 1006, Abstract, [0003], [0004].

166. Munro proposes “[m]ethods for generating natural language processing systems.” Ex. 1006, Title. To this effect, Munro explains that its invention

encompasses “[m]ethods [that] are presented for generating a natural language model. The method may comprise: ingesting training data representative of documents to be analyzed by the natural language model, generating a hierarchical data structure comprising at least two topical nodes within which the training data is to be subdivided into by the natural language model, selecting a plurality of documents among the training data to be annotated, generating an annotation prompt for each document configured to elicit an annotation about said document indicating which node among the at least two topical nodes said document is to be classified into, receiving the annotation based on the annotation prompt; and generating the natural language model using an adaptive machine learning process configured to determine patterns among the annotations for how the documents in the training data are to be subdivided according to the at least two topical nodes of the hierarchical data structure “ Ex. 1006, Abstract.

167. As described above, Munro is clearly directed to the generation of natural language processing systems, the analysis of documents in hierarchical structures, the annotation of documents, and the generation of the natural language model that is strictly designed to handle documents. As I describe in more detail below, machine learning for natural language processing and for object recognition in images use different types of data and use different techniques to analyze and process that data.

B. Cox and Munro Do Not Teach Elements 1[b-iii], 8[b-iii], or 15[b-iii] (“generating summary statistics describing the state of mind of the user in the image based on the received responses from the plurality of human observers”)

168. In my opinion, the combination of Cox and Munro (as Tesla and Dr. Janét propose) does not teach elements 1[b-iii], 8[b-iii], or 15[b-iii], which each require “generating summary statistics describing the state of mind of the user in the image based on the received responses from the plurality of human observers.”

169. Tesla and Dr. Janét allege that “[t]he combination of Cox and Munro teaches that Cox’s generated statistical data from responses of human annotators regarding the emotional state of a user in a training image would have included tallies to aggregate the number of annotators who similarly classified the emotional state of the user in the training image.” Pet. at 52-53; Ex. 1003, ¶ 169.

170. To the extent that Tesla and Dr. Janét suggest that Cox’s statistical data includes such tallies, I disagree with Tesla and Dr. Janét. Instead, as best I understand Tesla and Dr. Janét’s allegations, they propose to combine Cox with Munro such that Cox’s “statistical data” includes the tallies that Munro teaches. See Pet. at 50; Ex. 1003, ¶ 163. In my opinion, there are multiple shortcomings with this theory.

171. First, neither Tesla nor Dr. Janét describe what “statistical data” they are referring to from Cox. Instead, Tesla and Dr. Janét just repeatedly refer to “Cox’s generated statistical data.” Pet. at 50, 52-53; Ex. 1003, ¶¶ 163, 164, 169. To the extent this refers to Tesla and Dr. Janét’s allegations (as in Ground 1 and elsewhere)

that Cox’s psychometric data, item-response curves, human-weighted loss function, and/or end-of-test statistics somehow constitute (all or part of) the claimed “summary statistics describing the state of mind of the user in the image based on the received responses from the plurality of human observers,” I disagree for the reasons I have stated above with respect to Ground 1.

172. Second, even with this understanding, I disagree with Tesla and Dr. Janét that the combination of Cox and Munro teaches elements 1[b-iii], 8[b-iii], or 15[b-iii]. Munro expressly indicates that the purpose of the aggregation it describes is to “analyze[] all of the annotations made to a single document to determine whether said document may be reliably annotated for use in training the natural language model.” Ex. 1006, [0140]. Below is Munro’s discussion of this aggregation:

At block 1220, in the event that multiple annotators supply annotations to the same document, in some embodiments, the natural language platform may be configured to conduct an annotation aggregation process that analyzes all of the annotations made to a single document to determine whether said document may be reliably annotated for use in training the natural language model. For example, if six annotators provided annotations to the same document, and three annotators classified the document as belonging to a first label, while the other three annotators classified the document as belonging to a second label, then the annotation aggregation process of block 1220 may determine that this documents is too ambiguous to be used in the adaptive machine

learning training process. As another example, if five of the six annotators provided a consistent annotation to the document, while a sixth annotator provided a different annotation, it may be determined that this document has sufficiently consistent annotations so as to be used in the adaptive machine learning process.

Id.

173. Nothing in Munro suggests that its tallies “describ[e] the state of mind of the user in the image,” as required by the challenged claims. And, nothing in Cox suggests (nor do Tesla or Dr. Janét even allege) the desirability or need for indicating whether the images in the training data can be reliably annotated for use in training the model, which is the purpose of these tallies in Munro’s natural language model.

174. Machine learning for Natural Language Processing (NLP) and object recognition in images (such as face/no face or autonomous driving applications) both aim to identify patterns in data, but they operate very differently on distinctively different types of data. NLP models often use techniques such as transformers and language embeddings to capture semantic relationships and focus on understanding, interpreting, and/or generating human language and human language models, where data (words and sentences) is sequential and context-dependent. For instance, the sentences “I swam across the river to get to the other bank” versus “I walked across the road to get cash from the bank” convey different distinct meanings and interpretations about the word “bank” which can only be determined by looking at

the context. The word “bank” in the first sentence unambiguously points to the bank of the river versus in the second sentence, it unambiguously points to a financial institution. This is determined from the context surrounding the sentence where “swam” and “river” indicate that “bank” refers to the side of a river whereas in the second sentence, the word “cash” unambiguously indicates that “bank” is a financial institution. In addition, the sequencing of the words in sentences is extremely important to the meaning of the sentence. For example, “the food is bad, not good at all” versus “the food is good, not bad at all” convey completely different distinct meaning without ambiguity about the food. Finally, the dataset is often functioning like a textbook answer key. Each document is presented to the model along with the “correct” answer. If two identical or nearly identical documents receive different labels (i.e. are annotated differently by human observers), the model is essentially being given contradictory instructions. For this reason, dataset curators frequently try to identify and remove inconsistently labeled documents (i.e. inconsistent annotations), because the disagreement is often viewed as evidence that somebody made a mistake or that the labeling guidelines were not followed consistently

175. In contrast, object recognition processes visual data, identifying and classifying objects within images based on spatial patterns, shapes, colors, and textures. While NLP models often use techniques such as transformers and language embeddings to capture semantic relationships, object recognition models typically

employ convolutional neural networks (CNNs) to learn visual features that are not sequential as one would expect with text sentences as described earlier.

176. The primary challenge in NLP is understanding meaning and context using the sequential ordering of the words, whereas object recognition must handle and address multiple variations in appearance, scale, lighting, positioning within the frame and viewpoint. Hence, systems designed for autonomous applications operate entirely differently and in different environments from NLP systems. They are not learning from a world with sequential style data (as one would expect with sentences) that is context driven as described earlier. Instead, they are learning from a world filled with uncertainty, incomplete information, and human behavior that cannot always be predicted with certainty. For instance, consider a frame of video with a pedestrian standing at the edge of a crosswalk. One annotator may look at the image and conclude that the pedestrian is preparing to cross because the person's body is oriented toward the roadway and their attention appears focused on traffic. Another annotator may conclude that the pedestrian is simply waiting, and does not intend to cross at this time but rather looking for someone across the street. Neither annotator is necessarily wrong. The image itself may not contain enough information to determine with certainty what the pedestrian will do next.

177. Importantly though, both of these scenarios are applicable in autonomous driving conditions. If an autonomous-vehicle developer were to discard

every example on which annotators disagreed, many of the most important real-world situations that could raise potential ambiguity would disappear from the training data. The remaining dataset would be populated primarily by easy cases, such as pedestrians who are obviously crossing, vehicles that are clearly turning, and cyclists who are unmistakably merging. The model would become very good at recognizing obvious situations while learning much less about the ambiguous situations that actually create risk on the road and must be dealt with on a regular basis while driving. From that perspective, inconsistent labels may provide evidence that a scene admits more than one reasonable interpretation.

178. Rather than treating the disagreement as error to be eliminated, autonomous machine learning designers may use it to estimate uncertainty, build probabilistic models, and/or identify scenarios that deserve special attention during training. Hence, the underlying philosophy is therefore different. In many natural-language classification tasks, inconsistent labels often suggest that the dataset is teaching the model conflicting rules as illustrated earlier with the bank and food examples. On the other hand, in autonomous-driving prediction tasks, inconsistent labels may instead reveal that the presented scenarios often raise ambiguity and hence a competent model should learn to reason about that uncertainty rather than ignore it.

179. The passage I quoted above from Munro relates to precisely this issue, as it specifically makes clear that in NLP systems like Munro's, if annotators disagree on the classification of a document, the document should be excluded from the training data. In contrast, nothing in Cox suggests excluding data from the training dataset because annotators disagree.

180. I have considered the portions of Cox and Munro (and cross-referenced sections of the Petition and Dr. Janét's declaration) identified by Tesla and Dr. Janét as part of their allegations with respect to this claim element. *See* Pet. at 52-53; Ex. 1003, ¶¶ 169-170 (both citing "EX1006, Abstract, ¶¶0005, 0042-0048, 0132-0142; EX1007, ¶¶0002-0006, 0014, 0035, 0048" and "Sections VI.A.2.f (element 1[b-iii]), VI.D.2"). None of those portions of Cox and Munro, nor any of the cross-referenced sections of the Petition or Dr. Janét's declaration, alter my opinions with respect to these claim elements.

181. The portions of the Petition and Dr. Janét's declaration discussed above relate to claim 1 of the '046 patent. Both the Petition and Dr. Janét's declaration merely refer back to these discussions when discussing element 8[b-iii] and 15[b-iii]. *See* Pet. at 54; Ex. 1003, ¶¶ 174, 175. In my opinion, Cox fails to teach or suggest elements 8[b-iii] and 15[b-iii] for the same reasons as discussed above.

182. Therefore, in my opinion, Cox and Munro do not teach or suggest elements 1[b-iii], 8[b-iii], or 15[b-iii] ("generating summary statistics describing the

state of mind of the user in the image based on the received responses from the plurality of human observers”).

C. Cox and Munro Do Not Teach Elements 1[b-iv], 8[b-iv], or 15[b-iv] (“storing the summary statistics in association with the image as part of the training data”)

183. In my opinion, the combination of Cox and Munro (as Tesla and Dr. Janét propose) does not teach elements 1[b-iv], 8[b-iv], or 15[b-iv], which each require “storing the summary statistics in association with the image as part of the training data.”

184. Tesla and Dr. Janét allege the following with respect to this element:

Cox and Munro teach element 1[b-iv] for at least the same reasons discussed in Section VI.A.2.g (element 1[b-iv]), except the generated summary statistics as taught by Munro would have been stored in Cox’s computer record in a database for the statistic’s associated image.

Pet at 53-54; Ex. 1003, ¶ 171.

185. Neither Tesla nor Dr. Janét make clear what “generated summary statistics as taught by Munro” they are referring to, but as best I understand their allegations (including based on their cross-reference to their discussion of element 1[b-iii]), Tesla and Dr. Janét propose to modify Cox in view of Munro to include “tallies to aggregate the number of annotators who similarly classified the emotional state of the user in the training image” (Pet. 53; Ex. 1003, ¶ 169) and to store these tallies in Cox’s “database of training objects.” (Pet. at 23; Ex. 1003, ¶ 80).

186. As Tesla and Dr. Janét discuss, these tallies are part of Munro’s performance of “an ‘aggregation process’ that involves determining how many human annotators classified the stimulus as belonging to a first label and how many annotators classified the stimulus as belonging to a second label.” Pet. at 49; Ex. 1003, ¶ 161. As Munro makes clear though, the purpose of that aggregation process is to “analyze[] all of the annotations made to a single document to determine whether said document may be reliably annotated for use in training the natural language model.” *Id.* Nothing in Munro teaches storing the result of that aggregation process (or any tally generated as part of or as a result of that process) in any database. Instead, the purpose of that aggregation process in Munro is to determine if the document should be in the dataset for training the natural language model. *Id.*

187. Moreover, as I discuss earlier, machine learning for NLP and for object recognition in images use different types of data and use different techniques to analyze and process that data, and a POSITA would not have used Munro’s aggregation process (which aims to exclude data from the training data when annotators differ in their categorization of a document) in machine learning algorithms/processes for object recognition in images.

188. I have considered the portions of the Petition and Dr. Janét’s declaration cross-referenced as part of their allegations with respect to this claim element. *See* Pet. at 53-54; Ex. 1003, ¶ 171. None of the cross-referenced sections of the Petition

or Dr. Janét's declaration (or the portions of Cox and/or Munro cited therein) alter my opinions with respect to these claim elements.

189. The portions of the Petition and Dr. Janét's declaration discussed above relate to claim 1 of the '046 patent. Both the Petition and Dr. Janét's declaration merely refer back to these discussions when discussing element 8[b-iii] and 15[b-iii]. *See* Pet. at 54; Ex. 1003, ¶¶ 174, 175. In my opinion, Cox fails to teach or suggest elements 8[b-iii] and 15[b-iii] for the same reasons as discussed above.

190. Therefore, in my opinion, Cox and Munro do not teach or suggest elements 1[b-iv], 8[b-iv], or 15[b-iv] ("storing the summary statistics in association with the image as part of the training data").

D. Cox and Munro Do Not Teach Elements 1[c], 8[c], or 15[c] ("the model configured to receive an input image showing a user and predict summary statistics describing a state of mind of the user in the input image")

191. In my opinion, the combination of Cox and Munro (as Tesla and Dr. Janét propose) does not teach elements 1[c], 8[c], or 15[c], which each require "the model configured to receive an input image showing a user and predict summary statistics describing a state of mind of the user in the input image."

192. Tesla and Dr. Janét allege the following with respect to this element:

As combined, the training of Cox's model would have been performed as discussed in [Ground 1, element 1[c]] except the training data includes the generated summary statistics as taught by Munro, and

Cox's model is trained on such training data in the same manner discussed in [Ground 1, (element 1[c])].

Pet at 54; Ex. 1003, ¶ 172.

193. Neither Tesla nor Dr. Janét make clear what “generated summary statistics as taught by Munro” they are referring to, but as best I understand their allegations (including based on their cross-reference to their discussion of element 1[b-iii]), Tesla and Dr. Janét propose to modify Cox in view of Munro to include “tallies to aggregate the number of annotators who similarly classified the emotional state of the user in the training image” (Pet. at 53; Ex. 1003, ¶ 169) and to store these tallies in Cox's training data (Pet. at 54; Ex. 1003, ¶ 172).

194. As Tesla and Dr. Janét discuss, these tallies are part of Munro's performance of “an ‘aggregation process’ that involves determining how many human annotators classified the stimulus as belonging to a first label and how many annotators classified the stimulus as belonging to a second label.” Pet. at 49; Ex. 1003, ¶ 161. As Munro makes clear though, the purpose of that aggregation process is to “analyze[] all of the annotations made to a single document to determine whether said document may be reliably annotated for use in training the natural language model.” *Id.* Nothing in Munro teaches storing the result of that aggregation process (or any tally generated as part of or as a result of that process) in any database. Moreover, nothing in Cox or Munro teaches that a model is configured to predict such tallies. Instead, the purpose of that aggregation process in Munro is to

determine if the document should be in the dataset for training the natural language model. *Id.* Moreover, as I discuss earlier, machine learning for NLP and for object recognition in images use different types of data and use different techniques to analyze and process that data, and a POSITA would not have used Munro's aggregation process (which aims to exclude data from the training data when annotators differ in their categorization of a document) in machine learning algorithms/processes for object recognition in images.

195. I have considered the portions of the Petition and Dr. Janét's declaration cross-referenced as part of their allegations with respect to this claim element. *See* Pet. at 54; Ex. 1003, ¶ 172. None of the cross-referenced sections of the Petition or Dr. Janét's declaration (or the portions of Cox and/or Munro cited therein) alter my opinions with respect to these claim elements.

196. The portions of the Petition and Dr. Janét's declaration discussed above relate to claim 1 of the '046 patent. Both the Petition and Dr. Janét's declaration merely refer back to these discussions when discussing element 8[b-iii] and 15[b-iii]. *See* Pet. at 54; Ex. 1003, ¶¶ 174, 175. In my opinion, Cox fails to teach or suggest elements 8[c] and 15[c] for the same reasons as discussed above.

197. Therefore, in my opinion, Cox and Munro do not teach or suggest elements 1[c], 8[c], or 15[c] ("the model configured to receive an input image

showing a user and predict summary statistics describing a state of mind of the user in the input image”).

E. A POSITA Would Not Have Been Motivated to Combine Cox and Munro in the Manner Tesla and Dr. Janét Allege

198. As I discuss in Sections XI.A and XI.B above, machine learning for NLP and for object recognition in images use different types of data and use different techniques to analyze and process that data. As a result, in my opinion, a POSITA would not have been motivated to combine Munro’s aggregation process (which aims to exclude data from the training data when annotators differ in their categorization of a document) in machine learning algorithms/processes for object recognition in images such as Cox for the same reasons.

XII. GROUND 5

199. I understand that Tesla alleges (and Dr. Janét opines) that claims 1-19 of the ’046 patent are obvious in view of Ellenbogen and Munro. *See* Pet. 55-76; Ex. 1003, ¶¶ 177-240. For the reasons I discuss below, I disagree. In my opinion, none of claims 1-19 are obvious in view of Ellenbogen and Munro.

A. Ellenbogen and Munro

200. In arriving at my opinions concerning Tesla and Dr. Janét’s assertions regarding claims 1-19 of the ’046 patent and Ellenbogen and Munro, I have considered both Ellenbogen and Munro’s disclosures from the perspective of a POSITA. In my discussion of Ground 4 above, I discussed aspects of Munro that are

relevant to my opinions. Below I discuss some aspects of Ellenbogen that are relevant to my opinions.

201. Ellenbogen is concerned with the design of a “[p]latform for gathering real-time analysis.” Ex. 1004, Title. Ellenbogen explains that its platform receives data “characterizing a request for agent computation of sensor data. The request includes a required confidence and required latency for completion of the agent computation. Agents to query are determined based on the required confidence. Data is transmitted to query the determined agents to provide analysis of the sensor data.” *Id.*, Abstract. Fig. 23 of Ellenbogen provides a flow diagram illustrating an example method for managing agent queries to ensure confidence in a result provided by an agent computation component. *Id.*, [0062]. Fig. 34 provides a flow diagram illustrating a method of training a machine computation component on the result provided from an agent computation component. *Id.*, [0133]. “[T]he agent computation component can execute the task by querying one or more agents in a pool of agents to perform the task, such as an image recognition task, answering a question regarding a characteristic of the image, and the like.” Ex. 1004, [0135]. “The agent computation component can query multiple agents and create a composite output and a composite confidence. The confidence measure can characterize a likelihood that the output of the agent computation component is correct.” *Id.* Ellenbogen explains that “a predictive model of a machine computation

component can be trained by providing the sensor data as input and the result from the agent computation component as a supervisory signal.” Ex. 1004, [0136].

B. Ellenbogen and Munro Do Not Teach Elements 1[b-iii], 8[b-iii], or 15[b-iii] (“generating summary statistics describing the state of mind of the user in the image based on the received responses from the plurality of human observers”)

202. In my opinion, the combination of Ellenbogen and Munro (as Tesla and Dr. Janét propose) does not teach elements 1[b-iii], 8[b-iii], or 15[b-iii], which each require “generating summary statistics describing the state of mind of the user in the image based on the received responses from the plurality of human observers.”

203. Tesla and Dr. Janét allege that, when combined with Ellenbogen, “Munro teaches that Ellenbogen’s composite output of query result responses regarding the behavior or activity of a person in a sample image would have included an aggregate of the number of annotators who classified the behavior or activity as having various labels in their responses.” Pet. at 64; Ex. 1003, ¶ 194.

204. As discussed above in Ground 4, this aggregate number of responses (i.e., the tallies that Tesla and Dr. Janét refer to) is performed as part of an aggregation process that is designed to “analyze[] all of the annotations made to a single document to determine whether said document may be reliably annotated for use in training the natural language model.” Ex. 1006, [0140]. Nothing about these tallies “describe[es] the state of mind of the user in the image” as required by the challenged claims. And, nothing in Ellenbogen suggests (nor do Tesla or Dr. Janét

even allege) the desirability or need for indicating whether the images in the training data can be reliably annotated for use in training the model, which is the purpose of these tallies in Munro's natural language model. Moreover, as I discuss earlier, machine learning for NLP and for object recognition in images use different types of data and use different techniques to analyze and process that data, and a POSITA would not have used Munro's aggregation process (which aims to exclude data from the training data when annotators differ in their categorization of a document) in machine learning algorithms/processes for object recognition in images.

205. I have considered the portions of Ellenbogen and Munro (and cross-referenced sections of the Petition and Dr. Janét's declaration) identified by Tesla and Dr. Janét as part of their allegations with respect to this claim element. *See* Pet. at 64-65; Ex. 1003, ¶¶ 192-195. None of those portions of Ellenbogen and Munro, nor any of the cross-referenced sections of the Petition or Dr. Janét's declaration, alter my opinions with respect to these claim elements.

206. The portions of the Petition and Dr. Janét's declaration discussed above relate to claim 1 of the '046 patent. Both the Petition and Dr. Janét's declaration merely refer back to these discussions when discussing element 8[b-iii] and 15[b-iii]. *See* Pet. at 74-75; Ex. 1003, ¶¶ 217, 233. In my opinion, Ellenbogen fails to teach or suggest elements 8[b-iii] and 15[b-iii] for the same reasons as discussed above.

207. Therefore, in my opinion, Ellenbogen and Munro do not teach or suggest elements 1[b-iii], 8[b-iii], or 15[b-iii] (“generating summary statistics describing the state of mind of the user in the image based on the received responses from the plurality of human observers”).

C. Ellenbogen and Munro Do Not Teach Elements 1[b-iv], 8[b-iv], or 15[b-iv] (“storing the summary statistics in association with the image as part of the training data”)

208. In my opinion, the combination of Ellenbogen and Munro (as Tesla and Dr. Janét propose) does not teach elements 1[b-iv], 8[b-iv], or 15[b-iv], which each require “storing the summary statistics in association with the image as part of the training data.”

209. Tesla and Dr. Janét allege the following with respect to this element:

Munro teaches to store the Ellenbogen-Munro composite output in association with the image as part of the training data. EX1003, ¶197. Specifically, Munro’s system includes a “client data store” (i.e., a database of training data) that contains “a repository of data that is used to train and ultimately generate the natural language model.” EX1006, ¶0061.

Pet at 66; Ex. 1003, ¶ 197.

210. As discussed above, the “Ellenbogen-Munro composite output” Tesla and Dr. Janét refer to is the hypothetical combination of “Ellenbogen’s composite output of query result responses regarding the behavior or activity of a person in a sample image” and “an aggregate of the number of annotators who classified the

behavior or activity as having various labels in their responses.” Pet. at 64; Ex. 1003, ¶ 194.

211. As Tesla and Dr. Janét discuss, these tallies are part of Munro’s performance of “an ‘aggregation process’ that involves determining how many human annotators classified the stimulus as belonging to a first label and how many annotators classified the stimulus as belonging to a second label.” Pet. at 49; Ex. 1003, ¶ 161. As Munro makes clear though, the purpose of that aggregation process is to “analyze[] all of the annotations made to a single document to determine whether said document may be reliably annotated for use in training the natural language model.” Ex. 1006, [0140]. Nothing in Munro teaches storing the result of that aggregation process (or any tally generated as part of or as a result of that process) in any database, let alone “in association with the image as part of the training data.” Instead, the purpose of that aggregation process in Munro is to determine if the document should be in the dataset for training the natural language model. *Id.* Moreover, as I discuss earlier, machine learning for NLP and for object recognition in images use different types of data and use different techniques to analyze and process that data, and a POSITA would not have used Munro’s aggregation process (which aims to exclude data from the training data when annotators differ in their categorization of a document) in machine learning algorithms/processes for object recognition in images.

212. I have considered the portions of Ellenbogen and Munro cited (and the portions of the Petition and Dr. Janét’s declaration cross-referenced) as part of their allegations with respect to this claim element. *See* Pet. at 65-67; Ex. 1003, ¶¶ 196-199. None of the portions of Ellenbogen and/or Munro cited therein or the cross-referenced sections of the Petition or Dr. Janét’s declaration alter my opinions with respect to these claim elements.

213. The portions of the Petition and Dr. Janét’s declaration discussed above relate to claim 1 of the ’046 patent. Both the Petition and Dr. Janét’s declaration merely refer back to these discussions when discussing element 8[b-iv] and 15[b-iv]. *See* Pet. at 74-75; Ex. 1003, ¶¶ 218, 234. In my opinion, Ellenbogen fails to teach or suggest elements 8[b-iv] and 15[b-iv] for the same reasons as discussed above.

214. Therefore, in my opinion, Ellenbogen and Munro do not teach or suggest elements 1[b-iv], 8[b-iv], or 15[b-iv] (“storing the summary statistics in association with the image as part of the training data”).

D. Ellenbogen and Munro Do Not Teach Elements 1[c], 8[c], or 15[c] (“training a model using the training data, the model configured to receive an input image showing a user and predict summary statistics describing a state of mind of the user in the input image”)

215. In my opinion, the combination of Ellenbogen and Munro (as Tesla and Dr. Janét propose) does not teach elements 1[c], 8[c], or 15[c], which each require “the model configured to receive an input image showing a user and predict summary statistics describing a state of mind of the user in the input image.”

216. As I discuss above, in their theory for elements 1[b-iii]/[b-iv], 8[b-iii]/[b-iv], and 15[b-iii]/[b-iv], Tesla and Dr. Janét allege that “Munro teaches that Ellenbogen’s composite output of query result responses regarding the behavior or activity of a person in a sample image would have included an aggregate of the number of annotators who classified the behavior or activity as having various labels in their responses” (which Tesla and Dr. Janét call the “Ellenbogen-Munro composite output”) constitutes the claimed “summary statistics.” Pet. at 64; Ex. 1003, ¶ 194. Here, however, Tesla and Dr. Janét allege that a different feature constitutes the claimed “predict[ing] summary statistics”—specifically, “predicting a multiclass behavior or activity prediction of a person in a training image and associated confidence scores.” Pet. at 69.

217. In my opinion, Tesla and Dr. Janét’s theory is contrary to the plain and ordinary meaning of the language of the challenged claims to a POSITA. A POSITA reading the challenged claims would read the “summary statistics” in this element as referring to the same “summary statistics” in elements 1[b-iii]/[b-iv], 8[b-iii]/[b-iv], and 15[b-iii]/[b-iv]. The “[b]” elements of the challenged claims are part of “generating training data,” and the “[c]” elements recite “training a model using the training data.” A POSITA would not interpret the challenged claims to recite generating one set of summary statistics in the “[b]” elements but having a model be configured to predict a different set of summary statistics. Thus, Tesla and Dr.

Janét's effort to read the "summary statistics" here onto a different feature of Ellenbogen defies the plain and ordinary meaning of the claim language.

218. In addition, even ignoring Tesla and Dr. Janét's improper interpretation of the claims, the "multiclass behavior or activity prediction" Tesla and Dr. Janét point to does not meet the claims' requirement that "the model [be] configured to receive an input image showing a user and predict summary statistics describing a state of mind of the user in the input image." The portion of Ellenbogen that Tesla and Dr. Janét rely on merely indicates that "[t]he object classifier can generate a score from 0.0 to 1.0 related to the confidence that the object is not present (0.0) or present (1.0)" and that "[t]he classification can be binary or multiclass." *See* Ex. 1004, [0166] (cited both at Pet. at 68 and Ex. 1003, ¶ 202). Generating such object present/not present classifications (even with associated 0.0-1.0 confidence values) is not "predict[ing] summary statistics describing a state of mind of the user in the input image," as those object present/not present classifications do not "describ[e] the state of mind of the user in the input image."

219. I have considered the portions of Ellenbogen and Munro cited (and the portions of the Petition and Dr. Janét's declaration cross-referenced) as part of their allegations with respect to this claim element. *See* Pet. at 67-70; Ex. 1003, ¶¶ 200-204. None of the portions of Ellenbogen and/or Munro cited therein or the cross-

referenced sections of the Petition or Dr. Janét's declaration alter my opinions with respect to these claim elements.

220. The portions of the Petition and Dr. Janét's declaration discussed above relate to claim 1 of the '046 patent. Both the Petition and Dr. Janét's declaration merely refer back to these discussions when discussing element 8[c] and 15[c]. *See* Pet. at 74-76; Ex. 1003, ¶¶ 219, 235. In my opinion, Ellenbogen fails to teach or suggest elements 8[c] and 15[c] for the same reasons as discussed above.

221. Therefore, in my opinion, Ellenbogen and Munro do not teach or suggest elements 1[c], 8[c], or 15[c] ("the model configured to receive an input image showing a user and predict summary statistics describing a state of mind of the user in the input image").

E. A POSITA Would Not Have Been Motivated to Combine Ellenbogen and Munro in the Manner Tesla and Dr. Janét Allege

222. As I discuss earlier, machine learning for NLP and for object recognition in images use different types of data and use different techniques to analyze and process that data. Given these differences, in my opinion, a POSITA would not have been motivated to combine Munro's aggregation process (which aims to exclude data from the training data when annotators differ in their categorization of a document) in machine learning algorithms/processes for object recognition in images such as Ellenbogen for the same reasons.

XIII. CONCLUSION

223. This declaration represents the opinions that I have formed up to date based on the material currently available to me. I reserve the right to revise, supplement, and/or amend my opinions stated in this declaration based on new information that becomes available to me and on my continuing analysis of the materials available to me.

224. I declare that all statements made herein of my knowledge are true, and that all statements made on information and belief are believed to be true, and that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine, imprisonment, or both, under Section 1001 of Title 18 of the United States Code.

Dated: June 15, 2026



Dr. Eli Saber, Ph.D.