



- (21) **International Patent Classification:**
G06N 20/00 (2019.01) *G06N 5/02* (2006.01)
G06N 5/04 (2006.01)

(21) **International Application Number:**
PCT/EP2021/071718

(22) **International Filing Date:**
04 August 2021 (04.08.2021)

(25) **Filing Language:** English

(26) **Publication Language:** English

(30) **Priority Data:**
20196394.9 16 September 2020 (16.09.2020) EP

(71) **Applicant:** **ABB SCHWEIZ AG** [CH/CH]; Brugg-
strasse 66, 5400 Baden (CH).

(72) **Inventors:** **KOTRIWALA, Arzam**; Am Galgbrunnen
4, 68526 Ladenburg (DE). **POTSCHKA, Andreas**; Am
Stadtspark 30, 38642 Goslar (DE). **KLOEPPER, Ben-**
jamin; Adlerstrasse 12, 68199 Mannheim (DE). **DIX, Mar-**
cel; Neckarpromenade 15/517, 68167 Mannheim (DE).

(74) **Agent:** **MAIWALD PATENTANWALTS- UND**
RECHTSANWALTSGESELLSCHAFT MBH; Elisen-
hof, Elisenstrasse 3, 80335 München (DE).

(81) **Designated States** (*unless otherwise indicated, for every*
kind of national protection available): AE, AG, AL, AM,
AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ,
CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO,
DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN,
HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN,
KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD,
ME, MG, MK, MN, 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, GH,

- (54) Title:** REMOVING UNDESIRABLE INFERENCES FROM A MACHINE LEARNING MODEL

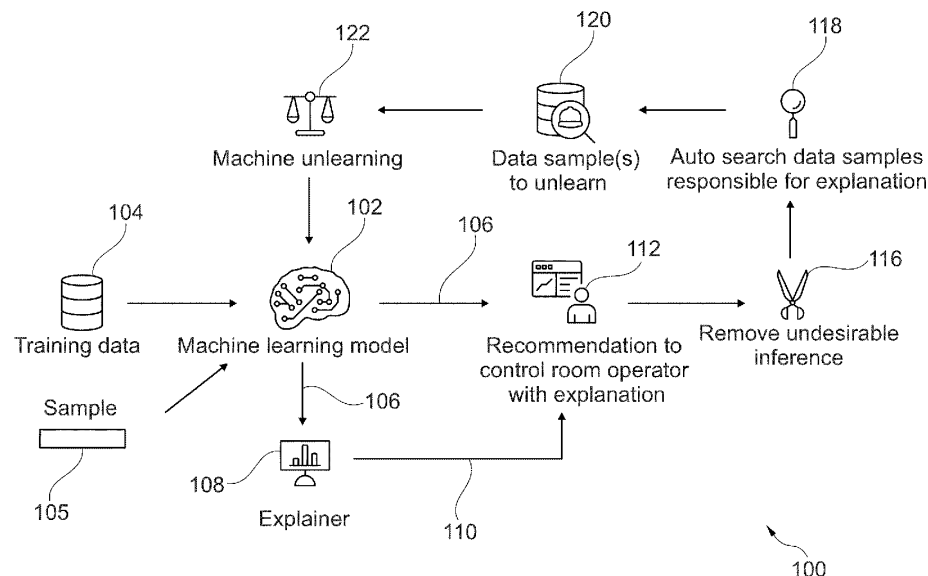


Fig. 2

- (57) Abstract:** There is provided a method and system (100) for removing undesirable inferences from a machine learning model (102). The system comprises a search component (118). The search component (118) is configured to receive a rejected explanation (110) of model output (106) provided by the machine learning model (102); identify data samples to unlearn (120) by selecting training samples from training data (104) that were used to train the machine learning model (102), the selected training samples being associated with explanations that are similar to the rejected explanation according to a calculated similarity measure; and pass the data samples to unlearn (120) to a machine unlearning unit (122).

GM, KE, LR, LS, MW, MZ, NA, RW, 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, 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).

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

Published:

— *with international search report (Art. 21(3))*