

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

(19) World Intellectual Property
Organization
International Bureau

(43) International Publication Date
10 November 2022 (10.11.2022)



(10) International Publication Number
WO 2022/235366 A1

- (51) **International Patent Classification:**
G06N 3/00 (2006.01) *G06N 20/20* (2019.01)
G06N 7/00 (2006.01)
- (21) **International Application Number:**
PCT/US2022/023522
- (22) **International Filing Date:**
05 April 2022 (05.04.2022)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
17/306,324 03 May 2021 (03.05.2021) US
- (71) **Applicant: RAYTHEON COMPANY** [US/US]; 870
Winter Street, Waltham, Massachusetts 02451-1449 (US).
- (72) **Inventors: WU, Peggy;** 2925 Southview Road, Ellicott
City, Maryland 21042 (US). **SRIVASTAVA, Kunal;** 239
Sterling Drive, Newington, Connecticut 06111 (US).
- (74) **Agent: DOYLE, David M. et al.;** Munck Wilson Mandala,
LLP, 600 Banner Place Tower, 12770 Coit Road, Dallas,
Texas 75251 (US).
- (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, JM, 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,

(54) **Title:** EXPLICIT ETHICAL MACHINES USING ANALOGOUS SCENARIOS TO PROVIDE OPERATIONAL GUARDRAILS

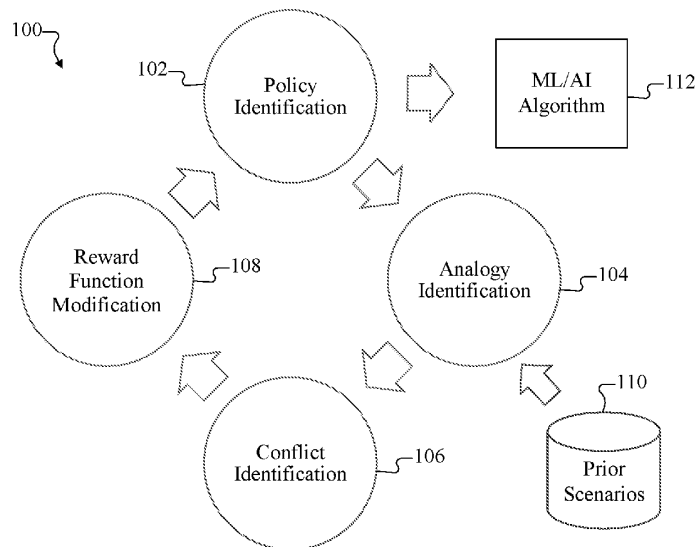


FIG. 1

(57) **Abstract:** An apparatus includes at least one memory (204, 210, 212) configured to store information associated with a current scenario (302a, 402a) to be evaluated by an ML/AI algorithm (112), where the information includes an initial reward function (302b, 402b) associated with the current scenario. The apparatus also includes at least one processor (202) configured to (i) identify one or more policies (304, 404) associated with one or more prior scenarios that are analogous to the current scenario, (ii) determine one or more differences between parameters that are optimized by the initial reward function and by one or more reward functions (306, 406) associated with the one or more prior scenarios, (iii) modify the initial reward function based on at least one of the one or more determined differences to generate a new reward function (308, 408), and (iv) generate a new policy (310, 410) for the current scenario based on the new reward function.

[Continued on next page]

WO 2022/235366 A1

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, 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))*