(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7083559号
(P7083559)

(45)発行日　令和4年6月13日(2022.6.13)

(24)登録日　令和4年6月3日(2022.6.3)

(51)国際特許分類

Ｇ０６Ｆ　21/55　(2013.01)		Ｇ０６Ｆ	21/55　３４０
Ｇ０６Ｎ　5/02　(2006.01)		Ｇ０６Ｎ	5/02　１２０

請求項の数　10　（全23頁）

(21)出願番号　特願2020-520029(P2020-520029)
(86)(22)出願日　平成30年10月10日(2018.10.10)
(65)公表番号　特表2021-500645(P2021-500645A)
(43)公表日　令和3年1月7日(2021.1.7)
(86)国際出願番号　PCT/IB2018/057830
(87)国際公開番号　WO2019/077440
(87)国際公開日　平成31年4月25日(2019.4.25)
審査請求日　令和3年3月23日(2021.3.23)
(31)優先権主張番号　15/786,888
(32)優先日　平成29年10月18日(2017.10.18)
(33)優先権主張国・地域又は機関
米国(US)

(73)特許権者　390009531
インターナショナル・ビジネス・マシーンズ・コーポレーション
ＩＮＴＥＲＮＡＴＩＯＮＡＬ　ＢＵＳＩＮＥＳＳ　ＭＡＣＨＩＮＥＳ　ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国１０５０４　ニューヨーク州　アーモンク　ニュー　オーチャード　ロード
Ｎｅｗ　Ｏｒｃｈａｒｄ　Ｒｏａｄ，　Ａｒｍｏｎｋ，　Ｎｅｗ　Ｙｏｒｋ　１０５０４，　Ｕｎｉｔｅｄ　Ｓｔａｔｅｓ　ｏｆ　Ａｍｅｒｉｃａ
(74)代理人　100112690
弁理士　太佐 種一

最終頁に続く

(54)【発明の名称】　コグニティブ仮想検出器

(57)【特許請求の範囲】
【請求項１】
敵対的な仮想対話を検出して軽減するための方法であって、
１つまたは複数のプロセッサが、
仮想エージェントと対話しているユーザによるユーザ通信を検出することと、
前記ユーザによって前記仮想エージェントとの対話中に実行される１つまたは複数のアクションに基づいて、検出された前記ユーザ通信に関連付けられたリスク・レベルを決定することと、
前記検出されたユーザ通信に関連付けられた前記決定されたリスク・レベルがリスク・レベルしきい値を超えているという決定に応答して、前記ユーザと前記仮想エージェントの間の対話に対する軽減プロトコルを開始することと、
前記仮想エージェントから前記ユーザへのより低い忠実度の応答を生成することと、を実行し、
前記軽減プロトコルが、前記ユーザによって前記仮想エージェントとの対話中に実行される前記アクションに基づいており、
前記より低い忠実度の応答が、前記検出されたユーザ通信に関連づけられた前記リスク・レベルが前記リスク・レベルしきい値を超える前の、前記仮想エージェントから前記ユーザへの元の応答の言語の精度を漸進的に希薄化した応答である、方法。
【請求項２】
前記軽減プロトコルを開始することが、

前記仮想エージェントから前記ユーザへの通信応答を変更することと、
前記ユーザと前記仮想エージェントの間の対話を、既定の対話ツリーに向けることとをさらに含み、
前記既定の対話ツリーが、前記仮想エージェントに関連付けられた秘密データを隠す応答プロトコルである、請求項１に記載の方法。
【請求項３】
前記軽減プロトコルを開始することが、
前記ユーザと前記仮想エージェントの間の対話を終了することと、
前記対話のデータを１つまたは複数の企業ネットワーク・データベースに報告することと、
前記対話のデータを格納することと、をさらに含み、
前記対話の前記データが、高リスク・レベルに関連付けられた用語を含んでいる、請求項１に記載の方法。
【請求項４】
前記軽減プロトコルを開始することが、
前記ユーザ通信に関連付けられた前記リスク・レベルが増加しているということを決定することと、
前記ユーザ通信に関連付けられた前記増加しているリスク・レベルに比例して、前記仮想エージェントから前記ユーザへの応答の期間を遅延させることと、をさらに含む、請求項１ないし３のいずれかに記載の方法。
【請求項５】
前記軽減プロトコルを開始することが、
前記ユーザの特徴を識別することであって、前記特徴が、前記仮想エージェントに対する抽出攻撃に関連付けられている、前記識別することと、
元のグランド・トゥルースを大まかに表し、かつ、攻撃者をだますことができる程度に十分近いデータを使用して、トレーニングされているハニーポット・モデルを生成することと、
前記ユーザを前記ハニーポット・モデルにリダイレクトすることと、
前記ユーザからデータを抽出することであって、前記ユーザから抽出された前記データが、専有データを抽出するために前記ユーザによって使用される手順を含んでいる、前記抽出することと、
をさらに含む、請求項１ないし４のいずれかに記載の方法。
【請求項６】
前記ユーザによって前記仮想エージェントとの対話中に実行される１つまたは複数のアクションに基づいて、前記検出されたユーザ通信に関連付けられた前記リスク・レベルを決定することが、
前記ユーザからさらに情報を取り出すためのプローブを始動することと、
前記プローブからの前記情報に基づいて、前記ユーザ通信に関連付けられた前記リスク・レベルを更新することと、
をさらに含む、請求項１ないし５のいずれかに記載の方法。
【請求項７】
１つまたは複数のプロセッサが、前記検出されたユーザ通信に関連付けられた前記決定されたリスク・レベルが前記リスク・レベルしきい値を超えているという決定に応答して、複数の前記軽減プロトコルを組み合わせて開始することをさらに実行し、
前記複数の軽減プロトコルが、前記仮想エージェントから前記ユーザへのより低い忠実度の応答を生成することと、前記ユーザと前記仮想エージェントの間の前記対話を終了することと、前記ユーザから情報を取り出すためのプローブを始動することと、前記仮想エージェントから前記ユーザへの応答の期間を遅延させることと、ハニーポット・モデルを生成することとから成る群から選択される、請求項１ないし６のいずれかに記載の方法。
【請求項８】
敵対的な仮想対話を検出して軽減するためのコンピュータ・プログラムであって、１つま

たは複数のプロセッサに、請求項１ないし７のいずれかに記載の方法の各ステップを実行させるためのコンピュータ・プログラム。
【請求項９】
請求項８に記載のコンピュータ・プログラムが記録されたコンピュータ可読記憶媒体。
【請求項１０】
敵対的な仮想対話を検出して軽減するためのコンピュータ・システムであって、
１つまたは複数のプロセッサと、
コンピュータ・プログラムを記録した１つまたは複数のコンピュータ可読記憶媒体と、
を備え、前記コンピュータ・プログラムが、請求項１ないし７のいずれかに記載の方法の各ステップを前記１つまたは複数のプロセッサの少なくとも１つに実行させるよう構成されている、コンピュータ・システム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、一般に、人工頭脳学の分野に関連しており、より詳細には、人工知能に関連している。
【背景技術】
【０００２】
人工知能では、知的エージェント（ＩＡ：intelligent agent）は、センサを介して観察し、アクチュエータ（すなわち、エージェント）を使用して環境に対して作用し、その活動を目標の達成に向ける（すなわち、経済学において定義されているように「合理的」である）、自律的実体である。知的エージェントは、目標を達成するために、知識を学習して使用することもある。知的エージェントは、極めて単純または極めて複雑であり、サーモスタットなどの反射的機械は知的エージェントである。
【０００３】
単純な反射的エージェントは、過去の他の知覚を無視し、現在の知覚のみに基づいて作動する。エージェントの機能は、条件を満たす場合にアクションを実行する、条件－アクション・ルールに基づく。エージェントの機能は、環境が完全に観察可能である場合にのみ成功する。一部の反射的エージェントは、アクチュエータがすでにトリガーされている条件をそれらが無視できるようにする、それらの現在の状態に関する情報も含むことができる。部分的に観察可能な環境内で動作している単純な反射的エージェントの場合、無限ループを避けることができないことが多い。
【０００４】
モデルに基づくエージェントは、部分的に観察可能な環境に対処することができる。その環境の現在の状態は、見ることができない世界の部分を表すある種の構造を維持するエージェント内に格納される。「世界がどのように動作するか」に関する知識は、世界のモデルと呼ばれ、そのため「モデルに基づくエージェント」という名前が付けられている。モデルに基づく反射的エージェントは、過去の知覚に依存し、それによって、現在の状態の観察されない側面の少なくとも一部を反映する、ある種の内部モデルを維持する必要がある。過去の知覚および環境に対するアクションの影響は、内部モデルを使用することによって決定され得る。次に、モデルに基づく反射的エージェントは、反射的エージェントと同じ方法でアクションを選択する。
【０００５】
目標に基づくエージェントは、「目標」情報を使用することによって、モデルに基づくエージェントの能力をさらに拡張する。目標情報は、望ましい状況を説明する。これによって、複数の可能性のうちから、目標の状態に達する可能性を選択するための方法をエージェントに与える。検索および計画立案は、エージェントの目標を達成するアクションのシーケンスを見つけることを専門とする人工知能の一分野である。目標に基づくエージェントは、決定を支援する知識が明示的に表され、変更され得るため、より柔軟である。
【０００６】

目標に基づくエージェントは、目標の状態と目標以外の状態のみを区別する。特定の状態がどの程度望ましいかの測定値を定義することが可能である。この測定値は、ある状態を、その状態の有用性の測定値にマッピングする有用性関数を使用することによって、取得され得る。より一般的な性能測定は、エージェントをどの程度満足させるかに厳密に従って、異なる世界の状態を比較できるようにする必要がある。有用性という用語は、エージェントがどの程度「満足しているか」を説明するために使用され得る。合理的な有用性に基づくエージェントは、アクションの結果の期待される有用性（すなわち、各結果の確率および有用性を前提として、エージェントが平均して得ることを期待する有用性）を最大化するアクションを選択する。有用性に基づくエージェントは、その環境をモデル化して追跡する必要があり、この作業は、知覚、表現、推論、および学習に関する大量の研究を伴っていた。

【０００７】

学習には、エージェントが最初に未知の環境内で動作することができ、初期の知識のみで可能な程度より有能になることができるという利点がある。改善を行う責任を負う「学習要素」と、外部のアクションを選択する責任を負う「実行要素」の間には、最も重量な区別がある。学習要素は、エージェントがどのように実行しているかに関する「批評者」からのフィードバックを使用し、今後より良く実行するために実行要素がどのように変更されるべきかを決定する。実行要素は、従来、エージェント全体であると考えられていたものであり、知覚を取り込んでアクションを決定する。学習エージェントの最後のコンポーネントは、「問題生成器」である。問題生成器は、新しい有益な経験につながるアクションを提案する責任を負う。

【０００８】

顧客または従業員との対話を処理するために、企業において仮想エージェントがますます配置されるようになっている。これらの仮想エージェントは、企業内のより多くの機能を引き受けるようになっているため、ますます攻撃（例えば、スパム攻撃、抽出攻撃、汚染攻撃、および回避攻撃）の対象になりつつある。

【発明の概要】

【発明が解決しようとする課題】

【０００９】

したがって、当技術分野において、前述の問題に対処する必要がある。

【課題を解決するための手段】

【００１０】

第１の態様から見ると、本発明は、敵対的な仮想対話を検出して軽減するためのコンピュータ実装方法を提供し、この方法は、１つまたは複数のプロセッサによって、仮想エージェントと対話しているユーザ通信を検出することと、１つまたは複数のプロセッサによって、検出されたユーザによって仮想エージェントとの対話中に実行される１つまたは複数のアクションに基づいて、検出されたユーザ通信に関連付けられたリスク・レベルを決定することと、検出されたユーザ通信に関連付けられた決定済みのリスク・レベルがリスク・レベルしきい値を超えているという決定に応答して、１つまたは複数のプロセッサによって、検出されたユーザと仮想エージェントの間の対話に対する軽減プロトコルを開始することとを含み、この軽減プロトコルは、検出されたユーザによって仮想エージェントとの対話中に実行されるアクションに基づく。

【００１１】

第１の態様から見ると、本発明は、敵対的な仮想対話を検出して軽減するためのコンピュータ・システムを提供し、このコンピュータ・システムは、１つまたは複数のコンピュータ・プロセッサと、１つまたは複数のコンピュータ可読記憶媒体と、１つまたは複数のコンピュータ・プロセッサのうちの少なくとも１つによって実行するための、１つまたは複数のコンピュータ可読記憶媒体に格納されたプログラム命令とを備えており、このプログラム命令が、仮想エージェントと対話しているユーザ通信を検出するためのプログラム命令と、検出されたユーザによって仮想エージェントとの対話中に実行される１つまたは複

数のアクションに基づいて、検出されたユーザ通信に関連付けられたリスク・レベルを決定するためのプログラム命令と、検出されたユーザ通信に関連付けられた決定済みのリスク・レベルがリスク・レベルしきい値を超えているという決定に応答して、検出されたユーザと仮想エージェントの間の対話に対する軽減プロトコルを開始するためのプログラム命令とを含み、この軽減プロトコルは、検出されたユーザによって仮想エージェントとの対話中に実行されるアクションに基づく。

【００１２】

さらに別の態様から見ると、本発明は、敵対的な仮想対話を検出して軽減するためのコンピュータ・プログラム製品を提供し、このコンピュータ・プログラム製品は、処理回路によって読み取り可能な、本発明のステップを実行するための方法を実行するためにこの処理回路によって実行される命令を格納している、コンピュータ可読記憶媒体を備えている。

【００１３】

さらに別の態様から見ると、本発明は、コンピュータ可読媒体に格納された、デジタル・コンピュータの内部メモリに読み込み可能なコンピュータ・プログラムを提供し、このコンピュータ・プログラムは、コンピュータ上で実行された場合に本発明のステップを実行するためのソフトウェア・コード部分を含んでいる。

【００１４】

本発明の一実施形態によれば、敵対的な仮想対話を検出して軽減するための方法が提供される。敵対的な仮想対話を検出して軽減するための方法は、仮想エージェントと対話しているユーザ通信を検出する１つまたは複数のプロセッサを含んでよい。この方法は、検出されたユーザによって仮想エージェントとの対話中に実行される１つまたは複数のアクションに基づいて、検出されたユーザ通信に関連付けられたリスク・レベルを決定する１つまたは複数のプロセッサをさらに含む。この方法は、検出されたユーザ通信に関連付けられた決定済みのリスク・レベルがリスク・レベルしきい値を超えているという決定に応答して、検出されたユーザと仮想エージェントの間の対話に対する軽減プロトコルを開始する１つまたは複数のプロセッサをさらに含み、この軽減プロトコルは、検出されたユーザによって仮想エージェントとの対話中に実行されるアクションに基づく。

【００１５】

以下では、次の図に示された好ましい実施形態を単に例として参照し、本発明が説明される。

【図面の簡単な説明】

【００１６】

【図１】本発明の実施形態に従って、分散データ処理環境を示す機能ブロック図である。

【図２】本発明の実施形態に従って、仮想エージェントとの敵対的な会話を検出して軽減するためのプログラムの動作可能なステップを示すフローチャートである。

【図３】本発明の実施形態に従って、仮想エージェントとの敵対的な会話を検出して軽減するためのプログラムの例を示す図である。

【図４】本発明の実施形態に従う、図１のサーバ・コンピュータなどのコンピュータ・システムのコンポーネントのブロック図である。

【発明を実施するための形態】

【００１７】

本発明の実施形態は、顧客または従業員との対話を処理するために、企業において仮想エージェント（例えば、仮想知的エージェント）がますます配置されるようになっているということを認識している。仮想エージェントは、ますます攻撃および悪用（例えば、帯域幅を飽和させるか、または運用経費を上昇させるためにボットによって生成されたスパム・トラフィックによる）の対象になりつつある。バックエンドの専有仮想エージェント・モデルは、専有モデルの機能をリバース・エンジニアリングするため、またはトレーニング・データから専有情報を抽出するための抽出攻撃に対して、脆弱である。実運用から継続的に学習する仮想エージェントは、汚染攻撃の対象になる。その場合、会話を脱線させるために、トレーニング・データ内の分布をシフトすることが採用される。コマンド制御

の状況において使用される仮想エージェントは、仮想エージェントを誤った方向に導くため、または欺くために、基礎になるモデルの弱点を学習して利用する攻撃者による、攻撃の対象になる。この例は、音声制御チャットボットへの隠された音声コマンドを介して、ユーザの電話の制御をこっそりと引き継ぐことである。

【００１８】

本発明の実施形態は、仮想エージェントに対する攻撃を検出するための現在の方法が、抽出攻撃、汚染攻撃、回避攻撃、および人間の行動をモデル化し、少量で行われる、より洗練されたスパム攻撃などの、意味アプリケーション・レベルの攻撃を検出するためには不十分であるということを、認識している。

【００１９】

本発明の実施形態は、３つのサブシステムを採用する方法を使用して、仮想エージェントの作業を監視する仮想エージェントと一緒に構築されたシステムを提供する。本発明の実施形態は、発話を解析して、疑わしいユーザの行動を検出するための、検出サブシステムを提供する。このシステムは、ユーザの行動を調べるための検出モデルの集合を含む。本発明の実施形態は、疑惑のレベルが上昇したときに、仮想エージェントの応答をリダイレクトするための、欺瞞サブシステムを提供する。本発明の実施形態は、情報収集プローブを介してユーザに関する学習を最大化するための、プロービング・サブシステムを提供する。敵対的なユーザの意図を明らかにする目的で、隠された対話が対話フローに投入される。

【００２０】

ここで、本発明に従って、図を参照して実施形態例が詳細に説明される。図１は、分散データ処理環境１００を示す機能ブロック図である。分散データ処理環境１００は、ネットワーク１８５を経由して相互接続されたコンピューティング・デバイス１１０およびサーバ１２０を含んでいる。

【００２１】

１つの実施形態では、コンピューティング・デバイス１１０が、グラフィカル・ユーザ・インターフェイス（ＧＵＩ：graphical user interface）１３０、Ｗｅｂブラウザ１５０、およびストレージ１６０を含んでいる。コンピューティング・デバイス１１０上のさまざまなプログラムは、Ｗｅｂブラウザ、電子メール・クライアント、セキュリティ・ソフトウェア（例えば、ファイアウォール・プログラム、地理位置情報プログラム、暗号化プログラムなど）、インスタント・メッセージ（ＩＭ：instant messaging）アプリケーション（アプリ）、および通信（例えば、電話）アプリケーションを含む。

【００２２】

コンピューティング・デバイス１１０は、デスクトップ・コンピュータ、ラップトップ・コンピュータ、タブレット・コンピュータ、専用コンピュータ・サーバ、スマートフォン、ウェアラブル・デバイス（例えば、スマート・ウォッチ、個人用フィットネス・デバイス、個人用安全装置）、または対話型ディスプレイを備える従来技術において知られた任意のプログラム可能なコンピュータ・システム、あるいは従来技術において知られた任意のその他のコンピュータ・システムであってよい。特定の実施形態では、コンピューティング・デバイス１１０は、データ・センターおよびクラウド・コンピューティング・アプリケーションにおいて一般的であるように、ネットワーク１８５を介してアクセスされたときにシームレスなリソースの単一のプールとして機能するクラスタ化されたコンピュータおよびコンポーネントを利用する、コンピュータ・システムを表す。一般に、コンピューティング・デバイス１１０は、機械可読のプログラム命令を実行し、ネットワークを介して他のコンピュータ・デバイスと通信することができる、任意のプログラム可能な電子デバイスまたはプログラム可能な電子デバイスの組み合わせを表す。

【００２３】

１つの実施形態では、グラフィカル・ユーザ・インターフェイス１３０がコンピューティング・デバイス１１０上で動作する。別の実施形態では、グラフィカル・ユーザ・インターフェイス１３０が、サーバベースの設定における別のコンピュータ上、例えば、サーバ

・コンピュータ（例えば、サーバ１２０）上で動作する。さらに別の実施形態では、グラフィカル・ユーザ・インターフェイス１３０が、ネットワーク１８５を介して相互接続されたサーバ・コンピュータ（例えば、サーバ１２０）と同時に、コンピューティング・デバイス１１０上で動作する。グラフィカル・ユーザ・インターフェイス１３０は、コンピューティング・デバイス１１０からの情報（プログラム２００によって収集または生成された情報など）にアクセスするために使用される任意のユーザ・インターフェイスであってよい。さらに、グラフィカル・ユーザ・インターフェイス１３０は、コンピューティング・デバイス１１０に情報（プログラム２００に入力するためにユーザによって提供された情報など）を提供するために使用される任意のユーザ・インターフェイスであってよい。一部の実施形態では、グラフィカル・ユーザ・インターフェイス１３０は、インターネットからリソースを取り出し、提示し、ネゴシエートするために使用される一般的Ｗｅｂブラウザを提示してよい。他の実施形態では、グラフィカル・ユーザ・インターフェイス１３０は、ユーザがコンピューティング・デバイス１１０でネットワーク１８５にアクセスできるようにする、ソフトウェアまたはアプリケーションであってよい。

【００２４】

さらに別の実施形態では、コンピューティング・デバイス１１０のユーザは、グラフィカル・ユーザ・インターフェイス（ＧＵＩ）への入力デバイスとして、かつソフトウェア・アプリケーションに関連付けられた複数のアイコンまたは実行中のソフトウェア・アプリケーションを描画する画像を提示する出力デバイス（すなわち、電子ディスプレイ）として働くタッチスクリーンを介して、グラフィカル・ユーザ・インターフェイス１３０と対話することができる。任意選択的に、ソフトウェア・アプリケーション（例えば、Ｗｅｂブラウザ）は、コンピューティング・デバイス１１０のＧＵＩ内で動作するグラフィカル・ユーザ・インターフェイス１３０を生成できる。グラフィカル・ユーザ・インターフェイス１３０は、マルチタッチ・ディスプレイと呼ばれる触覚センサ・インターフェイス（例えば、タッチスクリーンまたはタッチパッド）を含むが、これに限定されない、複数の入出力（Ｉ／Ｏ：input/output）デバイスから入力を受け取る。グラフィカル・ユーザ・インターフェイス１３０とインターフェイスするＩ／Ｏデバイスは、コンピューティング・デバイス１１０に接続されてよく、コンピューティング・デバイス１１０は、有線ネットワーク通信（例えば、ＵＳＢポート）または無線ネットワーク通信（例えば、赤外線、ＮＦＣなど）を利用して動作してよい。コンピューティング・デバイス１１０は、本発明の実施形態に従って、図４に関してさらに詳細に示され、説明されるように、コンポーネントを含んでよい。

【００２５】

Ｗｅｂブラウザ１５０は、インターネットから情報リソースを取り出し、提示し、トラバースするために使用される一般的Ｗｅｂブラウザであってよい。一部の実施形態では、Ｗｅｂブラウザ１５０は、モバイル・デバイス用に設計されたＷｅｂブラウザであってよい。他の実施形態では、Ｗｅｂブラウザ１５０は、デスクトップ・コンピュータ、ＰＣ、またはラップトップなどの従来型のコンピューティング・デバイス用に設計されたＷｅｂブラウザであってよい。一般に、Ｗｅｂブラウザ１５０は、コンピューティング・デバイス１１０のユーザがネットワーク１８５を経由してＷｅｂページにアクセスできるようにする任意のアプリケーションまたはソフトウェアであってよい。示されている環境では、Ｗｅｂブラウザ１５０がコンピューティング・デバイス１１０上に存在する。他の実施形態では、Ｗｅｂブラウザ１５０または同様のＷｅｂブラウザが、ネットワーク１８５を経由してＷｅｂページにアクセスできる他のコンピューティング・デバイス上に存在してよい。

【００２６】

コンピューティング・デバイス１１０上にあるストレージ１６０（例えば、データベース）は、コンピューティング・デバイス１１０によってアクセスされて利用されるデータを格納できる、任意の種類のストレージ・デバイスを表す。他の実施形態では、ストレージ１６０が、コンピューティング・デバイス１１０内の複数のストレージ・デバイスを表す。ストレージ１６０は、アカウント情報、認証のための認証情報、ユーザの嗜好、好まし

いユーザのリスト、以前に訪問されたＷｅｂサイト、訪問されたＷｉ－Ｆｉポータルの履歴、およびコンピューティング・デバイスの位置の履歴などの、ただしこれらに限定されない、情報を格納する。

【００２７】

一般に、ネットワーク１８５は、コンピューティング・デバイス１１０間の通信をサポートする接続およびプロトコルの任意の組み合わせであることができる。例えば、ネットワーク１８５は、ローカル・エリア・ネットワーク（ＬＡＮ：local area network）、インターネットなどの広域ネットワーク（ＷＡＮ：wide area network）、セルラー・ネットワーク、またはこれらの任意の組み合わせを含むことができ、有線接続、無線接続、または光ファイバ接続、あるいはその組み合わせをさらに含むことができる。

【００２８】

サーバ１２０は、デスクトップ・コンピュータ、ラップトップ・コンピュータ、タブレット・コンピュータ、専用コンピュータ・サーバ、スマートフォン、または従来技術において知られた任意のその他のコンピュータ・システムであってよい。特定の実施形態では、サーバ１２０は、データ・センターおよびクラウド・コンピューティング・アプリケーションにおいて一般的であるように、ネットワーク１８５を介してアクセスされたときにシームレスなリソースの単一のプールとして機能するクラスタ化されたコンピュータおよびコンポーネントを利用する、コンピュータ・システムを表す。一般に、サーバ１２０は、機械可読のプログラム命令を実行し、ネットワークを介して他のコンピュータ・デバイスと通信することができる、任意のプログラム可能な電子デバイスまたはプログラム可能な電子デバイスの組み合わせを表す。１つの実施形態では、サーバ１２０は、データベース１７０およびプログラム２００を含んでいる。

【００２９】

１つの実施形態では、サーバ１２０は、サーバ１２０とコンピューティング・デバイス１１０の間でハンドシェイク・プロセスを開始できる。ハンドシェイクは、チャネルを経由する通常の通信が開始する前に２つの実体間で確立される通信チャネルのパラメータを動的に設定する、ネゴシエーションの自動化されたプロセスである。ハンドシェイクは、チャネルの物理的確立に従い、通常の情報の転送に先行する。ハンドシェイクは、パラメータを設定するためにユーザの介入を必要とせず、通信チャネルを経由した異種のコンピューティング・システムまたは機器の接続を容易にする。１つの例では、サーバ１２０は、コンピューティング・デバイス１１０上のプログラムにアクセスするために、サーバ１２０が通信チャネルを確立したいということを示すメッセージをコンピューティング・デバイス１１０に送信することによって、ハンドシェイク・プロセスを開始する。

【００３０】

データベース１７０は、サーバ１２０によって読み取られてよい保存場所であってよい。データベース１７０は、サーバ１２０によってアクセスされて利用されるデータを格納できる任意の種類のストレージ・デバイスを表す。他の実施形態では、データベース１７０が、サーバ１２０内の複数のストレージ・デバイスを表す。データベース１７０は、アカウント情報、認証のための認証情報、ユーザの嗜好、好ましいユーザのリスト、以前に訪問されたＷｅｂサイト、訪問されたＷｉ－Ｆｉポータルの履歴、およびコンピューティング・デバイスの履歴、ならびにサーバにアクセスするコンピューティング・デバイス上にある情報などの、ただしこれらに限定されない、情報を格納する。

【００３１】

仮想エージェント・プログラム１７５は、サーバ１２０上のプログラムである。一実施形態では、仮想エージェント・プログラム１７５は、オンライン顧客サービス担当者の役割を果たす、擬人化された外観を有するアニメーション化された人工知能の仮想キャラクターである。１つの例では、仮想エージェント・プログラム１７５は、業種および領域のコンテンツを使用して事前にトレーニングされているため、ユーザが行おうとする会話の多くをすでに理解している。仮想エージェント・プログラム１７５は、コグニティブ技術を適用して、トレーニング済みの業種および領域の知識を使用して、個人向けの状況に応じ

た顧客体験を提供する。別の実施形態では、仮想エージェント・プログラム１７５は、ユーザとの知的な会話を導き、ユーザの質問に応答し、適切な非言語行動を実行する。仮想エージェント・プログラム１７５は、関与の指標を使用して、仮想エージェント・プログラム１７５が１人または複数のユーザと行う会話について、さらに理解することができる。別の実施形態では、仮想エージェント・プログラム１７５は、ホテルのロビー、レストラン、または事務所の受付で顧客を迎えるホログラフィック・プロジェクションとして現れてもよい。

【００３２】

検出器１８０は、プログラム２００のサブ・プログラムであり、ユーザによって仮想エージェント・プログラム１７５に発行された要求と、仮想エージェント・プログラム１７５による応答とを傍受し、プログラム２００の対話異常検出サブシステムとして機能する。一実施形態では、検出器１８０は、仮想エージェント・プログラム１７５によるログ・エントリの要求および応答を抽出し、検出器１８０は、異常検出モデルの集合を生成するために、ログ・エントリを入力として使用する。この実施形態では、検出器１８０は、異なる検出戦略に固有の検出モデルを生成することができる。１つの例では、検出器１８０が、自然言語を解析するためのマルコフ・モデル、モデルへの照会の分布を追跡して、高リスクのレベルの情報公開を決定するための情報漏洩モデル、時間タグを検査して、人間であることが疑わしいほど速い回答を検出するためのタイミング・モデル、認識された意図での低信頼度スコアの発生を警告する信頼度モデル、およびユーザとの対話における特定の対象の状態の発生などの対話進行標識の有無を識別する対話進行モデル、といった検出モデルの各々を生成してマージする。検出器１８０は、個々の検出モデルを、応答の疑わしさの決定を支援するように重み付けされた単一のリスク・スコアに結合する。

【００３３】

ボット・シールド・データベース（Bot shield database）１８２は、プログラム２００に存在し、プログラム２００によって排他的に使用されるデータベースである。ボット・シールド・データベース１８２は、プログラム２００によってアクセスされて利用されるデータを格納できる任意の種類のストレージ・デバイスを表す。他の実施形態では、ボット・シールド・データベース１８２が、プログラム２００内の複数のストレージ・デバイスを表す。ボット・シールド・データベース１８２は、コンテキスト情報、アカウント情報、認証のための認証情報、ユーザの嗜好、および好ましいユーザのリストなどの、ただしこれらに限定されない、情報を格納する。例えば、ボット・シールド・データベース１８２は、単語が疑わしい活動に関連するということを示唆する高リスク・レベルに関連付けられた単語および語句を格納する。ボット・シールド・データベース１８２は、個々の検出モデルからのデータの履歴を格納する。例えば、ボット・シールド・データベース１８２は、高リスクのユーザ、中リスクのユーザ、および低リスクのユーザに関連付けられた、タイミング検出モデルに対する応答のタイミング・パターンの１つまたは複数の履歴を格納する。別の例では、ボット・シールド・データベース１８２は、マルコフ検出モデルからの情報を格納する。この例では、ボット・シールド・データベース１８２は、不適切としてフラグ付けされた、仮想エージェントとユーザの間のありそうにない対話の履歴情報を格納する。ボット・シールド・データベース１８２は、ユーザと仮想エージェントの間の対話状態からの推移頻度の履歴を格納する。

【００３４】

別の実施形態では、ボット・シールド・データベース１８２が、個々の異常検出モデル、および１つまたは複数の配置の異常検出モデルの集合での異常の検出における有効性の履歴情報を格納する。別の実施形態では、ボット・シールド・データベース１８２が、異常検出サブシステムにおける追加の検出モデルの拡張および更新の履歴情報、軽減システムにおける軽減応答、およびプロービング・システムにおけるプローブ選択を格納する。１つの例では、プログラム２００が、新しい回避マルウェアの更新情報、および仮想エージェントに対する新しい攻撃を正常に軽減できる更新された防御を受信する。ボット・シールド・データベース１８２は、新しい回避マルウェアの情報、およびマルウェアに対する

更新された防御を格納する。
【００３５】
欺瞞エンジン１９０は、仮想エージェント・プログラム１７５によるユーザへの応答の忠実度を自動的に調整し、可能性のある攻撃を防ぐための、プログラム２００のサブ・プログラムである。欺瞞エンジン１９０は、ユーザとの元の対話フローを変更せずに、モデルの応答の忠実度または正確さを変更することによって、攻撃を軽減する。欺瞞エンジン１９０は、現在のユーザのリスク・スコアに従って応答の忠実度レベルを選択することによって、ユーザに提供されるモデルの応答の正確さを変更する。ユーザのリスクが高いほど、高リスクのユーザに提供されるモデルの応答の正確さが低くなる。例えば、欺瞞エンジン１９０が応答の忠実度を変更する前のモデルの応答が「クレジット・カード番号＃＃＃－＃＃＃＃－＃＃＃＃を入力したことを確認してください」である場合、欺瞞エンジン１９０は、ユーザによる高リスクの活動に応答して、ユーザのリスク・レベルと一致するように応答の忠実度を変更する。この例では、欺瞞エンジン１９０は、モデルの応答の忠実度を「確認のためにクレジット・カード番号を再入力してください」に変更する。ユーザのリスク・スコアが特定の事前に定義されたしきい値を超えた場合に、欺瞞エンジン１９０がトリガーされる。一実施形態では、欺瞞エンジン１９０が、仮想エージェント・プログラム１７５に対するユーザの応答に基づくユーザの特定のリスク・レベルおよび特定のしきい値を超えたという事実に基づいて、軽減アクションをトリガーする。
【００３６】
別の実施形態では、欺瞞エンジン１９０が、前述したようにモデルの応答の忠実度を変更することによって対話フローを安全な対話にリダイレクトすることに加えて、仮想エージェント・プログラム１７５に対する攻撃を軽減することができる。一実施形態では、欺瞞エンジン１９０が、漸進的なモデルの希薄化などの複数の戦略を使用して、より低い忠実度の応答を作成する。１つの例では、欺瞞エンジン１９０が、以前にトレーニングされたモデルを希薄化されたモデルの基準（ベースライン）として使用する。欺瞞エンジン１９０は、以前にトレーニングされたモデルを、元のモデルにおける真の基準（ベースライン・トゥルース）の不正確なバージョンにする。別の実施形態では、欺瞞エンジン１９０が、ランダムな誤った応答を徐々に挿入する。１つの例では、攻撃者によって収集される統計データを混乱させるために、欺瞞エンジン１９０が、時々、ランダムに選ばれた不正確な応答をユーザに返す。
【００３７】
別の実施形態では、欺瞞エンジン１９０が、ユーザをハニーポット・モデルにリダイレクトする。この例では、欺瞞エンジン１９０が、元のモデルの機能を模倣するモデルを使用するが、このモデルは、元の真の基準（ベースライン・トゥルース）を大まかに表す、ただし攻撃者をだますことができる程度に十分似ているデータを使用して、トレーニングされている。欺瞞エンジン１９０による欺瞞的応答は、攻撃者においてすでに抽出された情報を無効化するのに役立つことができる。別の例では、欺瞞エンジン１９０が、ユーザとの元の対話フローを変更せずに、モデルの応答の忠実度を変更する。この例では、欺瞞エンジン１９０が、仮想敵対者における情報の蓄積を遅くするか、または中断させる。別の例では、欺瞞エンジン１９０が、会話を人間の応答者に拡大する。この例では、欺瞞エンジン１９０が、ユーザのリスク・スコアに基づいて、人間の応答者に会話に介入させるための通知を開始する。
【００３８】
プローブ１９５は、コンピュータと人間を区別する完全に自動化された公開チューリング・テスト（Ｃａｐｔｃｈａ：Completely Automated Public Turing test to tell Computers and Humans Apart）の隠された対話の一般化を使用する、プログラム２００のサブ・プログラムである。Ｃａｐｔｃｈａは、ユーザが人間であるかどうかを判定するための計算において使用されるチャレンジ／レスポンス型テストの一種である。一実施形態では、プローブ１９５が、仮想エージェント・プログラム１７５を介してプローブをユーザに送信し、プローブの応答を評価し、この応答を考慮してユーザのリスク・スコアを更

新する。１つの例では、プローブ１９５が、時々、ユーザのリスク・スコアに基づいてプローブを投入する。ユーザのリスク・スコアが、低リスクとある程度のリスクの間の境界にある場合、プローブ１９５は、仮想エージェント・プログラム１７５とユーザの会話に介入し、ユーザが人間であることを探ることができる。プローブ１９５は、応答から収集されたデータおよびプローブ１９５から得られた情報を、ボット・シールド・データベース１８２に直接追加する。

【００３９】

１つの実施形態では、プログラム２００がサーバ１２０上で動作する。別の実施形態では、プログラム２００が、サーバベースの設定における別のコンピュータ上、例えば、サーバ・コンピュータ（図示されていない）上で動作する。さらに別の実施形態では、プログラム２００が、ネットワーク１８５を介して相互接続されたサーバ１２０と同時に、コンピューティング・デバイス１１０上で動作する。プログラム２００は、仮想エージェントとの敵対的な会話を検出して軽減するための能力を提供する。プログラム２００は、コンピューティング・デバイス１１０と通信するために、Ｗｉ－Ｆｉ技術、Ｂｌｕｅｔｏｏｔｈ、近距離無線通信（ＮＦＣ：Near Field Communication）タグ、グローバル・システム・フォー・モバイル・コミュニケーションズ（ＧＳＭ：Global System for Mobile Communications）、および全地球測位システム（ＧＰＳ：Global Positioning System）技術を利用できる。

【００４０】

一実施形態例では、プログラム２００が、コンピューティング・デバイス１１０上の１つまたは複数のアプリケーション内のコード・スニペットとして動作する。コード・スニペットは、スニペットとアプリケーション（例えば、サーバ１２０上のＷｅｂブラウザ・アプリケーションによってホストされるプログラム２００）の間の対話性の範囲を定義する。例えば、プログラム２００は、Ｗｅｂブラウザ１５０内の機能であり、プログラム２００のプロセスは、プログラム２００によって開始されたときに、Ｗｅｂブラウザ１５０の動作中に自動的に（すなわち、ユーザの介入なしで）発生する。動的なコード・スニペット要素は、スクリプトのサポートを提供する。変数が、サーバ１２０を介したプログラム２００、グラフィカル・ユーザ・インターフェイス１３０、Ｗｅｂブラウザ１５０、および仮想エージェント・プログラム１７５の間の対話を可能にする。

【００４１】

一実施形態では、プログラム２００が、会話のセキュリティを実現するために仮想エージェント・プログラム１７５の対話システムとインターフェイスすることができる独立した異常検出システムとして実装され得る。プログラム２００は、会話による先行するモデルの照会を介して、会話の文脈を活用することによって、異常な疑わしい会話を検出する。１つの例では、プログラム２００が、会話ログに対して動作する監視機能として、仮想エージェントのプラグインとして動作できる。異常検出サブシステムは、操作ダッシュボードに異常監視結果を供給するためのスタンドアロン・プログラムとして使用できる。この例では、欺瞞エンジン１９０およびプローブ１９５がユーザとの会話フローを操作するときに、欺瞞エンジン１９０およびプローブ１９５が対話ランタイムと統合されるか、または連携する。各サブシステムは拡張可能であり、さまざまな攻撃者との遭遇から学習することができる。拡張可能であるということは、追加の検出モデルをプログラム２００に実装することができ、追加の軽減応答をプログラム２００に追加することができ、追加のプローブ選択をプログラム２００に追加することができるということを意味する。

【００４２】

別の実施形態では、プログラム２００が、モデルのアプリケーション・プログラム・インターフェイス（ＡＰＩ：application program interface）レベルで異常を監視して検出するためのモデルのセキュリティとして、機能する。プログラム２００は、ユーザまたは組織の嗜好に従って、モデル固有の検出を実現することができる。

【００４３】

一実施形態では、プログラム２００は、ユーザと仮想エージェント・プログラム１７５の

遭遇に関して、中リスクの値を検出する。１つの例では、プログラム２００は、検出器１８０、欺瞞エンジン１９０、およびプローブ１９５を利用して、ユーザが、プログラム２００による軽減のしきい値をすれすれで超えたことを決定する。この例では、プログラム２００は、ボット・シールド・データベース１８２を調べることに基づいて、どの軽減手順が疑わしいユーザに対する最良の応答かを決定する。プログラム２００は、類似するユーザとの対話の履歴を解析し、良いユーザを抑止してしまうリスクに対して特定のプローブを採用することの強度および有用性に、重み付けする。

【００４４】

図２は、本発明の実施形態に従って、仮想エージェントとの敵対的な会話を検出して軽減するためのプログラム２００を示すフローチャートである。

【００４５】

ステップ２１０で、プログラム２００が対話のリスク値を決定する。１つの実施形態では、プログラム２００が、検出器１８０を使用して、ユーザと仮想エージェントの間の会話を監視しながら、ユーザからの１つまたは複数の発話を解析する。実施形態例では、プログラム２００が、各仮想エージェントの応答を傍受し、仮想エージェントの応答のログ・エントリを抽出する。プログラム２００は、仮想エージェントの応答のログ・エントリを、「Ｎ」個の異常検出モデルの集合を呼び出すための入力として使用し、各モデルが異なる異常検出戦略を実装する。集合内の個々の検出モデルは、同時に実行されて、個別にリスク値を計算し、このリスク値が、重み付き集合関数を使用して集合のリスク・スコアにマージされる。この集合は、ありそうにない対話（起こりそうにない推移）を検出するためのマルコフ・モデル、時間タグを検査して人間であることが疑わしいほど速い回答を検出するためのタイミング・モデル、認識された意図での低信頼度スコアの発生を警告する信頼度監視モデル、および仮想エージェント・プログラム１７５との対話の文脈に基づいて対話のリスク・スコア値を高めることがある対話における特定の対象の状態（例えば、「何かを販売する」という語句）の発生などの、対話進行標識の有無を識別する対話進行モデルのうちの１つまたは複数、およびその他のモデルを含んでよい。マルコフ・モデルは、対話状態からの推移頻度を使用して対話フロー・グラフから構築することができ、情報漏洩追跡モデルは、モデルの特徴空間内のモデルへの照会の分布を追跡して、情報公開のリスクレベル（すなわち、公開される情報が、敵対者によってモデルの機能を複製するのに十分であるか）を決定する。

【００４６】

１つの例では、プログラム２００が、マルコフ検出モジュールを個別に利用して、ユーザのリスク値を計算する。この例では、仮想エージェントへの対話ログ・エントリが既知の攻撃の特徴であるため、プログラム２００がユーザに対して高リスク値を計算する。このユーザによる対話ログ・エントリは、既知の攻撃のパターンに従っているため、侵入的であると見なされる。プログラム２００は、既知の攻撃に対するログ・エントリの類似性を使用して、ユーザに対して高リスク値を計算する。一実施形態では、プログラム２００が、重み付き集合関数を使用して、個々の検出モデルからのリスク・スコアを単一のリスク・スコアＲにマージする。この関数の重みは、時間と共に適応されてよい。プログラム２００は、マージされたリスク値を使用して、仮想エージェントにおけるユーザのリスク・スコアを更新する。プログラム２００は、対話ログ・エントリを使用して、集合内のすべての異常検出モデルを増加的に更新する。１つの例では、プログラム２００が、対話ログ・エントリを受信し、２つの異常検出モデルの組み合わせを利用してユーザのリスク・スコアを計算する。プログラム２００は、タイミング異常検出モデルを利用して、ユーザによる応答時間が既知の攻撃者に関連付けられたタイミング・パターンと一致するということを決定する。プログラム２００は、タイミング異常検出モデルに基づいて個々のリスク・スコア「ｒ１」を割り当てる。プログラム２００は、対話進行異常検出モデルを利用して、仮想エージェントへの対話ログ・エントリが既知の攻撃者の対話進行パターンと一致するということを決定する。プログラム２００は、対話進行異常モデルに基づいて個々のリスク・スコア「ｒ２」を割り当てる。プログラム２００は、リスク・スコア「ｒ１」お

よびリスク・スコア「ｒ２」を結合して、結合されたリスク・スコア値「Ｒ＝ｆ（ｒ１，ｒ２）」を計算し、「ｆ」は重み付き集合関数である。

【００４７】

図３は、本発明の実施形態に従って、仮想エージェントに対する攻撃を検出して軽減するように動作するプログラム２００の例を示している。この実施形態では、プログラム２００が、サーバ１２０上で動作し、領域モデル３２０内の極秘データに対してトレーニングされた秘密情報へのアクセスを許可または可能にするために、盗まれたモデル３１０（敵対者３１５）との仮想看護師３７５の対話を監視する。この実施形態では、盗まれたモデル３１０は、秘密診断情報である。敵対者３１５は、仮想看護師との対話を開始して続行しているユーザである。プログラム２００は、敵対者３１５と仮想看護師の間の対話を保護するボット・シールド防御（bot shield defense）である。領域モデル３２０は、患者に固有の秘密診断情報である。仮想看護師３７５は、患者と会話して、アップロードされた画像を含む症状の説明を収集する、仮想エージェント・チャットボットのフロントエンドである。仮想看護師３７５は、診断医療モデルを使用して、領域モデル３２０からの秘密の最終診断の応答を患者に提供する。

【００４８】

１つの例では、図３に示されているように、プログラム２００が、敵対者３１５によって仮想看護師３７５に入力された発話を解析する。この例では、プログラム２００が、領域モデル３２０を探査し、領域モデル３２０のリバース・エンジニアリングに関する情報を抽出する。プログラム２００は、盗まれたモデル３１０が抽出攻撃を介して生成されており、盗まれたモデル３１０が、領域モデル３２０のトレーニング・データから機密情報を抽出しようとしているということを決定する。プログラム２００は、前に説明された１つまたは複数の異常検出モデルを使用して、仮想看護師３７５が攻撃を受けていることを確認する。例えば、プログラム２００が、仮想看護師３７５をだまそうとしている敵対者３１５からの入力を検出したため、プログラム２００は、敵対者３１５が回避攻撃を使用していることを識別する対話進行モデルを使用する。

【００４９】

判定ステップ２２０で、プログラム２００が、リスク値がしきい値を超えているかどうかを判定する。１つの実施形態では、プログラム２００が、ボット・シールド・データベース１８２を介してアクセスされる発話の履歴に基づいて、現在検出されている発話と比較して、リスク・スコアがしきい値を超えたことを（ステップ２１０から）決定する。プログラム２００は、ステップ２１０で計算された値に基づき、集合のリスク・スコアにマージされてリスク値「Ｒ」を生成するように計算されたリスク値に応じて、リスク値がしきい値を超えたかどうかを判定する。

【００５０】

１つの例では、図３に関して、（ステップ２１０で決定された）プログラム２００が同時に動作して、敵対者３１５と仮想看護師３７５の間の会話のセキュリティ・レベルと、盗まれたモデル３１０と領域モデル３２０の間のＡＰＩレベルでのモデルのセキュリティ・レベルとで、複数のレベルの検出を提供する。プログラム２００は、この例では組織によって決定されたしきい値を超えるスコア「Ｒ」を評価する。領域モデル３２０は、極秘データに対してトレーニングされており、秘密情報を含んでいる。プログラム２００は、敵対的な入力に基づいて、仮想看護師３７５に応答してプログラム２００によって決定されたスコア「Ｒ」を敵対者３１５にも割り当てる。

【００５１】

１つの例では、仮想看護師３７５のホストが、３／１０の組織のリスク・スコアしきい値を設定する。仮想看護師３７５が極秘健康情報（領域モデル３２０）を保護するため、組織のリスク・スコアは、極秘データを保護していない組織のしきい値と比較して、低く設定される。１０が最高のリスク・スコアであり、１が最低のリスク・スコアである。プログラム２００は、個々の値に基づいて、高リスク値Ｒ（「Ｒ＝ｆ（ｒ１，ｒ２）」）を割り当て、「ｆ」は、個々の異常検出モデルを使用して計算される重み付き集合関数である

。この例では、攻撃者であるという高いリスクを敵対者３１５が示しているということが、決定される。この例では、プログラム２００が、ボット・シールド・データベース１８２からの履歴情報を介して、会話の異常な文脈に基づいて、対話進行モデルのスコア「ｒ１」を割り当てた。仮想看護師３７５は、健康情報および医療データに対してトレーニングされており、敵対者３１５が、病気および治療の情報などの個人情報に関連する対話を開始している。プログラム２００は、タイミング異常検出モデル「ｒ２」によって評価されて計算されたときに、敵対者３１５が過度に速い応答を仮想看護師３７５の基本的質問に対して提供しているということを決定する。プログラム２００は、高リスク値「Ｒ」＝１０／１０を計算し、１０は、高リスクの対話に対する組織のしきい値（３）を超えている。

【００５２】

ステップ２３０で、プログラム２００がアクセスを許可する。さらに具体的には、リスク・スコアがしきい値を超えていないということの決定（判定ステップ２２０の「いいえ」の分岐）に応答して、プログラム２００が仮想エージェント・プログラム１７５に対するアクセスを許可する（ステップ２３０）。この例では、プログラム２００が、ステップ２１０で決定されたリスク値がしきい値「Ｒ」を満たしていないということを決定する。１つの例では、プログラム２００が、ユーザによる仮想エージェント・プログラム１７５への各発話を解析し、応答が人間によるものであるということを決定する。プログラム２００は、ボット・シールド・データベース１８２にアクセスし、高い信頼度で人間であるということが決定された類似する発話の履歴を調査する。発話の正確さおよび許容できる応答の履歴との応答の一致に基づいて、プログラム２００は、仮想プログラム１７５に対するユーザのアクセスを許可する。

【００５３】

別の実施形態では、ステップ２２０の結果として、プログラム２００が、ユーザのリスク・スコアが低いということを決定する。１つの例では、プログラム２００が、ユーザに対して１／１０のリスク・スコアを計算し、ユーザによって示されたリスクに基づいてアクションを開始する組織のしきい値が、５／１０である。この例では、プログラム２００が、ユーザおよび仮想エージェントが会話を続行することを許可する。プログラム２００は、ユーザに対して低リスク値を計算した後に、ユーザと仮想エージェントの間での途切れず、中断されない会話を許可しながら、監視モードに移行する。プログラム２００は、監視モードでの会話の監視を続行し、仮想エージェントとのさらなる対話に基づいてユーザのリスク・スコアを再計算することができる。プログラム２００が、会話の文脈、または発話の過去の侵入的パターンの出現に基づいて、リスク・スコアを再び高めた場合、プログラム２００は、再計算されたリスク・スコアに応答して軽減アクションを開始できる。

【００５４】

ステップ２４０で、プログラム２００が、可能性のある損害を防ぐ。さらに具体的には、リスク値がしきい値を超えたということの決定（判定ステップ２２０の「はい」の分岐）に応答して、プログラム２００が異常軽減サブシステムを開始する（ステップ２４０）。実施形態例では、異常軽減サブシステムは、プログラム２００が「良い」ユーザまたは人間のユーザを抑止することがあるリスクと比較した、抑止の一形態をユーザに提供することの強度および有用性の解析に基づいて、可能性のある損害を防ぎ、可能性のある損害を防ぐための可能性のある経路を選択するように動作する。

【００５５】

１つの実施形態では、プログラム２００が、決定された特定のリスク・レベル、および決定されたリスク・レベルに応じて超えられた特定のしきい値に基づいて、欺瞞エンジン１９０およびプローブ１９５を介して軽減アクションを始動する。プログラム２００は、対話フローを変更し、ユーザとの会話を、以前に決定された対話ツリーの安全な領域にリダイレクトする。プログラム２００は、Ｒ値に対して、仮想看護師３７５からの応答の忠実度を調整できる。１つの例では、仮想看護師３７５は、ユーザへの仮想看護師３７５の応答に関して、２つの忠実度レベルを有する。仮想看護師３７５は、プログラム２００を介

して、ユーザの質問に対する元の高忠実度の仮想看護師３７５の応答を開始するか、または仮想看護師３７５は、ユーザに対する低忠実度の応答を開始して、ユーザによる高リスクの応答を軽減する。プログラム２００は、「私はこれに対してトレーニングされていません～詳細については、１－８００…に電話してください」と述べることによって会話を終了する軽減アクションを始動する。プログラム２００は、ステップ２１０で割り当てられた「Ｒ」値に基づいて、ユーザのリスク・スコアが「高リスク」のしきい値を超えるたびに介入する。プログラム２００は、対話フローを変更し、仮想看護師３７５の応答を以前に生成された低忠実度の応答にリダイレクトして、敵対者３１５と仮想看護師３７５の間の対話を軽減する。別の例では、プログラム２００が、プローブに対するユーザの応答または計算されたリスク・スコアに従って、ユーザへの応答を遅延させる。この例では、仮想看護師３７５に対するユーザの発話が、高リスクの発話の既知のパターンに次第に似てくる。プログラム２００は、ユーザのリスク・スコアを、各高リスクの応答に比例して増やす。リスク・スコアが高くなるため、プログラム２００は、応答をユーザに返送する前に、より長い遅延を導入する。

【００５６】

別の実施形態では、プログラム２００が、仮想看護師３７５の元の対話フローを変更せずに、仮想看護師３７５によるモデルの応答の忠実度を変更することによって、可能性のある攻撃を軽減して防ぐことができる。プログラム２００は、仮想敵対者の情報の蓄積を遅くするか、または中断させるように、仮想看護師３７５によるモデルの応答の忠実度を変更する。プログラム２００は、各保護されたモデルを、より低い忠実度のモデルの集合内に挿入する。さまざまな実施形態では、プログラム２００が、現在のユーザのリスク・スコアに従う忠実度のモデルを選択することによって、実際のモデルの応答を決定する。ステップ２１０で決定されるユーザのリスクが高いほど、応答のレベルが低くなる。１つの例では、リスク・スコアＲに応じて忠実度レベルＦが決定される。忠実度レベル１は最高のレベルであり、仮想エージェント・プログラム１７５による、低リスク・スコアを有する人間のユーザに対する元の応答と一致し、忠実度レベル２はより低く、忠実度レベル３はさらに低く、最大の忠実度レベルは、組織の嗜好またはユーザの嗜好によって決定されるＮである。

【００５７】

この別の実施形態では、プログラム２００が、認識された攻撃者に対するより低い忠実度の応答をもたらす追加のモデル作成することによって、より低い忠実度の応答を生成できる。プログラム２００は、漸進的なモデルの希薄化を使用できる。モデルの希薄化は、以前にトレーニングされたモデルを希薄化されたモデルのグランド・トゥルースとして使用する方法である。その結果、プログラム２００は、欺瞞エンジン１９０を介して、以前にトレーニングされたモデルを、元のモデルにおける真の基準（ベースライン・トゥルース）の不正確なバージョンにする。プログラム２００は、各低忠実度の応答を、徐々に低い忠実度の応答に、無限につなぐことができる。

【００５８】

追加の実施形態例では、図３に関して、リスク値がしきい値を超えたことの決定（判定ステップ２２０の「はい」の分岐）に応答して、プログラム２００がスパム攻撃に応答して軽減アクションを実行する。この例では、プログラム２００が、しきい値に対する特徴的な応答に基づいて、仮想看護師３７５によって提示される高忠実度の質問を決定する。プログラム２００は、次第に忠実度が低くなる質問を敵対者３１５に提供することによって、敵対者３１５が攻撃者であるということを仮想看護師が検証するのを支援する。プログラム２００は、忠実度レベル３の質問を敵対者３１５に提示する。プログラム２００によって提示された質問に対する敵対者３１５の全く不満足な回答に基づいて、プログラム２００は、忠実度レベル１０の質問を敵対者３１５に提示する。プログラム２００は、プログラム２００によって提示された質問に対する不可解な応答に基づいて、敵対者３１５が、仮想エージェントの運用経費を上昇させようとしているボットによって生成されたスパム・トラフィックであるということを決定する。

【００５９】
別の実施形態では、プログラム２００は、ランダムな誤った応答を挿入して、可能性のある損害を防ぐことができ、プログラム２００は、仮想エージェント・プログラム１７５がユーザと対話する際のトーンおよび方法を変更することができる。１つの例では、攻撃者によって収集されようとしている統計データを混乱させるために、プログラム２００が、時々、ランダムな不正確な応答を返す。プログラム２００は、望ましい忠実度レベルに従って、ランダムな応答の比率を適応させることができる。
【００６０】
追加の実施形態例では、図３に関して、リスク値がしきい値を超えたことの決定（判定ステップ２２０の「はい」の分岐）に応答して、プログラム２００が特徴的な専有モデルの機能の抽出攻撃に応答して軽減アクションを実行する。この例では、プログラム２００が、質問の特徴および仮想看護師３７５に対する応答に基づいて、敵対者３１５が専有モデルの機能の情報を収集しようとしているということを決定する。敵対者３１５は、仮想看護師３７５による応答の決定木の領域全体に関連する迅速な一連の質問を、仮想看護師３７５に提示する。敵対者３１５は、「はい」の応答を含む一連の回答を仮想看護師３７５に提示する。次に、敵対者３１５は、「いいえ」の応答を含めて、仮想看護師３７５によって提示された同じ質問に対する同じ一連の回答を提示する。プログラム２００は、特徴的な攻撃方法を認識し、仮想看護師３７５が、ランダムな無関係の応答を敵対者３１５にランダムに提供し、敵対者３１５が仮想看護師３７５の専有モデルの機能を収集するのを防ぐことを、支援する。
【００６１】
別の実施形態では、プログラム２００が、ユーザをハニーポット・モデルにリダイレクトすることができる。１つの例では、プログラム２００が、元のモデルの機能を模倣するために「ハニーポット」を作成して使用するが、このハニーポットは、元のグランド・トゥルースを大まかに表す、ただし攻撃者をだますことができる程度に十分近いデータを使用して、トレーニングされている。図３に関して、プログラム２００は、領域モデル３２０に似ている仮想「ハニーポット」を生成できる。この例では、攻撃者は、「ハニーポット」をリバース・エンジニアリングすることによって抽出攻撃を実行するため、実際の領域モデル３２０内の情報を攻撃者が実際に捕捉することが防がれる。この例では、盗まれたモデル３１０は、欺瞞（すなわち、「ハニーポット」）モデルのコピーになる。
【００６２】
追加の実施形態例では、図３に関して、リスク値がしきい値を超えたことの決定（判定ステップ２２０の「はい」の分岐）に応答して、プログラム２００が、専有情報を仮想看護師３７５のトレーニング・データから抽出するための敵対者３１５による攻撃に応答して軽減アクションを実行する。この例では、プログラム２００が、敵対者３１５が領域モデル３２０を抽出しようとしているということを決定する。プログラム２００は、敵対者３１５が仮想看護師３７５のセキュリティ・プロトコルを回避するために仮想看護師３７５に提示する特徴的な質問に基づいて、プログラム２００が、欺瞞エンジン１９０を介して、高対話型ハニーポットとして仮想看護師３７５を再作成するように仮想看護師３７５を適応させるということを、決定する。ここで、仮想看護師３７５は、高対話型ハニーポットとして機能し、敵対者３１５のツールおよび情報の抽出に使用される手法に関する詳細な情報を収集する。プログラム２００は、実際のシステム（ハニーポット・モデルとして一時的に目的を変更された仮想看護師３７５）を敵対者３１５に提示し、仮想看護師システムのルート権限を敵対者３１５に付与し、ハニーポット・システムへのアクセスを敵対者３１５に対して許可する。プログラム２００は、敵対者３１５の抽出攻撃に関する詳細な情報を収集し、抽出攻撃の特徴を作り出し、敵対者３１５のプロフィールおよび攻撃方法をボット・シールド・データベース１８２に格納する。
【００６３】
別の実施形態では、プログラム２００が、プローブ１９５を使用してユーザのリスク・スコアを素早く作り出すか、または不確定のユーザをさらに解析し、プローブ１９５による

さらなる解析に応じてリスク・スコアを割り当てる。プログラム２００は、正常な会話において発生する妥当性によって、使用可能なプローブをランク付けする。プログラム２００は、ユーザの現在のリスク・スコアに基づいてプローブを時々投入し、プログラム２００は、ユーザの現在のリスク・スコアに基づいて、プローブの強度および必要性、ならびに投入の頻度を調整できる。プログラム２００は、プローブに対するユーザによる応答を評価し、それに応じてリスク・スコアを更新するか、またはプログラム２００は、さらにプローブを採用することができる。プログラム２００は、プローブ１９５によって提供される情報に応じて、情報をボット・シールド・データベース１８２に追加する。プログラム２００は、プローブ１９５を介して、プローブに応答してユーザからさらに情報を得るために、要求または照会のその他の形態を送信する。プログラム２００は、プローブ１９５を介して、ユーザと仮想看護師３７５の間の会話に介入し、ユーザに割り当てられたリスク・スコアをさらに評価することができる。１つの例では、プログラム２００は、会話に介入し、「Ｃａｐｔｃｈａ」などの１つまたは複数のプローブを介してユーザが人間であることを探るために、ユーザに対して直接要求する。

【００６４】

１つの例では、プログラム２００は、新しいユーザ、または応答の特徴の履歴がボット・シールド・データベース１８２に格納されていないユーザに応答して、プローブ１９５を採用する。プローブ１９５は、プログラム２００によって低頻度で関与させられてよい。別の例では、プログラム２００は、ユーザによる口頭の発話に対するプローブを採用し、「すみません、私はこれに対してトレーニングされていません。言い換えてください」、「Ｘという意味でしたか？」などの語句を挿入する（Ｘは、以前のユーザの発話に関連していないということを仮想エージェント・プログラム１７５が強く確信している何かである（すなわち、消極的確認））か、またはプログラム２００は、プローブ１９５を利用して、「はい」または「いいえ」を超える回答を必要とする、現在の文脈に関連する余分な質問（例えば、自動車保険に関する会話における「自動車を初めて取得したのはいつですか」）を採用することができる。

【００６５】

追加の実施形態例では、図３に関して、リスク値がしきい値を超えたことの決定（判定ステップ２２０の「はい」の分岐）に応答して、プログラム２００が汚染攻撃に応答して軽減アクションを実行する。この例では、仮想看護師３７５が、実運用を通じて継続的に学習する。仮想看護師３７５が「良い」ユーザと多くの対話を行うほど、仮想看護師３７５の機能が改善され、仮想看護師３７５がさらに進化する。この例では、プログラム２００は、敵対者３１５による発話に基づいて、敵対者３１５が仮想看護師３７５のトレーニング・データを変更しているということを決定する。プログラム２００は、敵対者３１５によって模倣され、継続されている話題が、もともと医療の話題を甘受していた会話を、無関係の話題の領域に脱線させているということを決定する。それに応じて、プログラム２００は、仮想看護師３７５を「セーフ・モード」に移行させる。仮想看護師３７５は、敵対者３１５に対して、領域モデル３２０などの専有モデルの機能を含む、秘密情報、機密情報、専有情報をいずれも公開しない。敵対者３１５が引き続き仮想看護師３７５を「汚染」しようとしている場合、プログラム２００は、接続を終了し、仮想看護師３７５と敵対者３１５の間の会話を打ち切る。

【００６６】

追加の実施形態例では、図３に関して、リスク値がしきい値を超えたということの決定（判定ステップ２２０の「はい」の分岐）に応答して、プログラム２００が、ユーザに応答して忠実度を減らす軽減アクション、誤った応答をユーザに提供する軽減アクション、ハニーポットの欺瞞方法を使用する軽減アクション、および１つまたは複数のプローブを使用してユーザのリスク・スコアを更新する軽減アクションを結合することによって、軽減アクションを実行する。

【００６７】

図４は、本発明の実施形態例に従って、サーバ１２０のコンポーネントのブロック図を示

している。図４は、単に１つの実施形態の例を提供しており、さまざまな実施形態を実装できる環境に関して、どのような制限も意味していないと理解されるべきである。図に示された環境に対して、多くの変更が行われてよい。

【００６８】

サーバ１２０は、キャッシュ４１６、メモリ４０６、永続的ストレージ４０８、通信ユニット４１０、および入出力（Ｉ／Ｏ）インターフェイス４１２の間の通信を提供する通信ファブリック４０２を含んでいる。通信ファブリック４０２は、プロセッサ（マイクロプロセッサ、通信プロセッサ、およびネットワーク・プロセッサなど）、システム・メモリ、周辺機器、およびシステム内の任意のその他のハードウェア・コンポーネントの間で、データまたは制御情報あるいはその両方を渡すために設計された、任意のアーキテクチャを使用して実装され得る。例えば、通信ファブリック４０２は、１つまたは複数のバスまたはクロスバ・スイッチを使用して実装され得る。

【００６９】

メモリ４０６および永続的ストレージ４０８は、コンピュータ可読記憶媒体である。この実施形態では、メモリ４０６はランダム・アクセス・メモリ（ＲＡＭ）を含んでいる。一般に、メモリ４０６は、任意の適切な揮発性または不揮発性のコンピュータ可読記憶媒体を含むことができる。キャッシュ４１６は、メモリ４０６から最近アクセスされたデータ、およびアクセスされたデータに近いデータを保持することによって、コンピュータ・プロセッサ４０４の性能を向上させる高速なメモリである。

【００７０】

プログラム２００は、永続的ストレージ４０８、およびキャッシュ４１６を介して各コンピュータ・プロセッサ４０４のうちの１つまたは複数によって実行するためのメモリ４０６に格納されてよい。一実施形態では、永続的ストレージ４０８は、磁気ハード・ディスク・ドライブを含んでいる。磁気ハード・ディスク・ドライブに対する代替または追加として、永続的ストレージ４０８は、半導体ハード・ドライブ、半導体ストレージ・デバイス、読み取り専用メモリ（ＲＯＭ）、消去可能プログラマブル読み取り専用メモリ（ＥＰＲＯＭ）、フラッシュ・メモリ、あるいはプログラム命令またはデジタル情報を格納できる任意のその他のコンピュータ可読記憶媒体を含むことができる。

【００７１】

永続的ストレージ４０８によって使用される媒体は、取り外し可能であってもよい。例えば、取り外し可能ハード・ドライブを、永続的ストレージ４０８に使用できる。その他の例としては、永続的ストレージ４０８の一部でもある別のコンピュータ可読記憶媒体に転送するためのドライブに挿入される、光ディスクおよび磁気ディスク、サム・ドライブ、ならびにスマート・カードが挙げられる。

【００７２】

これらの例において、通信ユニット４１０は、他のデータ処理システムまたはデバイスとの通信を提供する。これらの例において、通信ユニット４１０は、１つまたは複数のネットワーク・インターフェイス・カードを含む。通信ユニット４１０は、物理的通信リンクまたは無線通信リンクのいずれか、あるいはその両方を使用して通信を提供できる。プログラム２００は、通信ユニット４１０を介して永続的ストレージ４０８にダウンロードされてよい。

【００７３】

Ｉ／Ｏインターフェイス４１２は、サーバ１２０に接続されてよい他のデバイスとのデータの入力および出力を可能にする。例えば、Ｉ／Ｏインターフェイス４１２は、キーボード、キーパッド、タッチスクリーン、またはその他の適切な入力デバイス、あるいはその組み合わせなどの、外部デバイス４１８への接続を提供してよい。外部デバイス４１８は、例えばサム・ドライブ、ポータブル光ディスクまたはポータブル磁気ディスク、およびメモリ・カードなどの、ポータブル・コンピュータ可読記憶媒体を含むこともできる。本発明の実施形態を実践するために使用されるソフトウェアおよびデータ（例えば、プログラム２００）は、そのようなポータブル・コンピュータ可読記憶媒体に格納することがで

き、Ｉ／Ｏインターフェイス４１２を介して永続的ストレージ４０８に読み込むことができる。Ｉ／Ｏインターフェイス４１２は、ディスプレイ４２０にも接続される。ディスプレイ４２０は、データをユーザに表示するためのメカニズムを提供し、例えば、コンピュータのモニタであってよい。

【００７４】

本明細書に記載されたプログラムは、アプリケーションに基づいて識別され、本発明の特定の実施形態において、そのアプリケーションに関して実装される。ただし、本明細書における特定のプログラムの名前は単に便宜上使用されていると理解されるべきであり、したがって、本発明は、そのような名前によって識別されたか、または暗示されたか、あるいはその両方によって示された特定のアプリケーションのみで使用するように制限されるべきではない。

【００７５】

本発明は、システム、方法、またはコンピュータ・プログラム製品、あるいはその組み合わせであってよい。コンピュータ・プログラム製品は、プロセッサに本発明の態様を実行させるためのコンピュータ可読プログラム命令を含んでいるコンピュータ可読記憶媒体を含んでよい。

【００７６】

コンピュータ可読記憶媒体は、命令実行デバイスによって使用するための命令を保持および格納できる有形のデバイスであることができる。コンピュータ可読記憶媒体は、例えば、電子ストレージ・デバイス、磁気ストレージ・デバイス、光ストレージ・デバイス、電磁ストレージ・デバイス、半導体ストレージ・デバイス、またはこれらの任意の適切な組み合わせであってよいが、これらに限定されない。コンピュータ可読記憶媒体のさらに具体的な例の非網羅的リストは、ポータブル・フロッピー（Ｒ）・ディスク、ハード・ディスク、ランダム・アクセス・メモリ（ＲＡＭ：random access memory）、読み取り専用メモリ（ＲＯＭ：read-onlymemory）、消去可能プログラマブル読み取り専用メモリ（ＥＰＲＯＭ：erasable programmable read-only memoryまたはフラッシュ・メモリ）、スタティック・ランダム・アクセス・メモリ（ＳＲＡＭ：static random access memory）、ポータブル・コンパクト・ディスク読み取り専用メモリ（ＣＤ－ＲＯＭ：compact disc read-only memory）、デジタル多用途ディスク（ＤＶＤ：digital versatile disk）、メモリ・スティック、フロッピー（Ｒ）・ディスク、パンチカードまたは命令が記録されている溝の中の隆起構造などの機械的にエンコードされるデバイス、およびこれらの任意の適切な組み合わせを含む。本明細書において使用されるとき、コンピュータ可読記憶媒体は、それ自体が、電波またはその他の自由に伝搬する電磁波、導波管またはその他の送信媒体を伝搬する電磁波（例えば、光ファイバ・ケーブルを通過する光パルス）、あるいはワイヤを介して送信される電気信号などの一過性の信号であると解釈されるべきではない。

【００７７】

本明細書に記載されたコンピュータ可読プログラム命令は、コンピュータ可読記憶媒体から各コンピューティング・デバイス／処理デバイスへ、またはネットワーク（例えば、インターネット、ローカル・エリア・ネットワーク、広域ネットワーク、または無線ネットワーク、あるいはその組み合わせ）を介して外部コンピュータまたは外部ストレージ・デバイスへダウンロードされ得る。このネットワークは、銅伝送ケーブル、光伝送ファイバ、無線送信、ルータ、ファイアウォール、スイッチ、ゲートウェイ・コンピュータ、またはエッジ・サーバ、あるいはその組み合わせを備えてよい。各コンピューティング・デバイス／処理デバイス内のネットワーク・アダプタ・カードまたはネットワーク・インターフェイスは、コンピュータ可読プログラム命令をネットワークから受信し、それらのコンピュータ可読プログラム命令を各コンピューティング・デバイス／処理デバイス内のコンピュータ可読記憶媒体に格納するために転送する。

【００７８】

本発明の動作を実行するためのコンピュータ可読プログラム命令は、アセンブラ命令、命

令セット・アーキテクチャ（ＩＳＡ：instruction-set-architecture）命令、マシン命令、マシン依存命令、マイクロコード、ファームウェア命令、状態設定データ、あるいは、Ｓｍａｌｌｔａｌｋ（Ｒ）、Ｃ＋＋などのオブジェクト指向プログラミング言語、および「Ｃ」プログラミング言語または同様のプログラミング言語などの従来の手続き型プログラミング言語を含む１つまたは複数のプログラミング言語の任意の組み合わせで記述されたソース・コードまたはオブジェクト・コードであってよい。コンピュータ可読プログラム命令は、ユーザのコンピュータ上で全体的に実行すること、ユーザのコンピュータ上でスタンドアロン・ソフトウェア・パッケージとして部分的に実行すること、ユーザのコンピュータ上およびリモート・コンピュータ上でそれぞれ部分的に実行すること、あるいはリモート・コンピュータ上またはサーバ上で全体的に実行することができる。後者のシナリオでは、リモート・コンピュータは、ローカル・エリア・ネットワーク（ＬＡＮ：local area network）または広域ネットワーク（ＷＡＮ：wide area network）を含む任意の種類のネットワークを介してユーザのコンピュータに接続されてよく、または接続は、（例えば、インターネット・サービス・プロバイダを使用してインターネットを介して）外部コンピュータに対して行われてよい。一部の実施形態では、本発明の態様を実行するために、例えばプログラマブル論理回路、フィールドプログラマブル・ゲート・アレイ（ＦＰＧＡ：field-programmable gate arrays）、またはプログラマブル・ロジック・アレイ（ＰＬＡ：programmable logic arrays）を含む電子回路は、コンピュータ可読プログラム命令の状態情報を利用することによって、電子回路をカスタマイズするためのコンピュータ可読プログラム命令を実行してよい。

【００７９】

本発明の態様は、本明細書において、本発明の実施形態に従って、方法、装置（システム）、およびコンピュータ・プログラム製品のフローチャート図またはブロック図あるいはその両方を参照して説明される。フローチャート図またはブロック図あるいはその両方の各ブロック、ならびにフローチャート図またはブロック図あるいはその両方に含まれるブロックの組み合わせが、コンピュータ可読プログラム命令によって実装され得るということが理解されるであろう。

【００８０】

これらのコンピュータ可読プログラム命令は、コンピュータまたはその他のプログラム可能なデータ処理装置のプロセッサを介して実行される命令が、フローチャートまたはブロック図あるいはその両方のブロックに指定される機能／動作を実施する手段を作り出すべく、汎用コンピュータ、専用コンピュータ、または他のプログラム可能なデータ処理装置のプロセッサに提供されてマシンを作り出すものであってよい。これらのコンピュータ可読プログラム命令は、命令が格納されたコンピュータ可読記憶媒体がフローチャートまたはブロック図あるいはその両方のブロックに指定される機能／動作の態様を実施する命令を含んでいる製品を備えるように、コンピュータ可読記憶媒体に格納され、コンピュータ、プログラム可能なデータ処理装置、または他のデバイス、あるいはその組み合わせに特定の方式で機能するように指示できるものであってもよい。

【００８１】

コンピュータ可読プログラム命令は、コンピュータ上、その他のプログラム可能な装置上、またはその他のデバイス上で実行される命令が、フローチャートまたはブロック図あるいはその両方のブロックに指定される機能／動作を実施するように、コンピュータ、その他のプログラム可能なデータ処理装置、またはその他のデバイスに読み込まれてもよく、それによって、一連の動作可能なステップを、コンピュータ上、その他のプログラム可能な装置上、またはコンピュータ実装プロセスを生成するその他のデバイス上で実行させる。

【００８２】

図内のフローチャートおよびブロック図は、本発明のさまざまな実施形態に従って、システム、方法、およびコンピュータ・プログラム製品の可能な実装のアーキテクチャ、機能、および動作を示す。これに関連して、フローチャートまたはブロック図内の各ブロックは、規定された論理機能を実装するための１つまたは複数の実行可能な命令を備える、命

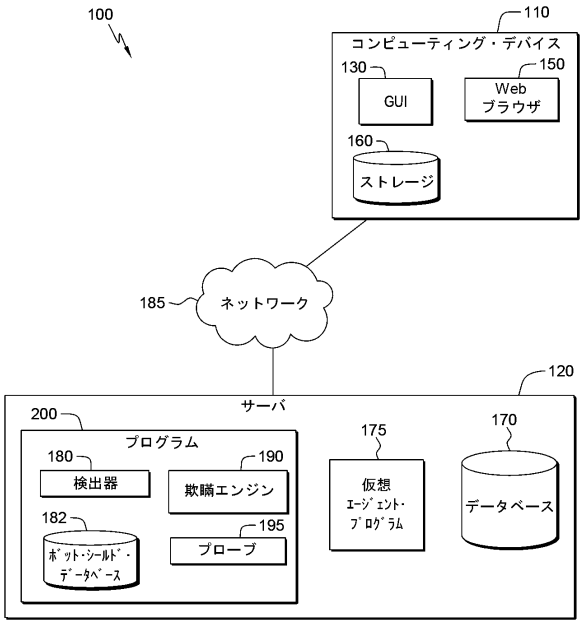
令のモジュール、セグメント、または部分を表してよい。一部の代替の実装では、ブロックに示された機能は、図に示された順序とは異なる順序で発生してよい。例えば、連続して示された２つのブロックは、実際には、含まれている機能に応じて、実質的に同時に実行されるか、または場合によっては逆の順序で実行されてよい。ブロック図またはフローチャート図あるいはその両方の各ブロック、ならびにブロック図またはフローチャート図あるいはその両方に含まれるブロックの組み合わせは、規定された機能または動作を実行するか、または専用ハードウェアとコンピュータ命令の組み合わせを実行する専用ハードウェアベースのシステムによって実装され得るということにも注意する。

【００８３】

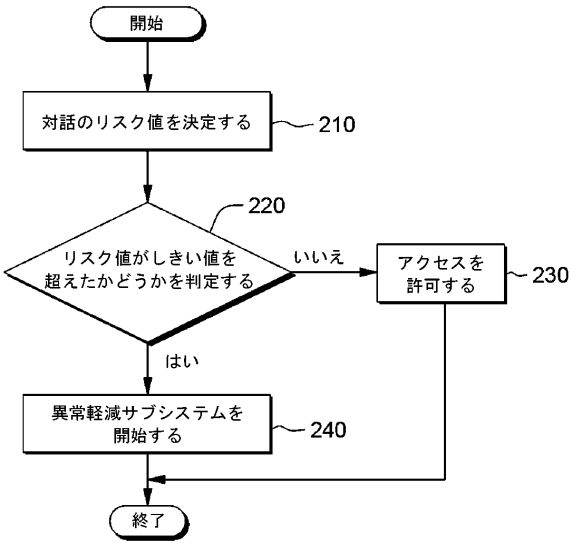
本発明のさまざまな実施形態の説明は、例示の目的で提示されているが、網羅的であることは意図されておらず、開示された実施形態に制限されない。本発明の範囲を逸脱することなく多くの変更および変形が可能であることは、当業者にとって明らかである。本明細書で使用された用語は、実施形態の原理、実際のアプリケーション、または市場で見られる技術を超える技術的改良を最も適切に説明するため、または他の当業者が本明細書で開示された実施形態を理解できるようにするため選択されている。

【図面】

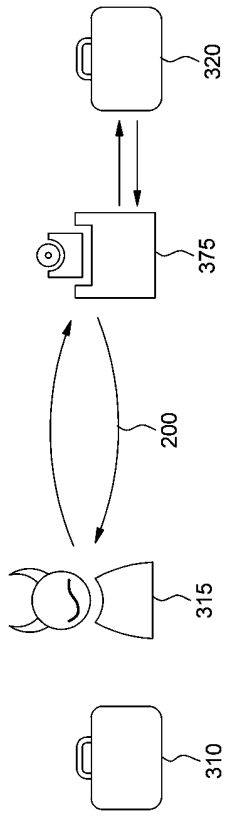
【図１】


100
110
コンピューティング・デバイス
130
GUI
150
Web
ブラウザ
160
ストレージ
185
ネットワーク
120
サーバ
200
プログラム
180
検出器
190
欺瞞エンジン
182
ﾎﾞｯﾄ･ｼｰﾙﾄﾞ･
ﾃﾞｰﾀﾍﾞｰｽ
195
プローブ
175
仮想
ｴｰｼﾞｪﾝﾄ･
ﾌﾟﾛｸﾞﾗﾑ
170
データベース

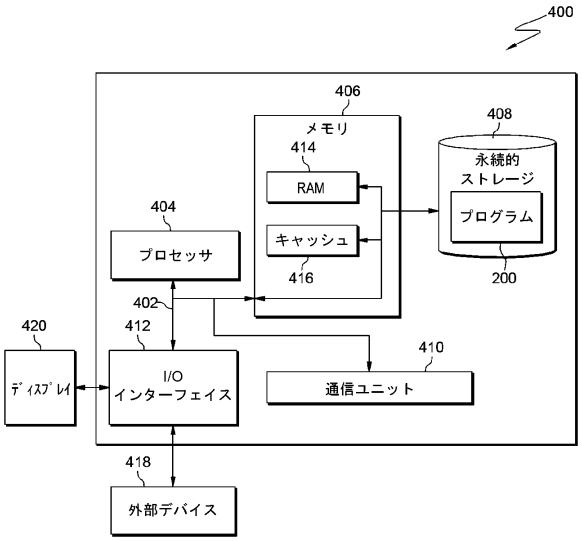
【図２】


開始
対話のリスク値を決定する
210
220
リスク値がしきい値を
超えたかどうかを判定する
いいえ
アクセスを
許可する
230
はい
異常軽減サブシステムを
開始する
240
終了

【図３】


320
375
200
315
310

【図４】


400
406
メモリ
414
RAM
キャッシュ
416
408
永続的
ストレージ
プログラム
200
404
プロセッサ
402
420
ﾃﾞｨｽﾌﾟﾚｲ
412
I/O
インターフェイス
410
通信ユニット
418
外部デバイス

フロントページの続き

(72)発明者　デスターバルト、イヴリン
アメリカ合衆国１０５９８　ニューヨーク州ヨークタウン・ハイツ　キッチャワン・ロード１１０１
(72)発明者　バウダート、ギヨーム、アントニン
アメリカ合衆国１０５９８　ニューヨーク州ヨークタウン・ハイツ　キッチャワン・ロード１１０１
(72)発明者　ピオルコウスキ、デイヴィッド、ジョン
アメリカ合衆国１０５９８　ニューヨーク州ヨークタウン・ハイツ　キッチャワン・ロード１１０１
(72)発明者　ドルビー、ジュリアン、ティモシー
アメリカ合衆国１０５９８　ニューヨーク州ヨークタウン・ハイツ　キッチャワン・ロード１１０１

審査官　松平 英
(56)参考文献　特表２００３－５２３５７８（ＪＰ，Ａ）
特表２０１６－５１１８４７（ＪＰ，Ａ）
特開２０１７－０７３１２５（ＪＰ，Ａ）
特表２０１３－５０３３７７（ＪＰ，Ａ）
特開２０１７－０２７１３４（ＪＰ，Ａ）
特開２０１７－１３８７２８（ＪＰ，Ａ）
特開平１１－３２７９０８（ＪＰ，Ａ）
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ１２／１４
２１／００－２１／８８
Ｇ０６Ｎ　５／００－７／０６
Ｇ０９Ｃ　１／００－５／００
Ｈ０４Ｋ　１／００－３／００
Ｈ０４Ｌ　９／００－９／４０