

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property
Organization
International Bureau

(43) International Publication Date
02 January 2025 (02.01.2025)



(10) International Publication Number
WO 2025/007117 A2

(51) International Patent Classification:
G06N 3/08 (2023.01) *G06N 20/00* (2019.01)

(21) International Application Number:
PCT/US2024/036378

(22) International Filing Date:
01 July 2024 (01.07.2024)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/523,959 29 June 2023 (29.06.2023) US

(71) Applicant: **VISA INTERNATIONAL SERVICE ASSOCIATION** [US/US]; P.O. Box 8999, San Francisco, California 94128 (US).

(72) Inventors: **LAI, Kwei-Herng**; Visa International Service Association, P.O. Box 8999, San Francisco, California 94128 (US). **YANG, Fan**; Visa International Service Association, P.O. Box 8999, San Francisco, California 94128 (US). **WANG, Lan**; Visa International Service Association, P.O. Box 8999, San Francisco, California 94128 (US). **MO-HAN, Vineeth, Rakesh**; Visa International Service Association, P.O. Box 8999, San Francisco, California 94128 (US). **YANG, Hao**; Visa International Service Association, P.O. Box 8999, San Francisco, California 94128 (US). **DAS, Mahashweta**; Visa International Service Association, P.O. Box 8999, San Francisco, California 94128 (US).

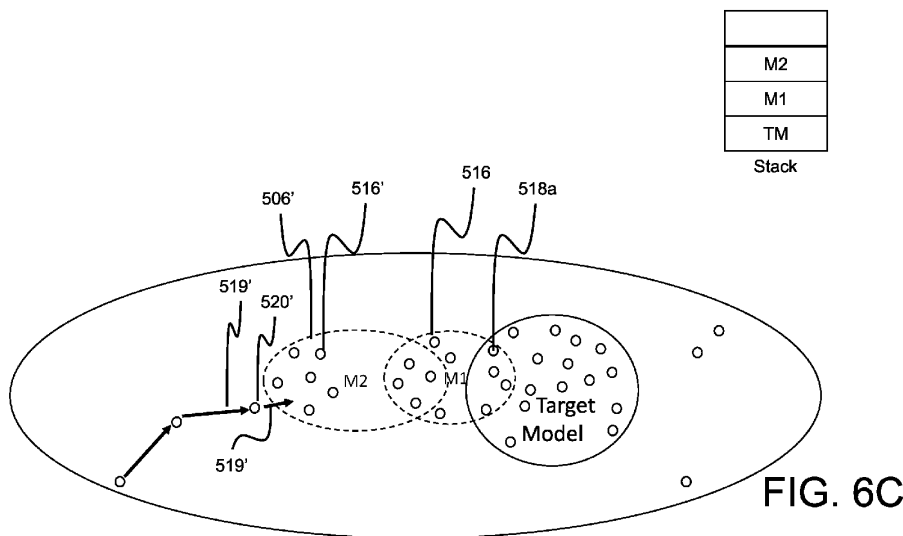
(74) Agent: **PREPELKA, Nathan, J.** et al.; The Webb Law Firm, One Gateway Center, 420 Ft. Duquesne Blvd., Suite 1200, Pittsburgh, Pennsylvania 15222 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM,

(54) Title: SYSTEM, METHOD, AND COMPUTER PROGRAM PRODUCT FOR GENERATING COUNTERFACTUAL INSTANCES



WO 2025/007117 A2



(57) Abstract: Systems, methods, and computer program products for generating counterfactual instances are provided. A system includes at least one processor configured to perturb a machine-learning model based on a plurality of input instances based on an outlier data point, determine that the perturbation based on the plurality of input instances has failed to change a result of the machine-learning model for the outlier data point, generate a plurality of query data points based on the outlier data point, create a training dataset based on each query data point of the plurality of query data points, train a candidate model derived from the machine-learning model based on the training dataset, perturb each candidate model of the plurality of candidate models based on a perturbation vector of a plurality of perturbation vectors, and generate a counterfactual output based on an output of at least one candidate model of the plurality of candidate models.

[Continued on next page]

AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*