(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111213161 B
(45)授权公告日 2023.11.10

(21)申请号 201880067127.1

(22)申请日 2018.10.10

(65)同一申请的已公布的文献号
申请公布号 CN 111213161 A

(43)申请公布日 2020.05.29

(30)优先权数据
15/786,888 2017.10.18 US

(85)PCT国际申请进入国家阶段日
2020.04.15

(86)PCT国际申请的申请数据
PCT/IB2018/057830 2018.10.10

(87)PCT国际申请的公布数据
WO2019/077440 EN 2019.04.25

(73)专利权人 国际商业机器公司
地址 美国纽约阿芒克

(72)发明人 E.杜斯特瓦尔德 G.A.鲍达尔特
D.J.皮奥科夫斯基 J.T.杜比

(74)专利代理机构 北京市柳沈律师事务所 11105
专利代理师 邸万奎

(51)Int.Cl.
H04L 9/40(2022.01)
H04L 51/02(2022.01)
G06N 5/025(2023.01)
G06F 40/30(2020.01)
G06N 20/20(2019.01)
G06N 3/006(2023.01)
G06N 7/01(2023.01)
G06F 16/9535(2019.01)

(56)对比文件
US 2016071517 A1,2016.03.10
CN 104704797 A,2015.06.10
Tianwei Zhang.Detection and Mitigation of Security Threats in Cloud Computing.《Princeton University》.2017,第1-7章.
Brian M. Bowen.Design and analysis of Decoy Systems for Computer Security.《Computer Science》.2011,摘要、第1-9章.

审查员 霍玉明

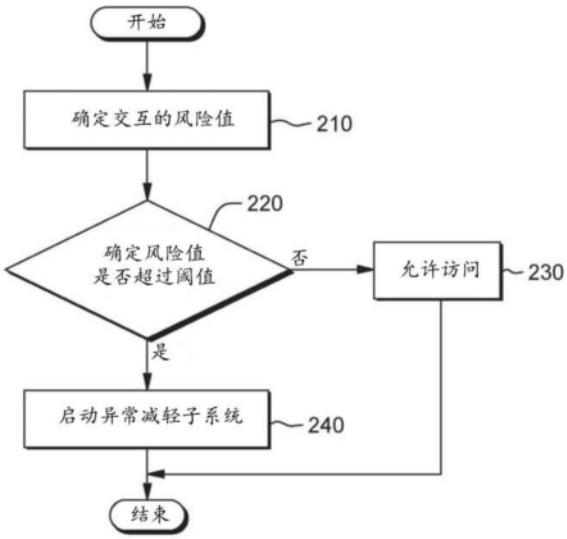
权利要求书3页 说明书14页 附图4页

(54)发明名称

认知虚拟检测器

(57)摘要

本发明的各方面公开了一种用于检测和减轻对抗虚拟交互的方法、计算机程序产品和系统。该方法包括一个或多个处理器检测与虚拟代理交互的用户通信。该方法还包括一个或多个处理器基于由检测到的用户在与虚拟代理交互时执行的一个或多个动作来确定与检测到的用户通信相关联的风险水平。该方法还包括一个或多个处理器响应于确定与检测到的用户通信相关联的所确定的风险水平超过风险水平阈值,发起关于检测到的用户与虚拟代理之间的交互的缓解协议,其中缓解协议基于由检测到的用户在与虚拟代理交互时执行的动作。



CN 111213161 B

1.一种用于检测和减轻对抗虚拟交互的计算机实现的方法,所述方法包括:

由一个或多个处理器检测与虚拟代理交互的用户通信;

由一个或多个处理器基于由检测到的用户在与所述虚拟代理交互时执行的一个或多个动作来确定与检测到的用户通信相关联的风险水平;以及

响应于确定与检测到的用户通信相关联的所确定的风险水平超过风险水平阈值,由一个或多个处理器发起关于所述检测到的用户与所述虚拟代理之间的交互的缓解协议,其中缓解协议基于由所述检测到的用户在与所述虚拟代理交互时执行的动作;

响应于发起关于所述检测到的用户与所述虚拟代理之间的交互的缓解协议,由一个或多个处理器生成从所述虚拟代理到所述用户的较低保真度响应,其中所述较低保真度响应是在所述检测到的用户超过所述风险水平阈值之前从所述虚拟代理到所述用户的原始响应的语言精度的渐进式减弱。

2.根据权利要求1所述的方法,其中,发起所述缓解协议还包括:

由一个或多个处理器改变从所述虚拟代理到所述用户的通信响应;以及

由一个或多个处理器将所述用户与所述虚拟代理之间的会话引导到预定会话树中,其中所述预定会话树是隐藏与所述虚拟代理相关联的机密数据的响应协议。

3.根据权利要求1所述的方法,其中,发起所述缓解协议还包括:

由一个或多个处理器终止所述用户与所述虚拟代理之间的交互;

由一个或多个处理器向一个或多个企业网络数据库报告所述交互的数据;以及

由一个或多个处理器存储所述交互的数据,其中会话的数据包括一个或多个频繁使用的术语,以及被设计为从所述虚拟代理提取信息的技术。

4.根据前述权利要求中任一项所述的方法,其中,发起所述缓解协议还包括:

由一个或多个处理器确定与所述用户相关联的所述风险水平正在增加;以及

由一个或多个处理器与和所述用户相关联的所述增加的风险水平成比例地延迟从所述虚拟代理到所述用户的所述响应的时间段。

5.根据前述权利要求1-3中的任一项所述的方法,其中,发起所述缓解协议还包括:

由一个或多个处理器识别所述用户的签名,其中所述签名与对所述虚拟代理的提取攻击相关联;

由一个或多个处理器生成与所述用户的通信相关联的机密信息的诱饵模型,其中所述机密信息的诱饵模型的格式化匹配机密信息的实际实例的格式化;

由一个或多个处理器将所述诱饵模型呈现给所述用户;以及

由一个或多个处理器从所述用户提取数据,其中从所述用户提取的所述数据包括由所述用户使用以提取专有数据的过程。

6.根据前述权利要求1-3中的任一项所述的方法,其中,基于由所述检测到的用户在与所述虚拟代理交互时执行的一个或多个动作来确定与所述检测到的用户通信相关联的所述风险水平还包括:

由一个或多个处理器激活探测器以从所述用户检索更多信息;以及

由一个或多个处理器基于来自所述探测器的所述信息来更新与所述用户通信相关联的所述风险水平。

7.根据前述权利要求1-3中任一项所述的方法,还包括:

响应于确定与检测到的用户通信相关联的所确定的风险水平超过所述风险水平阈值，由一个或多个处理器组合地发起多个减轻动作，其中，所述减轻动作是从由以下各项组成的组中选择的：生成从所述虚拟代理到所述用户的较低保真度响应，终止所述用户与所述虚拟代理之间的交互，激活探测器以从所述用户检索信息，延迟从所述虚拟代理到所述用户的响应的时间段，以及生成机密信息的诱饵模型。

8.一种计算机可读介质，其上存储有可加载到数字计算机的内部存储器中的计算机程序，包括软件代码部分，当所述程序在计算机上运行时，用于执行根据权利要求1至7中任一项所述的方法。

9.一种用于检测和减轻对手虚拟交互的计算机系统，所述计算机系统包括：

一个或多个计算机处理器；

一个或多个计算机可读存储介质；

存储在所述一个或多个计算机可读存储介质上以供所述一个或多个计算机处理器中的至少一个计算机处理器执行的程序指令，所述程序指令包括：

用于检测与虚拟代理交互的用户通信的程序指令；

用于基于由检测到的用户在与虚拟代理交互时执行的一个或多个动作来确定与检测到的用户通信相关联的风险水平的程序指令；

用于响应于确定与检测到的用户通信相关联的所确定的风险水平超过风险水平阈值，发起关于所述检测到的用户与所述虚拟代理之间的交互的缓解协议的程序指令，其中，所述缓解协议基于由所述检测到的用户在与所述虚拟代理交互时执行的动作；以及

用于响应于发起关于所述检测到的用户与所述虚拟代理之间的交互的缓解协议，由一个或多个处理器生成从所述虚拟代理到所述用户的较低保真度响应的程序指令，其中所述较低保真度响应是在所述检测到的用户超过所述风险水平阈值之前从所述虚拟代理到所述用户的原始响应的语言精度的渐进式减弱。

10.根据权利要求9所述的计算机系统，其中，发起所述缓解协议还包括存储在所述一个或多个计算机可读存储介质上的程序指令，所述程序指令在由处理器执行时使得所述处理器：

改变从所述虚拟代理到所述用户的通信响应；以及

将所述用户和所述虚拟代理之间的会话引导到预定会话树中，其中所述预定会话树是隐藏与所述虚拟代理相关联的机密数据的响应协议。

11.根据权利要求9所述的计算机系统，其中，发起所述缓解协议还包括存储在所述一个或多个计算机可读存储介质上的程序指令，所述程序指令在由处理器执行时使得所述处理器：

终止所述用户与所述虚拟代理之间的交互；

将所述交互的数据报告给一个或多个企业网络数据库；以及

存储所述交互的数据，其中会话的数据包括一个或多个频繁使用的术语，以及被设计为从所述虚拟代理提取信息的技术。

12.根据权利要求9至11中的任一项所述的计算机系统，其中，发起所述缓解协议还包括存储在所述一个或多个计算机可读存储介质上的程序指令，所述程序指令在由处理器执行时使所述处理器：

确定与所述用户相关联的所述风险水平正在增加;以及

与和所述用户相关联的增加的风险水平成比例地延迟从所述虚拟代理到所述用户的所述响应的时间段。

13.根据权利要求9至11中的任一项所述的计算机系统,其中,发起所述缓解协议还包括存储在所述一个或多个计算机可读存储介质上的程序指令,所述程序指令在由处理器执行时使所述处理器:

识别所述用户的签名,其中所述签名与对所述虚拟代理的提取攻击相关联;

生成与所述用户的通信相关联的机密信息的诱饵模型,其中所述机密信息的诱饵模型的格式化匹配机密信息的实际实例的格式化;

将所述诱饵模型呈现给所述用户;以及

从所述用户提取数据,其中从所述用户提取的所述数据包括由所述用户用于提取专有数据的过程。

14.根据权利要求9至11中的任一项所述的计算机系统,其中,基于由所述检测到的用户在与所述虚拟代理交互时执行的一个或多个动作来确定与所述检测到的用户通信相关联的所述风险水平还包括存储在所述一个或多个计算机可读存储介质上的程序指令,所述程序指令在由处理器执行时使所述处理器:

激活探测器以从所述用户检索更多信息;以及

基于来自所述探测器的所述信息来更新与所述用户通信相关联的所述风险水平。

15.根据权利要求9至11中任一项所述的计算机系统,存储在所述一个或多个计算机可读存储介质上,当由处理器执行时,使所述处理器:

响应于确定与所述检测到的用户通信相关联的所确定的风险水平超过所述风险水平阈值,组合地发起多个减轻动作,其中,所述减轻动作是从由以下各项组成的组中选择的:生成从所述虚拟代理到所述用户的较低保真度响应,终止所述用户与所述虚拟代理之间的交互,激活探测器以从所述用户检索信息,延迟从所述虚拟代理到所述用户的响应的时间段,以及生成机密信息的诱饵模型。

认知虚拟检测器

技术领域

[0001] 本发明一般地涉及控制论领域，更具体地涉及人工智能。

背景技术

[0002] 在人工智能中，智能代理(IA)是自主实体，其通过传感器观察并使用致动器对环境进行作用(即，其是代理)，并将其活动引向实现目标(即，其是“合理的”，如经济学中所定义的)。智能代理还可以学习或使用知识来实现其目标。它们可以非常简单或非常复杂：反射机器，例如恒温器，是智能代理。

[0003] 简单的反射代理仅基于当前感知起作用，忽略感知历史的其余部分。代理功能基于条件-动作规则：如果条件，则动作。这个代理功能只有在环境完全可观察时才成功。一些反射代理还可以包含关于它们的当前状态的信息，这允许它们忽略其致动器已经被触发的条件。对于在部分可观察环境中运行的简单反射代理，无限循环通常是不可避免的。

[0004] 基于模型的代理可以处理部分可观察环境。其当前状态被存储在代理内部，维持描述世界上不能被看到的部分的某种结构。这种关于“世界如何工作”的知识被称为世界的模型，因此被称为“基于模型的代理”。基于模型的反射代理应当保持某种内部模型，该内部模型取决于感知历史，并且由此反映当前状态的至少一些未被观察的方面。可以通过使用内部模型来确定感知历史和动作对环境的影响。然后，它以与反射代理相同的方式选择动作。

[0005] 基于目标的代理通过使用“目标”信息进一步扩展基于模型的代理的能力。目标信息描述了期望的情况。这允许代理在多种可能性中进行选择的方式，选择达到目标状态的一个可能性。搜索和规划是人工智能的子域，其致力于寻找实现代理目标的动作序列。基于目标的代理更灵活，因为支持其决定的知识被明确地表示并且可以被修改。

[0006] 基于目标的代理仅在目标状态与非目标状态之间进行区分。可以定义特定状态的期望程度的度量。这个度量可以通过使用将状态映射到该状态的效用度量的效用函数来获得。更一般的性能度量应该允许根据不同世界状态使代理准确地有多快乐来比较这些不同世界状态。术语“效用”可用于描述代理“快乐”的程度。基于合理效用的代理选择最大化动作结果的预期效用的动作，即，给定每个结果的概率和效用，代理平均预期得出什么。基于效用的代理需要建模并跟踪其环境、涉及对感知、表示、推理和学习的大量研究的任务。

[0007] 学习具有的优点是，其允许代理最初在未知环境中操作，并且变得比其初始知识单独可能允许的更有能力。最重要的区别在于负责进行改进的“学习元件”和负责选择外部动作的“性能元件”之间。学习元件使用来自“评论者”的关于代理如何做的反馈，并确定在将来应如何修改性能元件以做得更好。性能元件是我们先前认为是整个代理的元件：它接收感知并决定动作。学习代理的最后一个组件是“问题产生器”，它负责建议将导致新的和信息丰富的体验的动作。

[0008] 虚拟代理正越来越多地部署在企业中以处理与客户或与雇员的交互。随着这些虚拟代理在企业中承担更多的功能，它们正日益成为攻击(例如，垃圾邮件、提取、中毒和规避

攻击)的目标。因此,在本领域中需要解决上述问题。

发明内容

[0009] 从第一方面来看,本发明提供了一种用于检测和减轻对抗虚拟交互的计算机实现的方法,该方法包括:由一个或多个处理器检测与虚拟代理交互的用户通信;由一个或多个处理器基于由检测到的用户在与虚拟代理交互时执行的一个或多个动作来确定与检测到的用户通信相关联的风险水平;以及响应于确定与检测到的用户通信相关联的所确定的风险水平超过风险水平阈值,由一个或多个处理器发起关于检测到的用户与虚拟代理之间的交互的缓解协议,其中缓解协议基于由检测到的用户在与虚拟代理交互时执行的动作。

[0010] 从第一方面来看,本发明提供了一种用于检测和减轻对抗虚拟交互的计算机系统,该计算机系统包括:一个或多个计算机处理器;一个或多个计算机可读存储介质;以及存储在一个或多个计算机可读存储介质上以供一个或多个计算机处理器中的至少一个计算机处理器执行的程序指令,程序指令包括:用于检测与虚拟代理交互的用户通信的程序指令;用于基于由检测到的用户在与虚拟代理交互时执行的一个或多个动作来确定与检测到的用户通信相关联的风险水平的程序指令;以及用于响应于确定与检测到的用户通信相关联的所确定的风险水平超过风险水平阈值而发起关于检测到的用户与虚拟代理之间的交互的缓解协议的程序指令,其中缓解协议基于由检测到的用户在与虚拟代理交互时执行的动作。

[0011] 从另一方面来看,本发明提供了一种用于检测和减轻对抗虚拟交互的计算机程序产品,该计算机程序产品包括计算机可读存储介质,该计算机可读存储介质可由处理电路读取并且存储用于由处理电路执行以执行用于执行本发明的步骤的方法的指令。

[0012] 从另一方面来看,本发明提供了一种存储在计算机可读介质上并且可加载到数字计算机的内部存储器中的计算机程序,包括软件代码部分,当所述程序在计算机上运行时,用于执行本发明的步骤。

[0013] 根据本发明的一个实施例,提供了一种用于检测和减轻对抗虚拟交互的方法。用于检测和减轻对手虚拟交互的方法可以包括一个或多个处理器检测与虚拟代理交互的用户通信。该方法还包括一个或多个处理器基于由检测到的用户在与虚拟代理交互时执行的一个或多个动作来确定与检测到的用户通信相关联的风险水平。该方法还包括一个或多个处理器响应于确定与检测到的用户通信相关联的所确定的风险水平超过风险水平阈值,发起关于检测到的用户与虚拟代理之间的交互的缓解协议,其中缓解协议基于由检测到的用户在与虚拟代理交互时执行的动作。

附图说明

[0014] 现在,将参考优选实施例仅通过示例的方式描述本发明,如以下附图所示:

[0015] 图1是示出根据本发明实施例的分布式数据处理环境的功能框图。

[0016] 图2是描述根据本发明的实施例的用于检测和减轻与虚拟代理的对抗会话的程序的操作步骤的流程图。

[0017] 图3示出了根据本发明的实施例的用于检测和减轻与虚拟代理的对抗会话的程序的示例。

[0018] 图4是根据本发明实施例的计算机系统(例如图1的服务器计算机)的组件的框图。

具体实施方式

[0019] 本发明的实施例认识到,虚拟代理(例如,虚拟智能代理)正越来越多地部署在企业中以处理与客户或与雇员的交互。虚拟代理正日益成为攻击和滥用的目标,例如,由机器人生成的垃圾邮件流量使带宽饱和或抬高操作成本。后端专有虚拟代理模型易受提取攻击,以便反向工程专有模型功能或从训练数据提取专有信息。持续从生产使用中学习的虚拟代理是中毒攻击的目标。在这种情况下,训练数据中的移位分布被用于驱使会话偏离路线。命令控制情形中使用的虚拟代理是攻击者学习和利用底层模型中的弱点来误导或欺骗虚拟代理的攻击目标。这样的一个例子是经由隐藏的语音命令将用户电话的控制隐蔽地接管到语音控制聊天机器人。

[0020] 本发明的实施例认识到,当前检测对虚拟代理的攻击的方法不足以检测语义应用层攻击,例如提取、中毒和规避攻击,以及对人类行为建模并以较低量进行的更复杂的垃圾邮件攻击。

[0021] 本发明的实施例提供了一种与虚拟代理一起构建的系统,该系统使用采用三个子系统的方法来监督虚拟代理的工作。本发明的实施例提供了一种用于分析话语并检测可疑用户行为的检测子系统。该系统包括用于检查用户行为的检测模型的集合。本发明的实施例提供了一种用于在提升的怀疑级别上重定向虚拟代理响应的欺骗子系统。本发明的实施例提供了一种探测子系统,以通过信息收集探测器最大化关于用户的学习。隐藏会话被注入会话流中,以便揭示对抗用户意图。

[0022] 现在将参照附图详细描述根据本发明的示例性实施例。图1是示出分布式数据处理环境100的功能框图。分布式数据处理环境100包括通过网络185互连的计算设备110和服务器120。

[0023] 在一个实施例中,计算设备110包括图形用户界面(GUI)130、网络浏览器150和存储160。计算设备110上的各种程序包括网络浏览器、电子邮件客户端、安全软件(例如,防火墙程序、地理定位程序、加密程序等)、即时消息(IM)应用(app)、以及通信(例如,电话)应用。

[0024] 计算设备110可以是台式计算机、膝上型计算机、平板计算机、专用计算机服务器、智能电话、可穿戴设备(例如,智能手表、个人健身设备、个人安全设备)、或本领域中已知的具有交互式显示器的任何可编程计算机系统、或本领域中已知的任何其他计算机系统。在某些实施例中,计算设备110表示利用集群计算机和组件的计算机系统,当通过网络185访问时,集群计算机和组件充当单个无缝资源池,这在数据中心中以及与云计算应用中是常见的。通常,计算设备110表示能够执行机器可读程序指令并经由网络与其他计算机设备通信的任何可编程电子设备或可编程电子设备的组合。

[0025] 在一个实施例中,图形用户界面130在计算设备110上运行。在另一实施例中,图形用户界面130在基于服务器的设置的另一计算机上运行;例如在服务器计算机(例如服务器120)上。在又一实施例中,图形用户界面130在计算设备110上运行,同时通过网络185与服务器计算机(例如,服务器120)互连。图形用户界面130可以是用于从计算设备110访问信息的任何用户界面,诸如由程序200收集或产生的信息。另外,图形用户界面130可以是用于向

计算设备110提供信息的任何用户界面，诸如由用户提供的用于输入到程序200的信息。在一些实施例中，图形用户界面130可以呈现用于从因特网检索、呈现和协商资源的通用网络浏览器。在其他实施例中，图形用户界面130可以是使得计算设备110处的用户能够访问网络185的软件或应用。

[0026] 在又一实施例中，计算设备110的用户可以通过触摸屏与图形用户界面130交互，该触摸屏既作为图形用户界面(GUI)的输入设备又作为呈现与软件应用相关联的多个图标或描绘执行中的软件应用的图像的输出设备(即，电子显示器)来执行。可选地，软件应用(例如，网络浏览器)可以生成在计算设备110的GUI内运行的图形用户界面130。图形用户界面130接受来自多个输入/输出(I/O)设备的输入，所述输入/输出设备包括但不限于被称为多点触摸显示器的触觉传感器接口(例如，触摸屏或触摸板)。与图形用户界面130对接的I/O设备可以连接到计算设备110，其可以利用有线(例如，USB端口)或无线网络通信(例如，红外、NFC等)来操作。根据本发明的实施例，计算设备110可以包括如参考图4进一步详细描绘和描述的组件。

[0027] 网络浏览器150可以是用于从因特网检索、呈现和遍历信息资源的通用网络浏览器。在一些实施例中，网络浏览器150可以是为移动设备设计的网络浏览器。在其他实施例中，网络浏览器150可以是为诸如台式计算机、PC或膝上型计算机等传统计算设备设计的网络浏览器。一般而言，网络浏览器150可以是使得计算设备110的用户能够通过网络185访问网页的任何应用或软件。在所描绘的环境中，网络浏览器150驻留在计算设备110上。在其它实施例中，网络浏览器150或类似的网络浏览器可以驻留在能够通过网络185访问网页的其它计算设备上。

[0028] 位于计算设备110上的存储160(例如，数据库)表示能够存储由计算设备110访问和利用的数据的任何类型的存储设备。在其他实施例中，存储160表示计算设备110内的多个存储设备。存储160存储信息，例如但不限于账户信息、认证证书、用户偏好、优选用户列表、先前访问的网站、访问的Wi-Fi门户的历史、以及计算设备的位置的历史。

[0029] 通常，网络185可以是将支持计算设备110之间的通信的连接和协议的任何组合。网络185可以包括，例如，局域网(LAN)、诸如因特网的广域网(WAN)、蜂窝网络、或前述的任何组合，并且还可以包括有线、无线和/或光纤连接。

[0030] 服务器120可以是台式计算机、膝上型计算机、平板计算机、专用计算机服务器、智能电话或本领域已知的任何其它计算机系统。在某些实施例中，服务器120表示利用集群计算机和组件的计算机系统，当通过网络185访问时，集群计算机和组件充当单个无缝资源池，这在数据中心中以及与云计算应用中是常见的。通常，服务器120表示能够执行机器可读程序指令并经由网络与其它计算机设备通信的任何可编程电子设备或可编程电子设备的组合。在一个实施例中，服务器120包括数据库170和程序200。

[0031] 在一个实施例中，服务器120能够发起服务器120与计算设备110之间的握手过程。握手是协商的自动过程，其在两个实体之间建立的通信信道上的正常通信开始之前动态地设置该信道的参数。握手跟随信道的物理建立并且在正常信息传输之前。握手便于在通信信道上连接异构计算系统或设备，而无需用户干预来设置参数。在一个示例中，服务器120通过向计算设备110发送消息来发起握手过程，该消息指示服务器120想要建立通信信道以便获得对计算设备110上的程序的访问。

[0032] 数据库170可以是可由服务器120读取的储存库。数据库170表示能够存储由服务器120访问和使用的数据的任何类型的存储设备。在其它实施例中，数据库170表示服务器120内的多个存储设备。数据库170存储信息，例如但不限于账户信息、认证证书、用户偏好、优选用户列表、先前访问的网站、访问的Wi-Fi门户的历史和计算设备的历史，以及位于访问服务器的计算设备上的信息。

[0033] 虚拟代理程序175是服务器120上的程序。在一个实施例中，虚拟代理程序175是具有拟人外观的生动的、人工智能虚拟角色，其用作在线客户服务代表。在一个示例中，虚拟代理175已经理解用户想要具有的许多会话，因为它是用行业和领域内容预先训练的。虚拟代理程序175应用认知技术来提供个性化的、情境化的客户体验，具有预先训练的行业和领域知识。在另一个实施例中，虚拟代理程序175引导与用户的智能会话，响应用户问题并执行适当的非口头行为。虚拟代理程序175能够使用参与度量来更多地理解虚拟代理程序175与一个或多个用户正在进行的会话。在另一实施例中，虚拟代理175也可以作为酒店大堂、餐馆或商务办公室接待中欢迎顾客的全息投影出现。

[0034] 检测器180是程序200的子程序，其截取用户向虚拟代理程序175发出的请求以及虚拟代理程序175的响应，并且用作程序200的会话异常检测子系统。在一个实施例中，检测器180提取虚拟代理程序175的日志条目请求和响应，并且检测器180使用日志条目作为输入来生成异常检测模型的集合。在该实施例中，检测器180能够生成对于不同检测策略唯一的检测模型。在一个示例中，检测器180生成并合并以下检测模型中的每一个：用于分析自然语言的马尔可夫模型，用于跟踪查询到模型的分布并确定信息暴露的风险级别的信息泄露模型，用于检查时间标签以检测人类的可疑快速回答的定时模型，用于标记所识别意图上的低置信度分数的出现的置信度模型，以及用于标识会话进展指示符的不存在或存在的会话进展模型，诸如与用户的会话中的特定目标状态的出现。检测器180将各个检测模型组合成单个风险分数，对该单个风险分数进行加权以辅助确定响应的可疑度。

[0035] 机器人防护数据库182是位于程序200上并由程序200排他地使用的数据库。机器人防护数据库182表示能够存储由程序200访问和利用的数据的任何类型的存储设备。在其他实施例中，机器人防护数据库182表示程序200内的多个存储设备。机器人防护数据库182存储诸如但不限于上下文信息、帐户信息、用于认证的证书、用户偏好、以及优选用户的列表等信息。例如，机器人防护数据库182存储与高风险水平相关联的、暗示这些词与可疑活动相关联的词和短语。机器人防护数据库182存储来自各个检测模型的数据的历史。例如，机器人防护数据库182存储与高风险、中等风险和低风险用户相关联的定时检测模型的定时响应模式的一个或多个历史。在另一示例中，机器人防护数据库182存储来自马尔可夫检测模型的信息。在该示例中，机器人防护数据库182存储虚拟代理和用户之间的不太可能的交互的历史信息，该历史信息被标记为不适当的。机器人防护数据库182存储来自用户和虚拟代理之间的会话状态的转换频率的历史。

[0036] 在另一实施例中，机器人防护数据库182存储在检测用于单独异常检测模型的异常中的有效性的历史信息，并且异常检测模型集合在一个或多个布置中。在另一实施例中，机器人防护数据库182存储异常检测子系统中的附加检测模型的扩展和更新、缓解系统中的缓解响应以及探测系统中的探测器选择的历史信息。在一个示例中，程序200接收新的规避恶意软件和能够成功地减轻对虚拟代理的新颖攻击的升级的防御的更新。机器人防护数

据库182存储新的规避恶意软件的信息以及对该恶意软件的升级的防御。

[0037] 欺骗引擎190是程序200的子程序，其自动调整虚拟代理程序175对用户的响应的保真度，以阻止潜在的攻击。欺骗引擎190通过改变模型响应的保真度或准确性而不改变与用户的原始会话流来减轻攻击。欺骗引擎190通过根据当前用户风险分数选择响应的保真度级别来改变给予用户的模型响应的准确性。用户风险越高，给予高风险用户的模型响应的精度越低。例如，如果在欺骗引擎190改变响应的保真度之前，模型响应是“请确认你输入的信用卡号###-####-####”，则欺骗引擎190响应于用户的高风险活动，改变与用户的风险级别一致的响应的保真度。在该示例中，欺骗引擎190将模型响应的保真度改变为“请重新输入你的信用卡号以确认”。如果用户风险得分超过某个预定阈值，则触发欺骗引擎190。在一个实施例中，欺骗引擎190基于用户的特定风险水平以及基于对虚拟代理程序175的用户响应而超过特定阈值的事实来触发缓解动作。

[0038] 在另一实施例中，欺骗引擎190除了如前所述通过改变模型响应的保真度来将会话流重定向到安全会话之外，还能够减轻对虚拟代理程序175的攻击。在一个实施例中，欺骗引擎190使用多个策略来创建较低保真度响应，诸如渐进式模型稀释。在一个示例中，欺骗引擎190使用先前训练的模型作为稀释模型的基线。欺骗引擎190使先前训练的模型成为原始模型中的基线事实的较不准确的版本。在另一个实施例中，欺骗引擎190逐渐插入随机的错误响应。在一个示例中，欺骗引擎190不时地向用户返回随机化的、不正确的响应，以扰乱攻击者收集的任何统计数据。

[0039] 在另一个实施例中，欺骗引擎190将用户重定向到蜜罐模型。在该示例中，欺骗引擎190使用模拟原始模型的功能的模型，但是用松散地表示原始基线事实的数据来训练，但是足够类似以欺骗攻击者。欺骗引擎190的欺骗响应可帮助使攻击者中的已提取信息无效。在另一个示例中，欺骗引擎190改变模型响应的保真度，而不改变与用户的原始会话流。在该示例中，欺骗引擎190减慢或破坏假设对手中的信息累积。在另一示例中，欺骗引擎190将会话升级给人类响应者。在该示例中，基于用户的风险分数，欺骗引擎190发起通知以激活人类响应者来干预会话。

[0040] 探测器195是程序200的子程序，其使用完全自动化的公共图灵测试(Completely Automated Public Turing test)的隐藏会话概括来区分计算机和人类(验证码)。验证码是一种在计算中用来确定用户是否是人类的质询-响应测试。在一个实施例中，探测器195经由虚拟代理程序175向用户发送探测，评估该探测响应，并根据该响应来更新用户的风险分数。在一个示例中，探测器195基于用户的风险分数注入偶然探测。如果用户风险分数在低风险和某风险之间的边界，则探测器195可以干预虚拟代理程序175和用户的会话，以证明该用户是人类。探测器195将从响应收集的数据和从探测器195导出的信息直接添加到机器人防护数据库182。

[0041] 在一个实施例中，程序200在服务器120上运行。在另一实施例中，程序200在基于服务器的设置中在另一计算机上运行，例如在未示出的服务器计算机上。在又一实施例中，程序200在同时通过网络185与服务器120互连的计算设备110上运行。程序200提供检测和减轻与虚拟代理的对手会话的能力。程序200能够利用Wi-Fi技术、蓝牙、近场通信标签(NFC)、全球移动通信系统(GSM)和全球定位系统技术(GPS)来与计算设备110通信。

[0042] 在示例实施例中，程序200作为计算设备110上的一个或多个应用内的代码片段来

操作。代码片段定义了片段与应用(例如,由服务器120上的网络浏览器应用托管的程序200)之间的交互性的范围。例如,程序200是网络浏览器150内的功能,并且程序200的处理在由程序200启动的网络浏览器150的运行期间自动发生(即,没有用户干预)。动态代码片段元素提供脚本支持。这些变量使得程序200之间能够通过服务器120、图形用户界面130、网络浏览器150和虚拟代理程序175进行会话。

[0043] 在一个实施例中,程序200能够被实现为独立的异常检测系统,其能够与虚拟代理程序175的会话系统进行接口连接以提供会话安全性。程序200通过利用通过会话的先前模型查询的会话上下文,来检测异常和可疑的会话。在一个示例中,程序200可作为虚拟代理的插件来运行,作为对会话日志操作的监视能力。异常检测子系统可以用作独立程序以向操作仪表板馈送异常监视结果。在该示例中,当欺骗引擎190和探测器195正在与用户操纵会话流时,欺骗引擎190和探测器195与会话运行时(runtime)集成或协作。每个子系统是可扩展的,并且能够从与各种攻击者的相遇中学习。可扩展意味着能够将附加的检测模型实现到程序200,能够将附加的缓解响应添加到程序200,并且能够将附加的探测选择添加到程序200。

[0044] 在另一个实施例中,程序200用作模型安全性,以在模型应用程序接口(API)级别监视和检测异常。程序200能够根据用户或机构偏好提供模型特定检测。

[0045] 在一个实施例中,程序200检测用户和虚拟代理程序175之间的相遇的中等风险值。在一个示例中,程序200利用检测器180、欺骗引擎190和探测器195确定用户刚刚通过程序200的用于减轻的阈值。在该示例中,程序200基于咨询机器人防护数据库182来确定哪个减轻过程是对可能可疑用户的最佳响应。程序200分析与相似用户的交互历史,并衡量采用特定探测器的强度和效用与阻止良好用户的风险。

[0046] 图2是描述根据本发明的实施例的程序200的流程图,该程序是用于检测和减轻与虚拟代理的对抗会话的程序。

[0047] 在步骤210中,程序200确定交互的风险值。在一个实施例中,程序200在监视用户和虚拟代理之间的会话的同时使用检测器180分析来自用户的一个或多个话语。在示例实施例中,程序200拦截每个虚拟代理响应,并且提取虚拟代理响应日志条目。程序200使用虚拟代理响应日志条目作为输入来调用“N”个异常检测模型的集合,其中每个模型实现不同的异常检测策略。集合中的各个检测模型并行运行,以单独计算风险值,使用加权的集合函数将该风险值合并到集合风险分数中。该集合可以包括以下模型中的一个或多个以及其它模型:检测不太可能的交互(不可能的转换)的马尔可夫模型、检查时间标签以检测人类的可疑快速回答的定时模型、标记所识别的意图上的低置信度分数的出现的置信度监视模型、以及标识会话进展指示符的不存在或存在的的会话进展模型,诸如会话中的某些目标状态的出现(例如,短语“出售某物”))可以基于与虚拟代理程序175的交互的上下文来提高交互的风险值分数。马尔可夫模型可以使用来自会话状态的转换频率、信息泄露跟踪模型来从会话流图中构建,以跟踪查询在模型的特征空间中对模型的分布,并确定信息暴露的风险级别(即,是足以复制对手的模式功能的所暴露的信息)。

[0048] 在一个示例中,程序200为单独利用马尔可夫检测模块的用户计算风险值。在该示例中,程序200计算用户的高风险值,因为到虚拟代理的会话日志条目是已知攻击的特征。用户的会话日志条目符合已知攻击的模式,并且被认为是侵入性的。程序200使用日志条目

与已知攻击的相似性来计算对用户的高风险值。在一个实施例中，程序200使用加权的集合函数将来自各个检测模型的风险评分合并为单个风险评分R。函数的权重可以是随着时间的推移而改变。程序200使用合并的风险值更新虚拟代理中的用户风险分数。程序200使用会话日志条目递增地更新集合中的所有异常检测模型。在示例中，程序200接收会话日志条目，并且利用两个异常检测模型的组合来计算用户的风险分数。程序200利用定时异常检测模型来确定用户的响应时间与关联于已知攻击者的定时模式一致。程序200基于定时异常检测模型分配个体风险分数“r1”。程序200利用会话进展异常检测模型来确定到虚拟代理的会话日志条目与已知攻击者的会话进展模式一致。程序200基于会话进展异常模型分配个体风险评分“r2”。程序200组合风险评分“r1”和风险评分“r2”以计算组合风险评分值“R＝f(r1,r2)”，其中“f”是加权的集合函数。

[0049] 图3示出了根据本发明的实施例的程序200的示例，该程序运行以检测和减轻对虚拟代理的攻击。在该实施例中，程序200在服务器120上运行，并监视虚拟护士375与被盗模型310、对手315的交互，以允许或允许访问在域模型320中的敏感数据上训练的机密信息。在该实施例中，被盗模型310是机密的诊断信息。对手315是正在发起并继续与虚拟护士会话的用户。程序200是防护对手315和虚拟护士之间的交互的机器人防护。域模型320是患者特有的机密诊断信息。虚拟护士375是与患者交谈以收集症状描述(包括上传的图像)的虚拟代理聊天机器人前端。虚拟护士375使用诊断医学模型来从域模型320给出对患者的机密的最终诊断响应。

[0050] 在一个示例中，如图3中所描绘的，程序200分析由对手315输入到虚拟护士375的话语。在该示例中，程序200探测域模型320以提取用于域模型320的反向工程的信息。程序200确定被盗模型310是经由提取攻击而生成的，并且被盗模型310试图从域模型320训练数据中提取敏感信息。程序200采用先前讨论的一个或多个异常检测模型来确认虚拟护士375处于攻击之下。例如，程序200使用会话进展模型，该会话进展模型识别对手315正在采用规避攻击，因为程序200检测到来自试图欺骗虚拟护士375的对手315的输入。

[0051] 在判定步骤220中，程序200确定风险值是否超过阈值。在一个实施例中，程序200基于通过机器人防护数据库182访问的话语的历史与当前检测到的话语相比来确定(从步骤210)风险分数已经超过阈值。程序200基于在步骤210中作为计算的风险值的函数计算的值，确定风险值是否超过阈值，所述计算的风险值被合并到集合风险分数中以生成风险值“R”。

[0052] 在一个示例中，关于图3，(在步骤210中确定的)程序200同时运行以提供对手315和虚拟护士375之间的会话安全级别的多级别检测以及被盗模型310和域模型320之间的API级别的模型安全级别。程序200评估超过在该示例中由机构所确定的阈值的“R”分数。域模型320是在敏感数据上训练的，并且包含机密信息。程序200还响应于虚拟护士375，基于程序200所确定的对手输入，向对手315分配“R”得分。

[0053] 在一个示例中，虚拟护士375的主人设置机构风险分数阈值3/10。虚拟护士375保护机密的敏感健康信息、域模型320，因此相对于不保护敏感数据的机构阈值，该机构风险分数被设置得较低。10是最高风险分数，并且1是最低风险评分。程序200基于个体值R＝f(r1,r2)分配高风险值R，其中“f”是使用个体异常检测模型计算的加权集合函数。在该示例中，对手315被确定为呈现为攻击者的高风险。在该示例中，程序200通过来自机器人防护数

据库182的历史信息，基于会话的异常上下文来分配会话进展模型分数“r1”。虚拟护士375在健康信息和健康护理数据上被训练，并且对手315正在发起与个人信息(例如疾病和治疗信息)相关的会话。程序200确定对手315提供对虚拟护士375的基线问题的过快响应，如由定时异常检测模型评估和计算的“r2”，程序200计算高风险值“R”＝10/10，其中10超过用于高风险交互的机构阈值3。

[0054] 在步骤230中，程序200允许访问。更具体地，响应于确定风险分数没有超过阈值(判定步骤220，“否”分支)，程序200允许访问虚拟代理程序175(步骤230)。在该示例中，程序200确定在步骤210中确定的风险值不满足阈值“R”。在一个示例中，程序200分析用户对虚拟代理程序175的每个话语，并确定响应是由人作出的。程序200访问机器人防护数据库182，并且查看具有高置信度的被确定为人类的相似话语的历史。基于话语的准确性以及响应与可接受响应的历史的一致性，程序200允许用户访问虚拟程序175。

[0055] 在另一个实施例中，作为步骤220的结果，程序200确定用户的风险分数为低。在示例中，程序200为用户计算1/10的风险分数，其中基于用户呈现的风险来发起动作的机构阈值是5/10。在该示例中，程序200允许用户和虚拟代理继续会话。在计算用户的低风险值之后，程序200进入岗哨模式，同时允许用户与虚拟代理之间的无缝、不间断的会话。程序200继续以岗哨模式监测会话，并且能够基于与虚拟代理的进一步交互来重新计算用户的风险分数。在程序200基于会话的上下文或历史上侵入性的话语模式的出现而再次提高风险分数的情况下，程序200能够响应于重新计算的风险分数而发起减轻动作。

[0056] 在步骤240中，程序200阻止潜在的危害。更具体地，响应于确定风险值超过阈值(判定步骤220，“是”分支)，程序200启动异常减轻子系统(步骤240)。在示例实施例中，异常减轻子系统运行以基于与程序200可能正在阻止“好”用户或人类用户的风险相比，对向用户提供一种形式的阻止的强度和效用的分析来阻止潜在危害并选择潜在路径以阻止潜在危害。

[0057] 在一个实施例中，程序200基于所确定的特定风险水平以及作为所确定的风险水平的函数而被传递的特定阈值，经由欺骗引擎190和探测器195来激活减轻动作。程序200改变对话流，并将与用户的会话重定向到先前确定的对话树的安全区域中。程序200可以相对于R值调整来自虚拟护士375的响应的保真度。在示例中，虚拟护士375具有虚拟护士375对用户的响应的两个保真度水平。虚拟护士375通过程序200启动对用户问题的原始的高保真度的虚拟护士375响应，或者虚拟护士375启动对用户的低保真度响应以减轻用户的高风险响应。程序200激活减轻动作，该减轻动作通过陈述“我没有对此进行训练-为了进一步帮助请拨打1-800…”来终止会话。基于在步骤210中分配的“R”值，程序200每次在用户风险分数超过“高风险”阈值时进行干预。程序200改变对话流并且将虚拟护士375的响应重定向到先前生成的低保真响应以减轻对手315和虚拟护士375之间的交互。在另一示例中，程序200根据用户对探测器的响应或所计算的风险分数来延迟对用户的响应。在该示例中，用户对虚拟护士375的话语变得越来越类似于已知模式的高风险话语。程序200与每个高风险响应成比例地增加用户的风险分数。随着风险分数变高，程序200在将响应发送回用户之前引入更长的延迟。

[0058] 在另一实施例中，程序200能够通过改变虚拟护士375的模型响应的保真度而不改变虚拟护士375的原始对话流程来减轻和阻止潜在攻击。程序200改变虚拟护士375的模型

响应的保真度，以减慢或中断假设对手的信息累积。程序200将每个受保护模型插入较低保真度模型的集合内。在各个实施例中，程序200通过根据当前用户风险分数选择保真度模型来确定实际模型响应。如在步骤210中确定的，用户风险越高，响应水平越低。在一个示例中，保真度水平F被确定为风险评分R的函数。保真度等级1是最高级别，并且与虚拟代理175对具有低风险分数的人类用户的原始响应一致，保真度等级2将更低，保真度等级3甚至更低，直到保真度等级N，如由机构或用户偏好所确定的。

[0059] 在该另一实施例中，程序200能够通过创建导致对所感知的攻击者的较低保真度响应的附加模型来生成较低保真度响应。程序200能够使用渐进模型稀释。稀释模型是使用先前训练的模型作为稀释模型的基础事实的方法。结果，程序200通过欺骗引擎190使先前训练的模型成为原始模型中的基线事实的较不准确的版本。程序200能够无限地将每个低保真度响应链接到逐渐降低的保真度响应。

[0060] 在另一示例实施例中，参考图3，响应于确定风险值超过阈值(判定步骤220，“是”分支)，程序200响应于垃圾邮件攻击执行减轻动作。在该示例中，程序200基于对虚拟护士375呈现的阈值高保真问题的签名响应来确定。程序200通过向对手315提供越来越低保真度的问题来帮助虚拟护士确认对手315是攻击者。程序200向对手315呈现保真度级别3的问题。基于对手315对程序200所呈现的问题的完全不满意的回答，程序200向对手315呈现保真度级别10的问题。基于对程序200所呈现的问题的不可理解的响应，程序200确定对手315是尝试抬高虚拟代理的操作成本的机器人生成的垃圾邮件流量。

[0061] 在另一个实施例中，程序200能够插入随机错误响应以阻止潜在的伤害，并且程序200能够改变虚拟代理程序175与用户交互的音调或方式。在一个示例中，程序200不时地返回随机的不正确响应，以中断攻击者试图收集的任何统计数据。程序200可以根据期望的保真度级别来调整随机响应的速率。

[0062] 在另一示例实施例中，参考图3，响应于确定风险值超过阈值(判定步骤220，“是”分支)，程序200响应于签名专有模型功能提取攻击而执行减轻动作。在该示例中，程序200基于对虚拟护士375的问题和响应的特征来确定对手315正在尝试获取专有模型功能信息。对手315向虚拟护士375提出与虚拟护士375的响应决策树的整个区域相关的一系列快速问题。对手315向虚拟护士375呈现具有“是”响应的一系列回答。然后对手315以“无”响应呈现对虚拟护士375提出的相同问题的相同答案系列。程序200识别签名攻击方法，并且帮助虚拟护士375随机地向对手315提供随机的不相关响应，以防止对手315搜集虚拟护士375的专有模型功能。

[0063] 在另一个实施例中，程序200能够将用户重定向到蜜罐模型。在一个示例中，程序200创建并使用“蜜罐”来模仿原始模型的功能，但是用松散地表示原始地面真实但足够接近以欺骗攻击者的数据来训练。参照图3，程序200能够生成与域模型320类似的虚拟“蜜罐”。在该示例中，攻击者将通过反向工程“蜜罐”来执行提取攻击，从而阻止攻击者实际捕获实际域模型320中的信息。在该示例中，被盗模型310将是欺骗的副本，即“蜜罐”模型。

[0064] 在另外的示例实施例中，关于图3，响应于确定风险值超过阈值(判定步骤220，“是”分支)，程序200响应于对手315的攻击而执行减轻动作，以从虚拟护士375的训练数据提取专有信息。在该示例中，程序200确定对手315正在尝试提取域模型320。程序200确定，基于对手315向虚拟护士375提出的以规避虚拟护士375的安全协议的签名问题，程序200，

通过欺骗引擎190，使虚拟护士375适于将虚拟护士375重新创建为高交互蜜罐。现在用作高交互蜜罐的虚拟护士375收集关于对手315的用于提取信息的工具和技术的深入信息。程序200向对手315呈现真实系统，虚拟护士375临时重新计划为蜜罐模型，给予对手315虚拟护士系统的根特权，并允许对手315访问蜜罐系统。程序200收集关于对手315的提取攻击的详细信息，开发用于提取攻击的签名，并将对手315的简档和攻击方法存储在机器人防护数据库182中。

[0065] 在另一个实施例中，程序200使用探测器195来为用户快速形成风险分数，或者进一步分析不确定的用户，并且根据探测器195的进一步分析来分配风险分数。程序200通过在正常会话中发生的似真性来对可用探测器进行排名。程序200基于用户的当前风险分数注入偶然的探测器，并且程序200可以基于用户的当前风险分数调整探测器的强度和需求以及注入的频率。程序200评估用户对探测器的响应，并相应地更新风险分数，或者程序200可以采用进一步的探测器。程序200根据由探测器195提供的信息将信息添加到机器人防护数据库182。程序200通过探测器195发出请求或其它形式的询问，以从用户取回响应于探测器的更多信息。程序200通过探测器195能够干预用户和虚拟护士375之间的会话，以进一步评估分配给用户的风险分数。在一示例中，程序200插入到对话中，并通过诸如“验证码(Captchas)”等一个或多个探测器直接请求用户证明该用户是人类。

[0066] 在示例中，程序200响应于新用户或没有存储在机器人防护数据库182中的响应的签名历史的用户而采用探测器195。探测器195可以低频率被程序200接合。在另一个示例中，程序200对用户口头表达的话语采用探测器，并插入短语，诸如“抱歉，我没有对此进行训练，你可以换个说法吗?”、“你指的是X?”(其中X是虚拟代理程序175高度确信X不是先前用户话语所关联的事物，即否定确认)，或者程序200可利用探测器195来采用需要多于“是”或“否”回答且与当前上下文相关的多余问题。(例如，在关于汽车保险的会话中的“你什么时候第一次获得你的汽车”)。

[0067] 在另外的示例实施例中，关于图3，响应于确定风险值超过阈值(判定步骤220，“是”分支)，程序200响应于中毒攻击执行减轻动作。在该示例中，虚拟护士375连续地通过生产使用来学习。虚拟护士375与“良好”用户的交互越多，虚拟护士375就运作地越好，并且虚拟护士375进化地越多。在该示例中，程序200基于对手315的话语确定对手315正在改变虚拟护士375的训练数据。程序200确定由对手315模仿并继续的会话主题正在驱动最初被分配给健康护理主题的会话偏离路线进入不相关的主题区域。作为响应，程序200引导虚拟护士375进入“安全模式”。虚拟护士375不向对手315泄露任何机密的、敏感的或专有的信息，包括任何专有模型功能，例如域模型320。在对手315继续尝试“毒害”虚拟护士375的情况下，程序200终止连接，从而中断虚拟护士375和对手315之间的会话。

[0068] 在另一示例实施例中，参考图3，响应于确定风险值超过阈值(判定步骤220，“是”分支)，程序200通过组合响应于用户而降低保真度、向用户给出错误响应、使用蜜罐欺骗方法、以及使用一个或多个探测器来更新用户的风险分数的缓解动作，来执行缓解动作。

[0069] 图4描述了根据本发明的说明性实施例的服务器120的组件的框图。应当理解，图4仅提供了一种实现的说明，而不暗示对其中可实现不同实施例的环境的任何限制。可以对所描述的环境进行许多修改。

[0070] 服务器120包括通信结构402，其提供高速缓存416、存储器406、永久性存储408、通

信单元410和输入/输出(I/O)接口412之间的通信。通信结构402可以用被设计成在处理器(诸如微处理器、通信和网络处理器等)、系统存储器、外围设备和系统内的任何其它硬件组件之间传递数据和/或控制信息的任何体系结构来实现。例如,通信结构402可以用一个或多个总线或纵横开关来实现。

[0071] 存储器406和永久性存储408是计算机可读存储介质。在该实施例中,存储器406包括随机存取存储器(RAM)。通常,存储器406可以包括任何合适的易失性或非易失性计算机可读存储介质。高速缓存416是通过保存来自存储器406的最近访问的数据和接近被访问数据的数据来增强计算机处理器404的性能的快速存储器。

[0072] 程序200可以被存储在永久性存储408和存储器406中,以便由相应计算机处理器404中的一个或多个经由高速缓存416执行。在一个实施例中,永久性存储408包括磁硬盘驱动器。作为磁硬盘驱动器的替代或补充,永久性存储408可包括固态硬盘驱动器、半导体存储设备、只读存储器(ROM)、可擦除可编程只读存储器(EPROM)、闪存、或能够存储程序指令或数字信息的任何其它计算机可读存储介质。

[0073] 永久性存储408所使用的介质也可以是可移动的。例如,可移动硬盘驱动器可以用于永久性存储408。其它示例包括光盘和磁盘、拇指驱动器和智能卡,它们被插入到驱动器中以便传送到也是永久性存储408的一部分的另一计算机可读存储介质上。

[0074] 在这些示例中,通信单元410提供与其他数据处理系统或设备的通信。在这些示例中,通信单元410包括一个或多个网络接口卡。通信单元410可以通过使用物理和无线通信链路中的一种或两种来提供通信。程序200可以通过通信单元410下载到永久存储器408。

[0075] I/O接口412允许与可连接到服务器120的其它设备输入和输出数据。例如,I/O接口412可以提供到诸如键盘、小键盘、触摸屏和/或一些其它合适的输入设备的外部设备418的连接。外部设备418还可以包括便携式计算机可读存储介质,例如拇指驱动器、便携式光盘或磁盘、以及存储卡。用于实践本发明的实施例的软件和数据(例如程序200)可以存储在这样的便携式计算机可读存储介质上,并且可以经由(一个或多个)I/O接口412加载到永久性存储408上。I/O接口412也连接到显示器420。显示器420提供向用户显示数据的机制,并且可以是例如计算机监视器。

[0076] 这里描述的程序是基于在本发明的特定实施例中实现它们的应用来标识的。然而,应当理解,这里的任何特定程序术语仅是为了方便而使用,因此本发明不应当限于仅在由这样的术语标识和/或暗示的任何特定应用中使用。

[0077] 本发明可以是系统、方法、和/或计算机程序产品。计算机程序产品可以包括计算机可读存储介质,其上载有用于使处理器实现本发明的各个方面的计算机可读程序指令。

[0078] 计算机可读存储介质可以是可以保持和存储由指令执行设备使用的指令的有形设备。计算机可读存储介质例如可以是——但不限于——电存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或者上述的任意合适的组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括:便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、静态随机存取存储器(SRAM)、便携式压缩盘只读存储器(CD-ROM)、数字多功能盘(DVD)、记忆棒、软盘、机械编码设备、例如其上存储有指令的打孔卡或凹槽内凸起结构、以及上述的任意合适的组合。这里所使用的计算机可读存储介质不被解释为瞬时信号本身,诸如无线电波或者其他自由传播的电磁波、通

过波导或其他传输媒介传播的电磁波(例如,通过光纤电缆的光脉冲)、或者通过电线传输的电信号。

[0079] 这里所描述的计算机可读程序指令可以从计算机可读存储介质下载到各个计算/处理设备,或者通过网络、例如因特网、局域网、广域网和/或无线网下载到外部计算机或外部存储设备。网络可以包括铜传输电缆、光纤传输、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配卡或者网络接口从网络接收计算机可读程序指令,并转发该计算机可读程序指令,以供存储在各个计算/处理设备中的计算机可读存储介质中。

[0080] 用于执行本发明操作的计算机程序指令可以是汇编指令、指令集架构(ISA)指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据、或者以一种或多种编程语言的任意组合编写的源代码或目标代码,所述编程语言包括面向对象的编程语言—诸如Smalltalk、C++等,以及常规的过程式编程语言—诸如“C”语言或类似的编程语言。计算机可读程序指令可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中,远程计算机可以通过任意种类的网络—包括局域网(LAN)或广域网(WAN)—连接到用户计算机,或者,可以连接到外部计算机(例如利用因特网服务提供商来通过因特网连接)。在一些实施例中,通过利用计算机可读程序指令的状态信息来个性化定制电子电路,例如可编程逻辑电路、现场可编程门阵列(FPGA)或可编程逻辑阵列(PLA),该电子电路可以执行计算机可读程序指令,从而实现本发明的各个方面。

[0081] 这里参照根据本发明实施例的方法、装置(系统)和计算机程序产品的流程图和/或框图描述了本发明的各个方面。应当理解,流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合,都可以由计算机可读程序指令实现。

[0082] 这些计算机可读程序指令可以提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器,从而生产出一种机器,使得这些指令在通过计算机或其它可编程数据处理装置的处理器执行时,产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中,这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作,从而,存储有指令的计算机可读介质则包括一个制造品,其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

[0083] 也可以把计算机可读程序指令加载到计算机、其它可编程数据处理装置、或其它设备上,使得在计算机、其它可编程数据处理装置或其它设备上执行一系列操作步骤,以产生计算机实现的过程,从而使得在计算机、其它可编程数据处理装置、或其它设备上执行的指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

[0084] 附图中的流程图和框图显示了根据本发明的多个实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上,流程图或框图中的每个方框可以代表一个模块、程序段或指令的一部分,所述模块、程序段或指令的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。在有些作为替换的实现中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,两个连续的方框实际上可以基本并行地执

行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的是，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或动作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

[0085] 以上已经描述了本发明的各实施例，上述说明是示例性的，并非穷尽性的，并且也不限于所披露的各实施例。在不偏离本发明范围的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择，旨在最好地解释各实施例的原理、实际应用或对市场中的技术的技术改进，或者使本技术领域的其它普通技术人员能理解本文披露的各实施例。

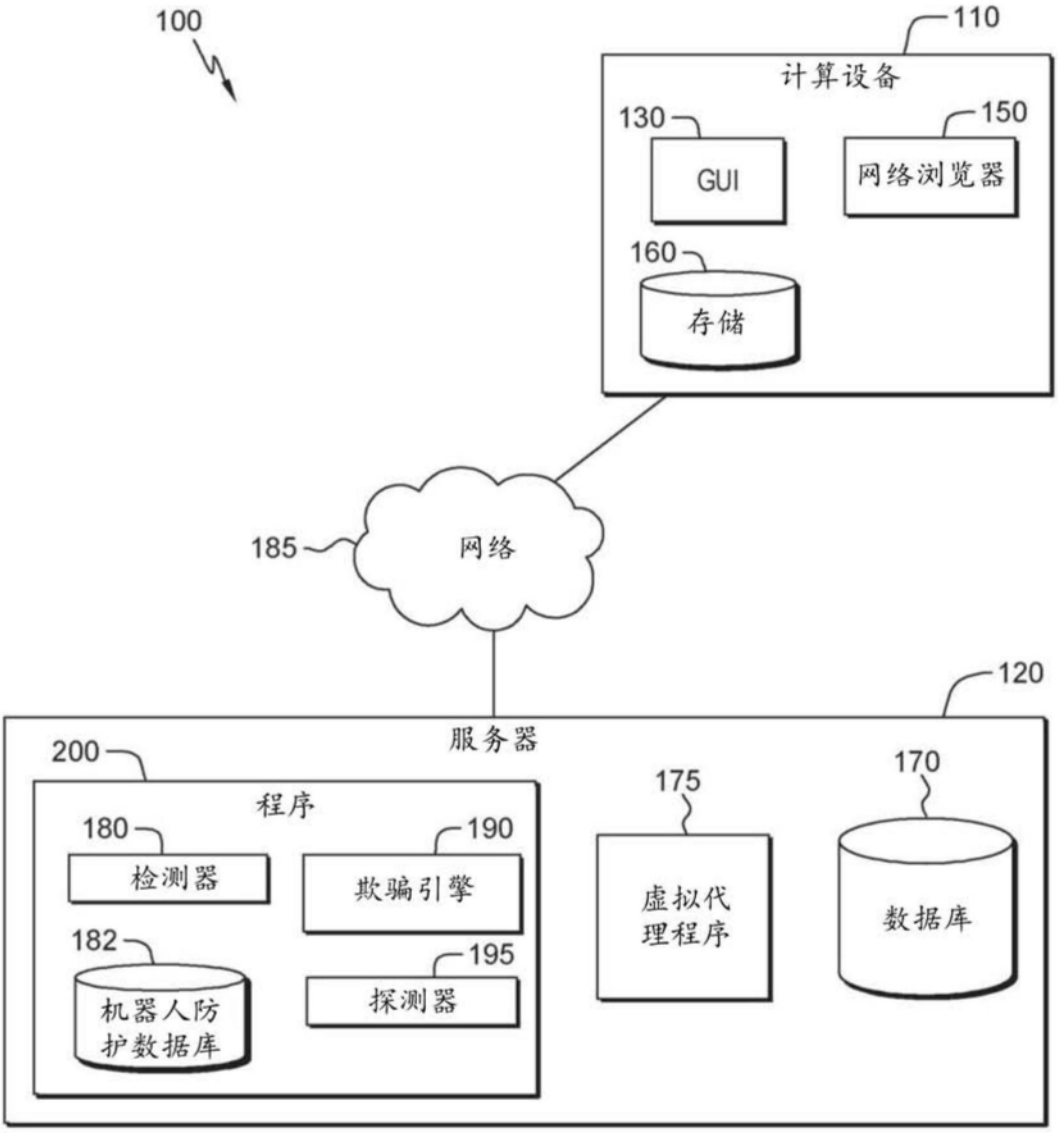

100
110
计算设备
130
GUI
150
网络浏览器
160
存储
185
网络
120
服务器
200
程序
180
检测器
190
欺骗引擎
182
机器人防护数据库
195
探测器
175
虚拟代理程序
170
数据库

图1

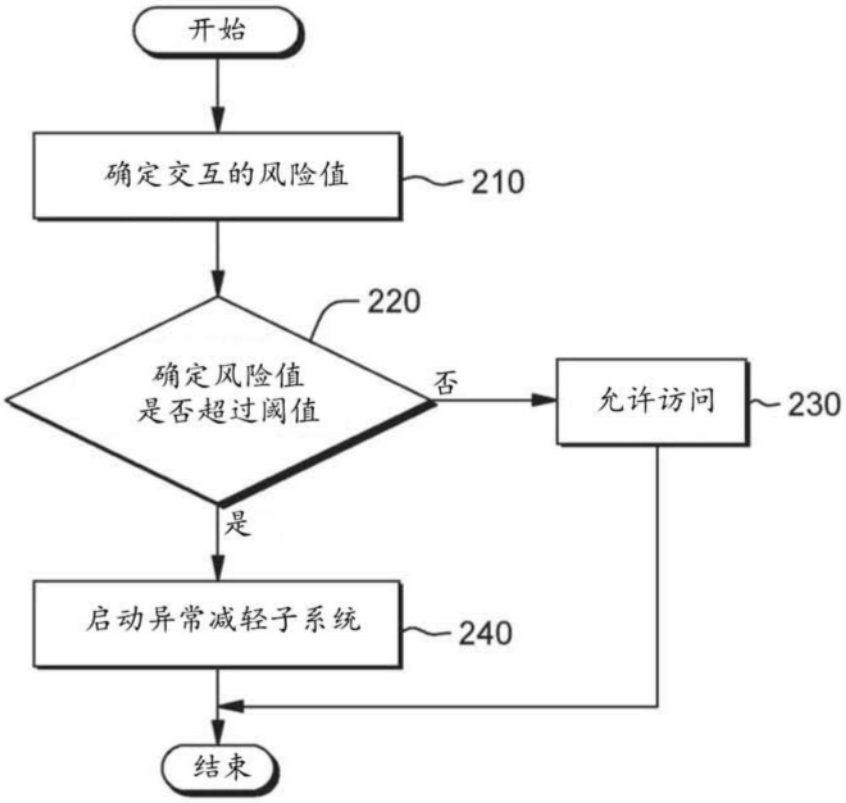

开始
确定交互的风险值
210
220
确定风险值
是否超过阈值
否
允许访问
230
是
启动异常减轻子系统
240
结束

图2

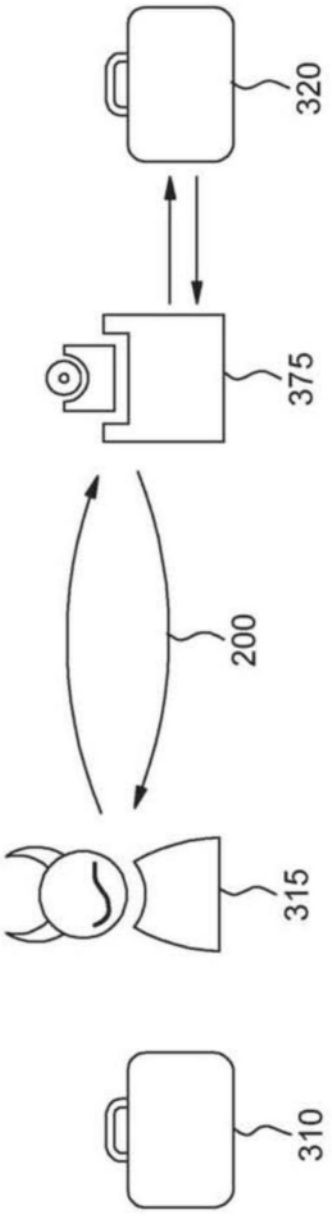


图3

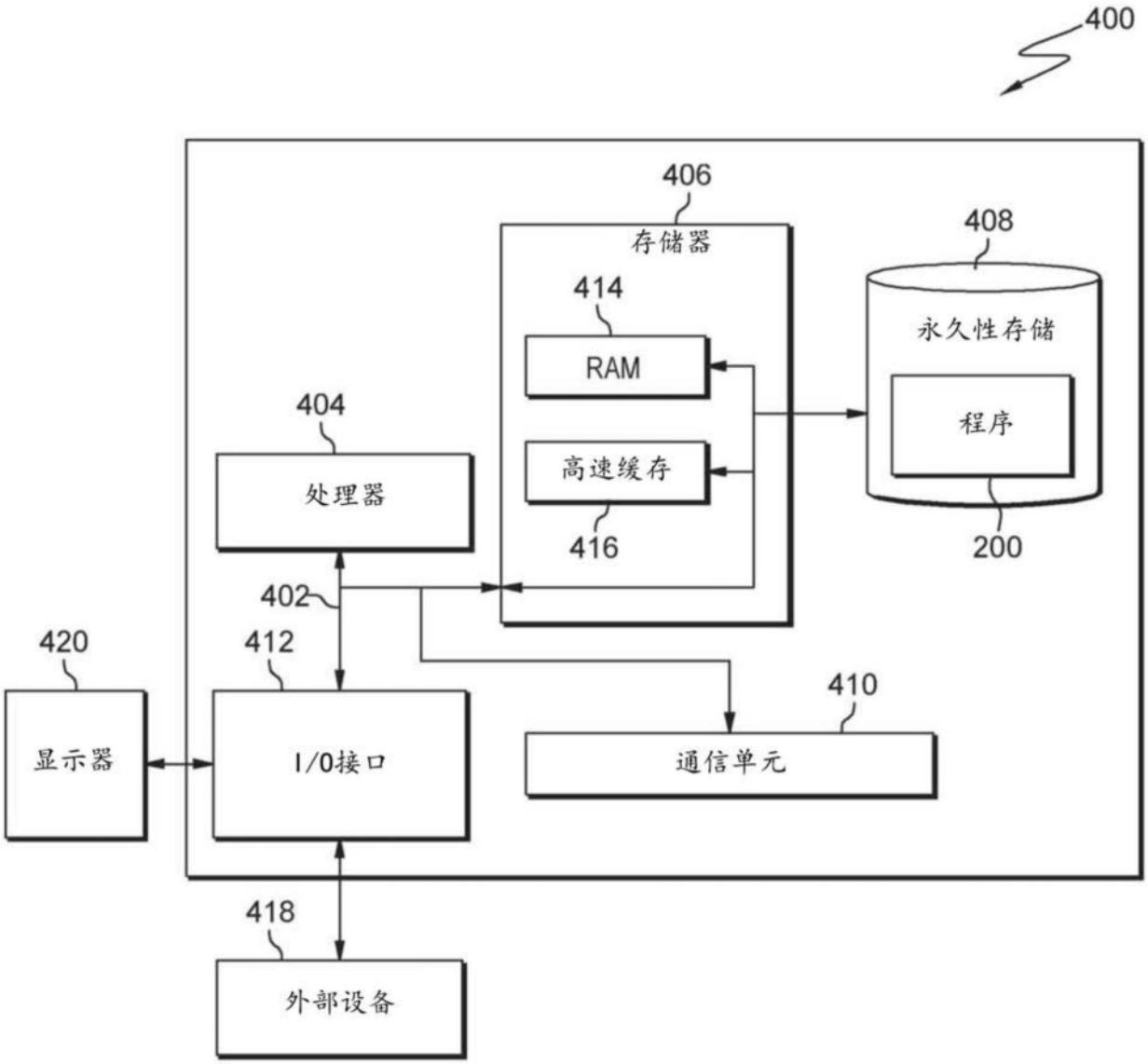

400
406
存储器
408
414
RAM
永久性存储
404
处理器
高速缓存
程序
416
200
402
420
412
410
显示器
I/O接口
通信单元
418
外部设备

图4