Title (en)

BIAS DETECTION AND EXPLAINABILITY OF DEEP LEARNING MODELS

Title (de)

VORSPANNUNGSDETEKTION UND ERKLÄRBARKEIT VON DEEP-LEARNING-MODELLEN

Title (fr)

DÉTECTION DE POLARISATION ET EXPLICATION DE MODÈLES D'APPRENTISSAGE PROFOND

Publication

EP 4066166 A1 20221005 (EN)

Application

EP 20771681 A 20200828

Priority

- US 201962954727 P 20191230
- US 2020048401 W 20200828

Abstract (en)

[origin: WO2021137897A1] System and method for latent bias detection by artificial intelligence modeling of human decision making using time series prediction data and events data of survey participants along with personal characteristics data for the participants. A deep Bayesian model solves for a bias distribution that fits a modeled prediction distribution of time series event data and personal characteristics data to a prediction probability distribution derived by a recurrent neural network. Sets of group bias clusters are evaluated for key features of related personal characteristics. Causal graphs are defined from dependency graphs of the key features. Bias explainability is inferred by perturbation in the deep Bayesian model of a subset of features from the causal graph, determining which causal relationships are most sensitive to alter group membership of participants.

IPC 8 full level

G06N 3/04 (2006.01); **G06N 5/00** (2006.01); **G06N 5/02** (2006.01); **G06N 5/04** (2006.01); **G06N 7/00** (2006.01)

CPC (source: EP)

G06N 3/044 (2023.01); **G06N 5/045** (2013.01); **G06N 7/01** (2023.01); G06N 5/01 (2023.01); G06N 5/02 (2013.01)

Citation (examination)

US 2016170996 A1 20160616 - FRANK ARI M [IL], et al

Designated contracting state (EPC)

AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR

Designated extension state (EPC)

BA ME

DOCDB simple family (publication)

WO 2021137897 A1 20210708; CN 114902239 A 20220812; EP 4066166 A1 20221005

DOCDB simple family (application)

US 2020048401 W 20200828; CN 202080090940 A 20200828; EP 20771681 A 20200828