

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

TESLA, INC.,

Petitioner,

v.

PERCEPTIVE AUTOMATA LLC,

Patent Owner.

Case No. IPR2025-01574

Patent No. 11,126,889

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	10
A.	Overview	10
B.	Claims.....	11
C.	Burden of Proof for Invalidity.....	12
D.	Claim Construction	13
E.	Person of Ordinary Skill in the Art	13
F.	Prior Art.....	13
G.	Anticipation	14
H.	Obviousness.....	15
V.	PERSON OF ORDINARY SKILL IN THE ART	18
VI.	THE '889 PATENT AND THE CHALLENGED CLAIMS	19
A.	Overview of the '889 Patent.....	19
B.	Claim Overview	20
VII.	CLAIM CONSTRUCTION	23
VIII.	GROUND 1	23
A.	Huval and Ross.....	24
B.	Huval and Ross Do Not Teach Elements 1[c], 2[c], or 12[c] ("generating a training dataset comprising summary statistics of user responses describing the state of minds of road users displayed in the plurality of images").....	26

C.	A POSITA Would Not Have Been Motivated to Combine Huval and Ross in the Manner Tesla and Dr. Janét Allege	30
IX.	GROUND 2	35
A.	Cox	35
B.	Huval, Ross, and Cox Do Not Teach Elements 1[c], 2[c], or 12[c] (“generating a training dataset comprising summary statistics of user responses describing the state of minds of road users displayed in the plurality of images”).....	40
C.	A POSITA Would Not Have Been Motivated to Combine Huval, Ross, and Cox in the Manner Tesla and Dr. Janét Allege	49
X.	GROUND 3	55
A.	Munro	55
B.	Huval, Ross, and Munro Do Not Teach Elements 1[d], 2[d], or 12[d] (“training, using the training dataset, a supervised learning based model configured to predict summary statistics describing a state of mind of a road user displayed in an input image”).....	57
C.	Huval, Ross, and Munro Do Not Teach Elements 1[f], 2[f], or 12[f] (“predicting, by the autonomous vehicle, using the supervised learning based model, summary statistics describing a state of mind of the road user in the new image”)	63
D.	A POSITA Would Not Have Been Motivated to Combine Huval, Ross, and Munro in the Manner Tesla and Dr. Janét Allege	65
XI.	GROUND 4	66
A.	Huval, Ross, Cox, and Munro Do Not Teach Elements 1[d], 2[d], or 12[d] (“training, using the training dataset, a supervised learning based model configured to predict summary statistics describing a state of mind of a road user displayed in an input image”).....	66
B.	Huval, Ross, Cox, and Munro Do Not Teach Elements 1[f], 2[f], or 12[f] (“predicting, by the autonomous vehicle, using the supervised learning based model, summary statistics describing a state of mind of the road user in the new image”)	69

C.	A POSITA Would Not Have Been Motivated to Combine Huval, Ross, Cox, and Munro in the Manner Tesla and Dr. Janét Allege	71
XII.	CONCLUSION.....	72

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. 11,126,889 (the “’889 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 1-20 of the ’889 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 1-4 of the Petition, render obvious claims 1-20 of the ’889 patent. In my opinion, the alleged prior art references, as presented in Grounds 1-4 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 '889 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 '889 patent.

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

<u>Document</u>	<u>Exhibit/Paper #</u>
'889 patent	1001
File history for the '889 patent	1002
Declaration of Jason Janét	1003
U.S. Patent Application Publication No. 2018/0373980 ("Huval")	1004
U.S. Patent No. 10,496,091 ("Ross")	1005

International Publication WO 2014/210334 ("Cox")	1006
U.S. Patent Application Publication No. 2016/0162456 ("Munro")	1009
Petition	1
Patent Owner's Preliminary Response	9
C. Aggarwal, "Neural Networks and Deep Learning", Springer ISBN 978-3-319-94462-3	2041
Kooij JFP, Schneider N, Flohr F, Gavrila 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	2043

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 '889 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 '889 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 '889 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 '889 patent.

29. As I discuss in detail below, it is my opinion that claims 1-20 of the '889 patent are not rendered obvious by the alleged prior art references, as presented in Grounds 1-4 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 '889 patent are patentable.

31. I have been informed 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 ’889 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 '889 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 two or three years of work experience in computer vision, image/video processing, 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 '889 Patent as of July 5, 2017 under either prong of Dr. Janét's definition. *See* Section II above and my CV.

VI. THE '889 PATENT AND THE CHALLENGED CLAIMS

A. Overview of the '889 Patent

54. The '889 patent describes “machine learning based prediction or human interactions with autonomous vehicles”, ‘Ex. 1001, Title. In essence, it describes “[s]ystems and methods for predicting user interaction with vehicles enabling motorists, cyclists, and pedestrians to anticipate the actions of other motorists, cyclists, and pedestrians.”, Ex. 1001, 1:57-60. To this effect, “[a] computing device receives an image and a video segment of a road scene, the first at least one of an image and a video segment being taken from a perspective of a participant in the road scene and then generates stimulus data based on the image and the video segment. Stimulus data is transmitted to a user interface and 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. The computing device aggregates a subset of the plurality of response data to form statistical data and a model is created based on the statistical data. The model is applied to another image or video segment and a prediction of user behavior in [] another image or video segment is generated”, Ex. 1001, Abstract

55. Hence, in summary, the proposed systems and methods of the ‘889 patent collect images and videos of motorists, pedestrians, cyclists, etc., gather a plurality of user responses that describe a state of mind of a road user, generate a training dataset comprising summary statistics of user responses, and use the above information to train supervised learning models that can infer future actions. When deployed in a vehicle, the system analyzes surrounding road users and estimates the probability of various behaviors—such as a pedestrian crossing the street or a cyclist entering traffic—allowing the vehicle to make safer and more informed driving decisions by anticipating human actions rather than simply reacting to them.

B. Claim Overview

56. Independent claims 1, 2, and 12 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 system comprising: a computer processor; and a non-transitory computer readable storage medium storing instructions that when executed by the computer processor cause the computer processor to perform steps of a computer-implemented method for controlling an autonomous vehicle based on a predicted state of mind of road users in a scene captured by a camera of the autonomous vehicle, the steps comprising:

1[a]	receiving a plurality of images displaying road scenes captured by one or more vehicles;
1[b]	receiving a plurality of user responses, each user response describing a state of mind of a road user displayed in one or more images;
1[c]	generating a training dataset comprising summary statistics of user responses describing the state of minds of road users displayed in the plurality of images;
1[d]	training, using the training dataset, a supervised learning based model configured to predict summary statistics describing a state of mind of a road user displayed in an input image;
1[e]	receiving, by an autonomous vehicle, a new image captured by a camera of the autonomous vehicle, the new image of a scene including a road user;
1[f]	predicting, by the autonomous vehicle, using the supervised learning based model, summary statistics describing a state of mind of the road user in the new image; and
1[g]	controlling the autonomous vehicle based on the prediction of the supervised learning based model.
Claim 2	Limitations
2[pre]	A computer-implemented method for controlling an autonomous vehicle based on a predicted state of mind of road users in a scene captured by a camera of the autonomous vehicle, the method comprising:
2[a]	receiving a plurality of images displaying road scenes captured by one or more vehicles;
2[b]	receiving a plurality of user responses, each user response describing a state of mind of a road user displayed in one or more images;

2[c]	generating a training dataset comprising summary statistics of user responses describing the state of minds of road users displayed in the plurality of images;
2[d]	training, using the training dataset, a supervised learning based model configured to predict summary statistics describing a state of mind of a road user displayed in an input image;
2[e]	receiving, by an autonomous vehicle, a new image captured by a camera of the autonomous vehicle, the new image of a scene including a road user;
2[f]	predicting, by the autonomous vehicle, using the supervised learning based model, summary statistics describing a state of mind of the road user in the new image; and
2[g]	controlling the autonomous vehicle based on the prediction of the supervised learning based model.
Claim 12	Limitations
12[pre]	A non-transitory computer readable storage medium storing instructions that when executed by a computer processor cause the computer processor to perform steps of a computer-implemented method for controlling an autonomous vehicle based on a predicted state of mind of road users in a scene captured by a camera of the autonomous vehicle, the steps comprising:
12[a]	receiving a plurality of images displaying road scenes captured by one or more vehicles;
12[b]	receiving a plurality of user responses, each user response describing a state of mind of a road user displayed in one or more images;
12[c]	generating a training dataset comprising summary statistics of user responses describing the state of minds of road users displayed in the plurality of images;

12[d]	training, using the training dataset, a supervised learning based model configured to predict summary statistics describing a state of mind of a road user displayed in an input image;
12[e]	receiving, by an autonomous vehicle, a new image captured by a camera of the autonomous vehicle, the new image of a scene including a road user;
12[f]	predicting, by the autonomous vehicle, using the supervised learning based model, summary statistics describing a state of mind of the road user in the new image; and
12[g]	controlling the autonomous vehicle based on the prediction of the supervised learning based model.

VII. CLAIM CONSTRUCTION

57. In my opinion, none of the terms of the '889 patent require any specific construction in order to resolve the issues presented in Tesla's petition and in my opinions as expressed below. In rendering the opinions set forth herein, I have applied the plain and ordinary meaning of all terms of the challenged claims to a person of ordinary skill in the art.

VIII. GROUND 1

58. I understand that Tesla alleges (and Dr. Janét opines) that claims 1-20 of the '889 patent are obvious in view of Huval and Ross. *See* Pet. at 17-55; Ex. 1003, ¶¶ 65-142. For the reasons I discuss below, I disagree. In my opinion, none of claims 1-20 are obvious in view of Huval and Ross.

A. Huval and Ross

59. In arriving at my opinions concerning Tesla and Dr. Janét's assertions regarding claims 1-20 of the '889 patent and Huval and Ross, I have considered Huval and Ross's disclosures from the perspective of a POSITA. Below I discuss some aspects of Huval and Ross that are relevant to my opinions.

60. Huval describes a “[m]ethod for training and refining an artificial intelligence”, Ex. 1004, Title. To this effect, Huval describes “training a neural network on a training set to identify objects in optical images; receiving a manual label attributed to an optical image recorded by a road vehicle during its operation by a human annotator; passing the optical image through the neural network to generate an automated label attributed to the optical image; in response to the manual label differing from the automated label, serving the optical image to a human annotator for manual confirmation of one of the manual label and the automated label; appending the training set with the optical image containing one of the manual label and the automated label based on confirmation received from the human annotator; and retraining the neural network, with the training set, to identify objects in optical images”, Ex. 1004, Abstract.

61. Ross describes “[b]ehavior and intent estimations of road users for autonomous vehicles” (Ex. 1005, Title). 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.

62. 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. Huval and Ross Do Not Teach Elements 1[c], 2[c], or 12[c] (“generating a training dataset comprising summary statistics of user responses describing the state of minds of road users displayed in the plurality of images”)

63. In my opinion, Huval and Ross do not teach elements 1[c], 2[c], or 12[c], which each require “generating a training dataset comprising summary statistics of user responses describing the state of minds of road users displayed in the plurality of images.”

64. Tesla and Dr. Janét allege that Huval teaches this element in two ways, either through (i) “combin[ing] the labels by determining an ‘average’ to generate a final label for each image,” or (ii) “[t]he training set of Huval . . . includ[ing] the manual labels that are ‘packaged’ into the training set.” Pet at 33-34; Ex. 1003, ¶¶ 91, 94. I disagree. I discuss each of these theories below and the reasons they are incorrect.

65. First, Huval’s combining and averaging of the labels is not a summary statistic of any kind, as required by the challenged claims. Tesla and Dr. Janét appear to assume that Huval’s averaging constitutes a summary statistic. To the contrary, the “averag[ing]” that Huval describes involves taking multiple human annotator responses that are sought and obtained because of conflicting manual and automated

labels to a given image and “averag[ing] or otherwise combin[ing] responses (e.g., object types, label locations, activation times, deactivation times, etc.) from these human annotators into a final manually-confirmed label for the optical image.” Ex. 1004, [0085]. Huval then gives the example of “calculat[ing] a binary combination (e.g., either the manual label or the automated label) of: a first conformation of one of the first manual label and the first automated label received from a first human annotator in the set; a second confirmation of one of the first manual label and the first automated label received from a second human annotator in the set; etc.” *Id.* As these disclosures make clear, the “averag[ing] or otherwise combin[ing] responses” (either through “calculat[ing] a binary combination” or otherwise) does not generate any kind of summary statistics, but instead generates a “final manually-confirmed label for the optical image”. *Id.*

66. The Petition and Dr. Janét incorrectly compare Huval’s description of this averaging to the ’889 patent’s discussion of “measurements of the central tendency of the distribution of scores like the mean, median, or mode.” Pet at 34; Ex. 1003, ¶ 92 (both citing Ex. 1001, 10:60-67, 6:8-24). Nothing in the cited portions of Huval describes computing a mean, median, or mode of any figures.

67. Second, Tesla and Dr. Janét incorrectly allege that “[t]he training set of Huval comprises summary statistics of [the] user responses at least because it includes the manual labels that are ‘packaged’ into the training set and reflect the

summarized statistics of the content of all accurate manual human labels that have been obtained and are used for training.” Pet. at 34-35; Ex. 1003, ¶ 94. The fact that Huval teaches that the training set includes the images and their labels does not mean that the training set includes any kind of summary statistics. Tesla and Dr. Janét incorrectly allege that “because these aggregated labels are a collection of measurements summarizing how data has been labeled, a POSITA would have recognized that these aggregated labels form statistics.” Pet. at 35; Ex. 1003, ¶ 94. In my opinion, the labels stored in Huval’s training set are merely the manual or automated labels for the content of the image (e.g., a car defined by a boundary box) and are not any kind of statistic, much less a summary statistic. Tesla and Dr. Janét do not point to anything that a POSITA would consider to be a statistic, much less explain how any such statistic is a summary statistic. For example, despite referencing the ’889 patent specification’s indication that “measurements of the central tendency of the distribution of scores like the mean, median, or mode” may constitute summary statistics, neither Tesla nor Dr. Janét attempt to explain how these labels in the training set include any mean, median, mode, or other measure of central tendency.” *See id.*; *see also* Pet at 34; Ex. 1003, ¶ 92 (both citing Ex. 1001, 10:60-67, 6:8-24).

68. I have considered the paragraphs of Huval and Ross 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 Huval's combining/averaging responses or training set constitute summary statistics. *See* Pet. at 33-37; Ex. 1003, ¶¶ 91-95. Instead, paragraph [0085] is Huval's lone description of any combining/averaging of responses, and the paragraphs that discuss Huval's training set do not teach any kind of statistics, much less any summary statistics. Both the Petition and Dr. Janét's declaration also cross-reference their respective discussions of element 2[b]. *See* Pet. at 37; Ex. 1003, ¶ 95. I have reviewed those sections, as well as the portions of Huval and Ross cited in them, and they do not alter my opinions expressed above.

69. The portions of the Petition and Dr. Janét's declaration discussed above relate to claim 2 of the '889 patent. Both the Petition and Dr. Janét's declaration merely refer back to these discussions when discussing elements 1[c] and 12[c]. *See* Pet. at 52, 54; Ex. 1003, ¶¶ 121, 138. In my opinion, Cox fails to teach or suggest elements 1[c] and 12[c] for the same reasons as discussed above.

70. Therefore, in my opinion, Huval and Ross do not teach or suggest elements 1[c], 2[c], or 12[c] ("generating a training dataset comprising summary statistics of user responses describing the state of minds of road users displayed in the plurality of images").

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

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

72. Tesla and Dr. Janét allege the following:

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

- Huval’s annotation labels provided by human annotators label observed road user actions in image data captured by vehicles, where such labeled actions correspond to road user intent;
- Huval’s machine learning model is (1) trained with such human annotation labels and (2) predicts a “set of possible intents or hypotheses identifying possible next actions (and predicted trajectories), a given point in time when the action is likely to occur, and associated likelihood values” for road users displayed in input image data; and
- Huval’s autonomous vehicle is controlled based on the predictions.

Pet. at 20-21; Ex. 1003, ¶ 72.

73. With respect to Tesla and Dr. Janét’s first combination, as best I understand their theory, they propose to modify Huval such that the annotation labels provided by human annotators (but not the automated labels) be labels that identify observed road user actions. *See* Pet. at 20; Ex. 1003, ¶ 72; *see also* Ex. 1004 at, e.g., [0008] (Huval describing use of both human and automated labels in its system), Ex.

1005 at, e.g., 9:41-10:10 (Ross describing labeling of road user intents). In my opinion, such a modification would not make sense to a POSITA, as it would result in the human annotations relating to road user intents but the automated annotations relating only to the identification and location of objects in the image not to mention the significant redesign of Huval's automated annotation neural network in order to predict road user intents in addition to labels. Tesla and Dr. Janét do not attempt to explain how modifying the human labels, but not the automated labels, would impact Huval's system as a whole.

74. Moreover, as Ross describes, its system works by looking at image/video data past the particular point in time to determine what the road user in fact did in a given scenario, rather than relying on either a human or automated prediction or judgment as to what the road user will do based on an image at a fixed point in time (as Huval's system does). *See* Ex. 1004 at, e.g., [0008], Ex. 1005 at, e.g., 9:41-10:10. Tesla and Dr. Janét do not make clear which of these two mechanisms their proposed combination of Huval and Ross would use to label images with road user intents. *See* Pet. at 20-26; Ex. 1003, ¶¶ 72-81. If the proposed combination relies on Huval's existing process of using human and automated labeling of the images, the prediction of road user intents will increase the conflict between such human and automated labels, requiring significant additional usage of the human annotation portals for label conflict resolution. Huval specifically teaches

away from requiring significant amounts of human annotation, as it drastically increases the cost of training and operating the system. *See* Ex. 1004, [0011] (“Similarly, manual labeling of features in optical data by one human annotator and subsequent checking of these labeled features by other human annotators . . . may further escalate costs to develop the training set, both in terms of human labor hours per label and transmission of the same optical data to multiple local computer systems.”). Conversely, if the proposed combination relies on Ross’s process of looking at image/video data past the particular point in time to determine what the road user in fact did in a given scenario, Huval’s system would have to be significantly reconfigured to provide the additional image/video data to make such a determination of what the road user in fact did. The use of that additional data would require significantly more storage space and processing power.¹ Nowhere do

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. To the extent it is performed by human annotators, this further teaches away from combining Huval and Ross because of the significant cost of additional human annotation. *See* Ex. 1004, [0011].

Tesla or Dr. Janét account for these realities in their proposed combination of Huval and Ross.

75. Tesla and Dr. Janét's second combination suffers from the same flaws. Training Huval's machine learning model trained with Ross's human annotation labels that identify road user intents does not make sense if the automated annotations in Huval are still limited to identifying and locating objects in the image. And, as I discuss above, training Huval's model in this manner using Huval's existing process of human and automated labeling of the images, the prediction of road user intents will increase the conflict between such human and automated labels, requiring significant additional usage of the human annotation portals for label conflict resolution, from which Huval specifically teaches away. *See* Ex. 1004, [0011]. As I also discuss above, the proposal to modify the training of Huval's model using Ross's process of looking at image/video data past the particular point in time to determine what the road user in fact did in a given scenario would require Huval's system to have been significantly reconfigured to provide the additional image/video data to make such a determination of what the road user in fact did. Nowhere do Tesla or Dr. Janét account for these realities in their proposed combination of Huval and Ross.

76. In addition, modifying Huval's system to predict road user intents instead of identifying and locating road users would require significant modification

to Huval's system. According to Tesla and Dr. Janét, the proposed modification would require Huval's model, instead of merely identifying and locating an object in an image, to predict several things—specifically, (a) a set of possible intents, (b) trajectories for each such possible intent), (c) the time within which each such possible intent may occur, and (d) the associated likelihood values for each prediction. Tesla and Dr. Janét do not explain how Huval's system would be reconfigured to allow it to make all of those predictions, or even whether a POSITA would have had a reasonable expectation of success in doing so.

77. Tesla and Dr. Janét's third combination is significantly flawed as well. Nowhere do Tesla or Dr. Janét account for how Huval's system would have to be reconfigured to not only make Ross's intent predictions, but also how Huval's system would have to be reconfigured in order to control an autonomous vehicle based on such predictions. Autonomous vehicle systems must be robust enough and have enough processing power to make such intent predictions and control the vehicle in very short amounts of time to allow the autonomous vehicle to react to the intent prediction in time (e.g., by steering/swerving, braking, etc.). As discussed above, Tesla and Dr. Janét's proposed modification would require Huval's model to predict several things—specifically, (a) a set of possible intents, (b) trajectories for each such possible intent), (c) the time within which each such possible intent may occur, and (d) the associated likelihood values for each prediction. Not only do Tesla

and Dr. Janét fail to explain how Huval's system would be reconfigured to allow it to make all of those predictions (or whether a POSITA would have had a reasonable expectation of success in doing so), but they also fail to explain how Huval's system would be reconfigured to make those predictions in a fast enough manner to allow the system to control an autonomous vehicle (e.g., by steering/swerving, braking, etc.) in time.

78. Therefore, in my opinion, a POSITA would not have been motivated to combine Huval and Ross in the manner Tesla and Dr. Janét allege.

IX. GROUND 2

79. I understand that Tesla alleges (and Dr. Janét opines) that claims 1-20 of the '889 patent are obvious in view of Huval, Ross, and Cox. *See* Pet. at 55-68; Ex. 1003, ¶¶ 143-171. For the reasons I discuss below, I disagree. In my opinion, none of claims 1-20 are obvious in view of Huval, Ross, and Cox.

A. Cox

80. In arriving at my opinions concerning Tesla and Dr. Janét's assertions regarding claims 1-20 of the '889 patent and Huval, Ross, and Cox, I have considered Huval, Ross, and Cox's disclosures from the perspective of a POSITA. In my discussion of Ground 1 above, I discussed aspects of Huval and Ross that are relevant to my opinions. Below I discuss some aspects of Cox that are relevant to my opinions.

81. As Cox indicates, its “invention relates to machine learning, in particular to machine learning augmented by measurements acquired from human populations.” Ex. 1006, [0001].

82. 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 ’889 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. 1006, [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.

83. 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. 1006, [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.*

84. 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. 1006, [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.

1006, [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. 1006, [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. 1006, [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. 1006, [0009].

85. 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. 1006, [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. 1006, [0035].

86. 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. 1006, [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.

87. 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. 1006, [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. Huval, Ross, and Cox Do Not Teach Elements 1[c], 2[c], or 12[c] (“generating a training dataset comprising summary statistics of user responses describing the state of minds of road users displayed in the plurality of images”)

88. In my opinion, the combination of Huval, Ross, and Cox does not teach elements 1[c], 2[c], or 12[c], which each require “generating a training dataset

comprising summary statistics of user responses describing the state of minds of road users displayed in the plurality of images.”

89. Tesla and Dr. Janét allege that this element is met because “[t]he Huval-Ross-Cox combination provides that psychometric data is generated from the Huval-Ross combination’s human annotator response labels regarding road user actions in image data and added to training data for Huval’s model.” Pet. at 63; Ex. 1003, ¶160. As best I understand Tesla and Dr. Janét’s theory for this element, they propose to modify the combination of Huval and Ross (as described in Ground 1) to instead include one or more of several features of Cox as allegedly meeting the claimed “summary statistics.” As I discuss below, each of their theories are incorrect.

Psychometric Data

90. Cox’s psychometric data is not “summary statistics of user responses” as required by the challenged claims. Cox makes clear that its psychometric data is acquired as the annotators classify each of the images (*see* Ex. 1006, [0007]) and that the 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” (*id.*, [0008]). In other words, Cox does not describe its psychometric data as being any kind of summary statistics of the user responses as a whole; instead, each piece of psychometric data relates to a particular annotator’s labeling of a particular image. At best, Tesla and Dr. Janét

allege (without explanation) that Cox’s psychometric dataset somehow “summarizes statistics on how human annotators annotated training images.” Pet. 63; Ex. 1003, ¶ 160. But Tesla and Dr. Janét’s identification of the psychometric dataset as a whole (i.e., every annotator’s response time to every image, the accuracy of every response, and the presentation time of every image to every annotator) is not a summary of any kind.

91. 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 of user responses.” *See* Pet. at 63; Ex. 1003, ¶ 160 (both citing Ex. 1006, “¶¶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.

92. Tesla and Dr. Janét also allege that “Cox’s description of ‘psychometric data’ is consistent with the ’889 Patent’s explanation.” Pet. at 63-64; Ex. 1003, ¶160. I have reproduced the portion of the ’889 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.

'889 patent, 6:8-43. (emphasis and italicization added).

93. 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 '889 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.” '889 patent, 6:11-18. The '889 patent also gives specific examples of summary statistics of user responses by indicating that they can be a “central tendency, variance, skew, kurtosis, scale, or histogram.” *Id.*, 6:23-24. 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” ('889 patent, 6:18-21) does not mean that such a time is a summary statistic of user responses.

Item-Response Curves

94. Tesla and Dr. Janét next allege that Cox’s “item-response curves” can constitute the claimed “summary statistics of user responses.” Pet. at 64; Ex. 1003, ¶ 161. I disagree.

95. 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. 1006, [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. 1006, [0006]. 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. Neither of these portions of Cox teach “generating a training dataset comprising” an item-response curve, as nothing in Cox teaches including any item-response curves in the training dataset. Tesla and Dr. Janét summarily allege that “the item response curve would be included as part of training data for the Huval-Ross model.” Pet. at 64; Ex. 1003, ¶ 161. Tesla and Dr. Janét cite to numerous paragraphs of Huval (i.e., “¶¶ 0008-0018, 0021-0025, 0029-0030, 0036-0050, 0085-0090, 0095-0098, 0118, Figs. 1, 3,

and 4”), but I have reviewed those paragraphs, and none of them teach including item-response curves in the training dataset.

96. I have considered the additional portions of Cox that Tesla and Dr. Janét identify in connection with this discussion. *See* Pet. at 64; Ex. 1003, ¶ 161 (both citing Ex. 1006, “¶¶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 are included in a training dataset as required by the challenged claims.

“distribution of data and labels”

97. Tesla and Dr. Janét next allege that Cox’s “distribution of data and labels” can constitute the claimed “summary statistics of user responses.” Pet. at 65; Ex. 1003, ¶ 162. I disagree.

98. Cox’s only mention of a “distribution of data and labels” is a reference to “the joint distribution of data and labels $P(x, y)$ ” when discussing a “recognition function.” Ex. 1006, [0035]. Nothing in Cox teaches generating such a “distribution of data and labels” for user responses in the training dataset. Tesla and Dr. Janét offer no explanation for how this “distribution of data and labels” meets this element of the challenged claims.

99. I have considered the additional portions of Cox that Tesla and Dr. Janét identify in connection with this discussion. *See* Pet. at 65; Ex. 1003, ¶ 162 (both

citing Ex. 1006, “¶0002-0009, 0010, 0014, 0017-0020, 0025-0035, 0048-0049, 0063, Claim 11”). Other than paragraph [0035], none of these portions of Cox discuss this “distribution of data and labels.”

“end-of-test statistics”

100. Tesla and Dr. Janét next allege that Cox’s “end-of-test statistics” can constitute the claimed “summary statistics of user responses.” Pet. at 65; Ex. 1003, ¶ 162. I disagree.

101. 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. 1006, [0048].

102. 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 are “summary statistics of user responses” or that they are included in the training dataset as required by the challenged claims.

103. I have considered the additional portions of Cox that Tesla and Dr. Janét identify in connection with this discussion. *See* Pet. at 65; Ex. 1003, ¶ 162 (both citing Ex. 1006, “¶0002-0009, 0010, 0014, 0017-0020, 0025-0035, 0048-0049, 0063, Claim 11”). None of those cited portions contains any disclosure that teaches or suggests that these end-of-test statistics are “summary statistics of user responses” as required by the challenged claims.

104. The portions of the Petition and Dr. Janét’s declaration discussed above relate to claim 2 of the ’889 patent. In discussing independent claims 1 and 12 for Ground 2, both the Petition and Dr. Janét’s declaration refer back to their respective discussions of Ground 1. *See* Pet. at 68; Ex. 1003, ¶ 171. I thus understand Tesla and Dr. Janét to not be relying on any disclosure from Cox as allegedly meeting any elements of independent claims 1 and 12 for Ground 2. In my opinion, the combination of Huval and Ross does not meet elements 1[c] and 12[c] for the reasons discussed above with respect to Ground 1.

105. Therefore, in my opinion, Huval, Cox, and Ross do not teach or suggest elements 1[c], 2[c], or 12[c] (“generating a training dataset comprising summary statistics of user responses describing the state of minds of road users displayed in the plurality of images”).

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

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

107. Tesla and Dr. Janét allege the following:

A POSITA would have been motivated to combine Huval and Ross as discussed in [Ground 1]. A POSITA would have been further motivated to combine Huval and Ross with Cox such that:

- Cox's psychometric data is generated from the Huval-Ross combination's human annotator response labels and used for model training;
- Cox's distribution of data and labels (e.g., statistics) and other statistical information are generated from the Huval-Ross combination's human annotator response labels and used for model training; and
- The training of the Huval-Ross model would have included comparing the predicted intents to the human labeled training data to identify prediction errors and optimize the Huval-Ross model such that the prediction errors are minimized.

Pet. at 57-58; Ex. 1003, ¶ 147.

108. As I discuss above with respect to Ground 1, a POSITA would not have been motivated to combine Huval and Ross in the manner Tesla and Dr. Janét allege. Those flaws in Tesla and Dr. Janét's theory apply equally to the combination of Huval, Ross, and Cox.

109. 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.

110. 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. 2041 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.

111. 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. 1006, [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.

112. 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.

113. 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. 1006, [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.

114. 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.

115. In addition, I disagree with each of the specific modifications Tesla and Dr. Janét allege a POSITA would have been motivated to make to Huval and Ross in view of Cox.

116. Tesla and Dr. Janét’s first alleged modification is that “Cox’s psychometric data is generated from the Huval-Ross combination’s human annotator response labels and used for model training.” Pet. at 57; Ex. 1003, ¶ 147. However, the combination of Huval and Ross already describes a process for training the model, as I discuss with respect to Ground 1 above. Tesla and Dr. Janét do not

describe what the addition of Cox's psychometric data to Huval and Ross adds to its training process or how they propose to use Cox's psychometric data in the training process. As I discuss above, Cox describes using the psychometric data as an input to its human-weighted loss function. But, Tesla and Dr. Janét do not appear to allege that a POSITA would have been motivated to combine Huval and Ross with Cox so as to implement Cox's human-weighted loss function.

117. Tesla and Dr. Janét's second alleged modification is that "Cox's distribution of data and labels (e.g., statistics) and other statistical information are generated from the Huval-Ross combination's human annotator response labels and used for model training." Pet. at 57; Ex. 1003, ¶ 147. However, as discussed above, Cox only mentions a "distribution of data and labels" once and does not describe using it in model training at all. Ex. 1006, [0035]. Thus, it is unclear how Tesla and Dr. Janét propose to use this "distribution of data and labels" in training the model. It is also unclear what Tesla and Dr. Janét are referring to by "other statistical information."

118. Tesla and Dr. Janét's third alleged modification is that "[t]he training of the Huval-Ross model would have included comparing the predicted intents to the human labeled training data to identify prediction errors and optimize the Huval-Ross model such that the prediction errors are minimized." Pet. at 57-58; Ex. 1003, ¶ 147. However, Huval and Ross each already teach processes for identifying

prediction errors and optimizing the model to minimize prediction errors. *See, e.g.*, Ex. 1004, [0008]-[0009], [0073]-[0098], Figs. 1, 3; Ex. 1005, 9:41-10:10. Thus, a POSITA would not have been motivated to combine Huval and Ross with Cox for this purpose.

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

X. GROUND 3

120. I understand that Tesla alleges (and Dr. Janét opines) that claims 1-20 of the '889 patent are obvious in view of Huval, Ross, and Munro. *See* Pet. 68-74; Ex. 1003, ¶¶ 172-189. For the reasons I discuss below, I disagree. In my opinion, none of claims 1-20 are obvious in view of Huval, Ross, and Munro.

A. Munro

121. In arriving at my opinions concerning Tesla and Dr. Janét's assertions regarding claims 1-20 of the '889 patent and Huval, Ross, and Munro, I have considered Huval, Ross, and Munro's disclosures from the perspective of a POSITA. In my discussion of Ground 1 above, I discussed aspects of Huval and Ross that are relevant to my opinions. Below I discuss some aspects of Munro that are relevant to my opinions.

122. In marked contrast to Huval, Ross, and the '889 patent, Munro is generally directed to natural language processing. Thus, instead of being concerned

with determining the intent of a road user in an image, Munro is directed to improving processes for analyzing the content of documents. *Compare, e.g.*, Ex. 1004, [0008]-[0009], Ex. 1005, 1:7-26 *and* Ex. 1001, 1:19-53 *with, e.g.*, Ex. 1009, Abstract, [0003], [0004].

123. Munro proposes “[m]ethods for generating natural language processing systems.” Ex. 1009, 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. 1009, Abstract.

124. 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. Huval, Ross, and Munro Do Not Teach Elements 1[d], 2[d], or 12[d] (“training, using the training dataset, a supervised learning based model configured to predict summary statistics describing a state of mind of a road user displayed in an input image”)

125. In my opinion, the combination of Huval, Ross, and Munro does not teach elements 1[d], 2[d], or 12[d], which each require “training, using the training dataset, a supervised learning based model configured to predict summary statistics describing a state of mind of a road user displayed in an input image.”

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

As combined, the training of the Huval-Ross model would have been performed as discussed in [Ground 1, element 2[d]], except the training data includes the generated summary statistics as taught by Munro, and the model is trained on such training data in the same manner discussed in [Ground 1, element 2[d]].

Pet. at 73; Ex. 1003, ¶ 183.

127. 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 discussion of element 2[c] for Ground 1), Tesla and Dr. Janét propose to modify Huval and Ross in view of Munro to include “tallies of the number of annotators who classified the behavior or activity in the images as having various labels in their responses” and to store these tallies in Huval’s training data. *See* Pet. at 72-73; Ex. 1003, ¶ 182.

128. 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 69; Ex. 1003, ¶ 173 (citing Ex. 1009, [0140]). 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 training dataset. Moreover, nothing in Huval, Ross, or Munro teaches that a model is configured to predict such tallies, as required by the challenged claims. 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.*

129. Moreover, machine learning for natural language processing (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.

130. Machine learning for natural language processing 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.

131. 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.

132. 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.

133. 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.

134. 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.

135. Munro's discussion of its aggregation process (Ex. 1009, [0140]) 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 Huval or Ross suggests excluding data from the training dataset because annotators disagree.

136. 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 73; Ex. 1003, ¶ 183. None of the cross-referenced sections of the Petition or Dr. Janét’s declaration (or the portions of Huval, Ross, and/or Munro cited therein) alter my opinions with respect to these claim elements.

137. The portions of the Petition and Dr. Janét’s declaration discussed above relate to claim 2 of the ’889 patent. Both the Petition and Dr. Janét’s declaration merely refer back to these discussions of element 2[d] when discussing elements 1[d] and 12[d]. *See* Pet. at 74; Ex. 1003, ¶¶ 187-188. In my opinion, Huval, Ross, and Munro fail to teach or suggest elements 1[d] and 12[d] for the same reasons as discussed above.

138. Therefore, in my opinion, Huval, Ross, and Munro do not teach or suggest elements 1[d], 2[d], or 12[d] (“training, using the training dataset, a supervised learning based model configured to predict summary statistics describing a state of mind of a road user displayed in an input image”).

C. Huval, Ross, and Munro Do Not Teach Elements 1[f], 2[f], or 12[f] (“predicting, by the autonomous vehicle, using the supervised learning based model, summary statistics describing a state of mind of the road user in the new image”)

139. In my opinion, the combination of Huval, Ross, and Munro does not teach elements 1[f], 2[f], or 12[f], which each require “predicting, by the

autonomous vehicle, using the supervised learning based model, summary statistics describing a state of mind of the road user in the new image.”

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

As combined, the Huval-Ross trained model would have operated on new images to determine the actions and possible intents of vehicles, bicyclists, and pedestrians in the images, teaching element 2[f] for the same reasons as discussed in [Ground 1, element 2[f]].

Pet at 74; Ex. 1003, ¶ 185.

141. As this allegation makes clear, Tesla and Dr. Janét allege that “determin[ing] the actions and possible intents of vehicles, bicyclists, and pedestrians in the images” constitutes the claimed “predicting . . . summary statistics describing a state of mind of the road user in the new image.” Pet at 74; Ex. 1003, ¶ 185. I disagree that determining actions and possible intents constitutes predicting summary statistics, as required by the challenged claims. Nothing about an action or a possible intent is any kind of statistic, much less a summary statistic.

142. 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 74; Ex. 1003, ¶ 185. None of the cross-referenced sections of the Petition or Dr. Janét’s declaration (or the portions of Huval, Ross, and/or Munro cited therein) alter my opinions with respect to these claim elements.

143. The portions of the Petition and Dr. Janét’s declaration discussed above relate to claim 2 of the ’889 patent. Both the Petition and Dr. Janét’s declaration merely refer back to these discussions of element 2[f] when discussing element 1[f] and 12[f]. *See* Pet. at 74; Ex. 1003, ¶¶ 187-188. In my opinion, Huval, Ross, and Munro fail to teach or suggest elements 1[f] and 12[f] for the same reasons as discussed above.

144. Therefore, in my opinion, Huval, Ross, and Munro do not teach or suggest elements 1[f], 2[f], or 12[f] (“predicting, by the autonomous vehicle, using the supervised learning based model, summary statistics describing a state of mind of the road user in the new image”).

D. A POSITA Would Not Have Been Motivated to Combine Huval, Ross, and Munro in the Manner Tesla and Dr. Janét Allege

145. As I discuss in Sections X.A and X.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 Huval and Ross for the same reasons.

XI. GROUND 4

146. I understand that Tesla alleges (and Dr. Janét opines) that claims 1-20 of the '889 patent are obvious in view of Huval, Ross, Cox, and Munro. *See* Pet. 75-78; Ex. 1003, ¶¶ 190-202. For the reasons I discuss below, I disagree. In my opinion, none of claims 1-20 are obvious in view of Huval, Ross, Cox, and Munro.

147. In arriving at my opinions concerning Tesla and Dr. Janét's assertions regarding claims 1-20 of the '889 patent and Huval, Ross, Cox, and Munro, I have considered Huval, Ross, Cox, and Munro's disclosures from the perspective of a POSITA. In my discussion of Grounds 1-3 above, I discussed aspects of Huval, Ross, Cox, and Munro that are relevant to my opinions.

A. Huval, Ross, Cox, and Munro Do Not Teach Elements 1[d], 2[d], or 12[d] (“training, using the training dataset, a supervised learning based model configured to predict summary statistics describing a state of mind of a road user displayed in an input image”)

148. In my opinion, the combination of Huval, Ross, Cox, and Munro does not teach elements 1[d], 2[d], or 12[d], which each require “training, using the training dataset, a supervised learning based model configured to predict summary statistics describing a state of mind of a road user displayed in an input image.”

149. Tesla and Dr. Janét allege the following for this element:

As combined, the training of Huval-Ross model would have been performed as discussed in [element 2[d] – Ground 1] and [element 2[d] – Ground 2], except the training data includes the generated summary statistics as taught by Munro, and the Huval-Ross model is trained on

such training data in the same manner discussed in [element 2[d] – Ground 1] and [element 2[d] – Ground 2].

Pet. at 76-77; Ex. 1003, ¶ 195.

150. It is unclear what Tesla and Dr. Janét’s theory is and what the Huval-Ross model is configured to predict. In discussing element 2[d] in Ground 1, Tesla and Dr. Janét allege that “Huval’s model is configured to predict a ‘set of possible intents or hypotheses identifying possible next actions (and predicted trajectories), a given point in time when the action is likely to occur, and associated likelihood values’ for road users (e.g., pedestrians, vehicles, bicyclists) displayed in input image data.” Pet. at 38; Ex. 1003, ¶ 97. In discussing element 2[d] in Ground 2, Tesla and Dr. Janét allege that these same features of Huval and Ross are what are predicted. Pet. at 65-66; Ex. 1003, ¶ 163. Here in Ground 4, however, Tesla and Dr. Janét appear to allege that “the training data includes the generated summary statistics as taught by Munro,” i.e. the tallies from Munro’s aggregation process. It is unclear whether Tesla and Dr. Janét allege that the model would be configured to predict those tallies, whether those tallies are merely used in the training to “compar[e] the predicted intents to the human labeled training data to identify prediction errors and optimize the Huval-Ross model” (Pet. at 66; Ex. 1003, ¶ 163), or both. In any case, I disagree with Tesla and Dr. Janét’s theory.

151. Munro does not teach configuring a model to predict its tallies. Instead, as I discuss above, Munro makes clear that the purpose of its 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. 1009, [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 training dataset. Moreover, nothing in Huval, Ross, or Munro teaches that a model is configured to predict such tallies, as required by the challenged claims. 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.*

152. Munro likewise does not teach that its tallies are used to “identify prediction errors and optimize the Huval-Ross model.” *See* Pet. at 66; Ex. 1003, ¶ 163. Instead, the tallies are used to determine whether a given document should be excluded from the training set. *See* Ex. 1009, [0140]. Nothing about this process identifies prediction errors—in fact, it is designed to excuse prediction “errors” by “determin[ing] whether said document may be reliably annotated for use in training the natural language model.” *Id.* This process likewise does not optimize the underlying model but instead enhances the quality of the training dataset used to train the model.

153. I have considered the 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 76-77; Ex. 1003, ¶ 195. None of the cross-

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

154. The portions of the Petition and Dr. Janét’s declaration discussed above relate to claim 2 of the ’889 patent. Both the Petition and Dr. Janét’s declaration merely refer back to these discussions of element 2[d] when discussing element 1[d] and 12[d]. *See* Pet. at 78; Ex. 1003, ¶¶ 200-201. In my opinion, Huval, Ross, Cox, and Munro fail to teach or suggest elements 1[d] and 12[d] for the same reasons as discussed above.

155. Therefore, in my opinion, Huval, Ross, Cox, and Munro do not teach or suggest elements 1[d], 2[d], or 12[d] (“training, using the training dataset, a supervised learning based model configured to predict summary statistics describing a state of mind of a road user displayed in an input image”).

B. Huval, Ross, Cox, and Munro Do Not Teach Elements 1[f], 2[f], or 12[f] (“predicting, by the autonomous vehicle, using the supervised learning based model, summary statistics describing a state of mind of the road user in the new image”)

156. In my opinion, the combination of Huval, Ross, Cox, and Munro does not teach elements 1[f], 2[f], or 12[f], which each require “predicting, by the autonomous vehicle, using the supervised learning based model, summary statistics describing a state of mind of the road user in the new image.”

157. For these elements, Tesla and Dr. Janét allege that “[a]s combined, the Huval-Ross model would have operated on new images to determine the actions and

possible intents of vehicles, bicyclists, and pedestrians in the images.” Pet. at 77; Ex. 1003, ¶ 197. As I discuss above with respect to Ground 3, determining actions and possible intents of vehicles, bicyclists, and pedestrians is not predicting summary statistics, as such actions/intents are not summary statistics. *See* Section X.C.

158. I have considered the 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 77; Ex. 1003, ¶ 197. None of the cross-referenced sections of the Petition or Dr. Janét’s declaration alter my opinions with respect to these claim elements.

159. The portions of the Petition and Dr. Janét’s declaration discussed above relate to claim 2 of the ’889 patent. Both the Petition and Dr. Janét’s declaration merely refer back to these discussions of element 2[f] when discussing element 1[f] and 12[f]. *See* Pet. at 78; Ex. 1003, ¶¶ 200-201. In my opinion, Huval, Ross, Cox, and Munro fail to teach or suggest elements 1[f] and 12[f] for the same reasons as discussed above.

160. Therefore, in my opinion, Huval, Ross, Cox, and Munro do not teach or suggest elements 1[f], 2[f], or 12[f] (“predicting, by the autonomous vehicle, using the supervised learning based model, summary statistics describing a state of mind of the road user in the new image”).

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

161. In my opinion, a POSITA would not have been motivated to combine Huval, Ross, Cox, and Munro in the manner Tesla and Dr. Janét allege.

162. Tesla and Dr. Janét allege the following:

A POSITA would have been motivated to combine Huval, Ross, Cox and Munro for the same reasons discussed in [Grounds 2 and 3]. The same motivations that would apply to add Munro to Huval and Ross apply equally to the combination of Huval, Ross, and Cox. Similarly, the same motivations that would apply to add Cox to Huval and Ross apply equally to the combination of Huval, Ross, and Munro.

Pet. at 75; Ex. 1003, ¶ 191.

163. As I discuss above, a POSITA would not have been motivated to combine Huval and Ross. *See* Section VIII.C. As I also discuss above, a POSITA would not have been motivated to combine Huval, Ross, and Cox. *See* Section IX.C. And as I further discuss above, a POSITA would not have been motivated to combine Huval, Ross, and Munro. *See* Section X.D. All of those issues with those combinations apply equally to Tesla and Dr. Janét's proposed combination of Huval, Ross, Cox, and Munro. Here too, Tesla and Dr. Janét do not address any of those identified issues, including for example Cox's focus on face detection and Munro's disclosure relating to NLP.

XII. CONCLUSION

164. 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.

165. 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 29, 2026



Dr. Eli Saber, Ph.D.