



(12) 发明专利申请

(10) 申请公布号 CN 119072701 A

(43) 申请公布日 2024. 12. 03

(21) 申请号 202380038476.1

(22) 申请日 2023.05.10

(30) 优先权数据

63/340,883 2022.05.11 US

(85) PCT国际申请进入国家阶段日

2024.11.05

(86) PCT国际申请的申请数据

PCT/EP2023/062432 2023.05.10

(87) PCT国际申请的公布数据

W02023/217868 EN 2023.11.16

(71) 申请人 渊慧科技有限公司

地址 英国

(72) 发明人 约拉姆·巴赫拉赫

安德里亚·塔凯蒂

伊恩·迈克尔·根普

雅诺什·克拉马尔

马特乌什·马利诺夫斯基

凯文·罗伯特·麦基

(74) 专利代理机构 中原信达知识产权代理有限

责任公司 11219

专利代理师 周亚荣 朴金丹

(51) Int.Cl.

G06Q 10/04 (2006.01)

G06N 3/02 (2006.01)

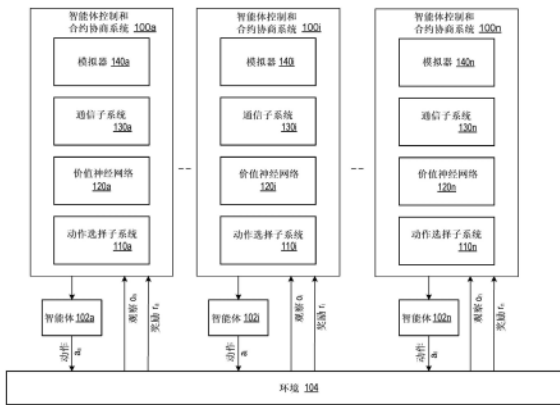
权利要求书4页 说明书19页 附图6页

(54) 发明名称

多智能体系统中智能体合作的合约协商

(57) 摘要

方法、系统和装置,包括在计算机存储介质上编码的计算机程序,用于使智能体能够以改进其集体效率的方式相互合作。智能体可以通过考虑其他智能体的行为来修改其行为,从而可以实现比每个智能体独立行动更好的总体结果。这是通过使智能体能够相互协商限制其相应的动作的合约来实现的。



1. 一种控制第一智能体在环境中执行任务的计算机实现的方法, 在所述环境中所述第一智能体与执行一个或多个其他任务的一个或多个其他智能体交互,

其中, 所述第一智能体具有动作选择子系统和价值神经网络, 所述动作选择子系统被配置为处理表征所述环境的状态的状态表示以生成用于选择要由所述智能体执行的动作的策略输出, 所述价值神经网络被配置为处理所述状态表示以确定状态值, 其中, 所述状态值定义所述智能体的从由所述状态表示表征的所述环境的所述状态的期望回报;

其中, 所述第一智能体在多个时间步中的每个时间步处:

获得表征当前时间步处的所述环境的所述状态的状态表示;

使用所述智能体的所述动作选择子系统处理所述状态表示以生成针对所述第一智能体的所述策略输出;

使用所述状态表示确定所述其他智能体的预测动作; 以及

使用针对所述第一智能体的所述策略输出, 取决于所述其他智能体的所述预测动作来选择供所述第一智能体执行的动作;

所述方法包括, 在所述时间步中的一个或多个时间步处:

所述第一智能体与第二智能体协商合约, 其中, 所述第二智能体是所述其他智能体中的一个, 所述合约定义了对可能由所述第一智能体选择的动作的限制以及对可能由所述第二智能体选择的动作的限制;

其中, 协商所述合约包括所述第一智能体通过以下操作来确定是否提议合约:

通过模拟 (i) 在有所提议的合约的情况下以及 ii) 在考虑所提议的合约中对所述动作的所述限制的影响的情况下所述第一智能体和所述其他智能体的未来动作的影响, 来确定一个或多个得分, 所述一个或多个得分表示在有所提议的合约的情况下和没有所提议的合约的情况下, 在后续时间步处的对所述第一智能体的所述环境的所述状态的期望值, 以及

取决于所述一个或多个得分将所提议的合约传送给所述第二智能体; 以及

响应于从所述第二智能体接收到接受所提议的合约的指示, 考虑所提议的合约中对所述第一智能体的所述动作的所述限制来选择供所述第一智能体执行的动作。

2. 根据权利要求1所述的方法, 进一步包括: 所述第一智能体确定是否接受所提议的合约; 并且其中, 在考虑所提议的合约中对所述第一智能体的所述动作的所述限制的情况下来选择所述第一智能体的所述动作进一步依赖于所述第一智能体接受所提议的合约。

3. 根据权利要求1或权利要求2所述的方法, 其中, 确定所述后续时间步处的对所述第一智能体的所述环境的所述状态的所述期望值包括: 使用所述动作选择子系统对所述第一智能体在该时间步处的一个或多个可能的动作进行采样, 并且, 对于所述一个或多个动作中的每个动作并且对于所述其他智能体中的每个其他智能体:

对在该时间步处该其他智能体的多个可能的动作进行采样; 以及

使用所述价值神经网络确定由所述智能体的所采样的动作确定的所述环境的预测的新状态的状态值; 并且其中

模拟所述未来动作的所述影响包括: 在所述其他智能体中的每个其他智能体的所采样的动作上对每个预测的新状态的状态值取平均, 以针对所述第一智能体的所采样的一个或多个可能的动作中的每个动作, 确定对所述第一智能体的所述环境的所述状态的所述期望

值。

4. 根据权利要求1、权利要求2或权利要求3所述的方法,其中,模拟所述智能体的所述未来动作包括,对于每个智能体:

通过i) 使用所述动作选择子系统,和或ii) 在所提议的合约应用于该智能体的情况下通过应用所提议的合约中的对所述动作的所述限制,针对该智能体选择一个或多个候选动作的集合。

5. 根据权利要求1至4中任一项所述的方法,其中,所述合约定义了动作的类,所述动作对可能由所述第一智能体和所述第二智能体两者选择的所述动作定义了相同或对应的限制;所述方法进一步包括:

确定第一得分和第二得分,所述第一得分表示在有所提议的合约的情况下在后续时间步处的对所述第一智能体的所述环境的所述状态的期望值,所述第二得分表示在没有所提议的合约的情况下在后续时间步处的对所述第一智能体的所述环境的所述状态的期望值;

比较所述第一得分和所述第二得分;以及

当所述第一得分超过所述第二得分时,确定传送所提议的合约。

6. 根据权利要求1至4中任一项所述的方法,其中,所述合约定义了必须由所述第一智能体选择的第一动作和必须由所述第二智能体选择的第二动作;其中,所述一个或多个得分包括协定得分,所述协定得分表示当所述第一智能体和所述第二智能体分别选择所述第一动作和所述第二动作时,在所述后续时间步处的对所述第一智能体和所述第二智能体两者的所述环境的所述状态的期望值,所述方法进一步包括:

确定相应的第一基线值和第二基线值,所述第一基线值和所述第二基线值分别表示在没有将所述第二智能体限制于所述第二动作的情况下以及在没有将所述第一智能体限制于所述第一动作的情况下,在所述后续时间步处的对所述第一智能体和所述第二智能体的所述环境的所述状态的所述期望值;以及

确定相对于所述基线值的所述协定得分。

7. 根据权利要求6所述的方法,其中,所述协定得分包括以下项的乘积:i) 当所述第一智能体和所述第二智能体分别选择所述第一动作和所述第二动作时,在所述后续时间步处的对所述第一智能体的所述环境的所述状态的所述期望值与所述第一基线值的差;以及ii) 当所述第一智能体和所述第二智能体分别选择所述第一动作和所述第二动作时,在所述后续时间步处的对所述第二智能体的所述环境的所述状态的所述期望值与所述第二基线值之间的差。

8. 根据权利要求6或权利要求7所述的方法,其中,所述第一基线值包括当所述第一智能体选择所述第一动作并且由所述第二智能体选择的动作不受限制时,在所述后续时间步处的对所述第一智能体的所述环境的所述状态的所述期望值;并且其中,所述第二基线值包括当所述第二智能体选择所述第二动作并且由所述第一智能体选择的动作不受限制时,在所述后续时间步处对所述第二智能体的所述环境的所述状态的所述期望值。

9. 根据权利要求6或权利要求7所述的方法,

其中,确定所述第一基线值包括:对于除了包括所述第一智能体和所述第二智能体的对之外的每个智能体对,针对所述后续时间步识别它们是否会达成合约,并且,当它们会达成合约时,将所述第一基线值确定为在如同所达成的合约已经确立一样的情况下在所述后

续时间步处的对所述第一智能体的所述环境的所述状态的所述期望值;以及

其中,确定所述第二基线值包括:对于除了包括所述第一智能体和所述第二智能体的对之外的每个智能体对,针对所述后续时间步识别它们是否会达成合约,并且,当它们将达成合约时,将所述第二基线值确定为在如同所达成的合约已经确立的情况下在所述后续时间步处的对所述第二智能体的所述环境的所述状态的所述期望值。

10.根据权利要求6至权利要求9中任一项所述的方法,进一步包括:通过确定多个候选提议合约并且选择具有最高协定得分的候选提议合约作为所提议的合约来确定针对所述第二智能体的所提议的合约。

11.根据权利要求6至权利要求10中任一项所述的方法,进一步包括:通过确定多个候选提议合约并且选择所述候选提议合约之一作为所提议的合约来确定所述第二智能体的所提议的合约,其中,确定所述候选提议合约包括:

通过使用所述动作选择子系统处理所述当前时间步处的所述状态表示以生成针对所述第一智能体的所述策略输出来确定所述第一智能体的第一候选动作集合;

通过使用所述动作选择子系统处理所述当前时间步处的所述状态表示以生成针对所述第二智能体的所述策略输出来确定所述第二智能体的第二候选动作集合;

生成动作对,所述动作对包括来自所述第一候选动作集合的第一候选动作和来自所述第二候选动作集合的第二候选动作;以及

通过经由以下操作来选择候选动作对的集合来确定所述候选提议合约:根据对度量来对所述动作对进行排名并且选择K个最高排名的动作对;

其中,每个候选提议合约包括所述候选动作对之一;以及

其中,动作对的所述对度量包括以下项的组合:当所述第一智能体执行所述动作对中的所述第一候选动作时,在所述后续时间步处的对所述第一智能体的所述环境的所述状态的所述期望值,以及当所述第二智能体执行所述动作对中的所述第二候选动作时,在所述后续时间步处的对所述第二智能体的所述环境的所述状态的所述期望值。

12.根据权利要求11中所述的方法,其中,所述组合是加权组合,以对所述第二智能体的所述提议合约的值进行加权。

13.根据权利要求1至12中任一项所述的方法,进一步包括:

由所述第一智能体检测到所述第二智能体在违约时间步处已经违反所述合约;并且,作为响应,

制裁所述第二智能体,其中,所述制裁包括所述第一智能体在所述违约时间步之后的时间步处修改与所述第二智能体的合约的所述协商。

14.根据权利要求1至13中任一项所述的方法,其中,所述环境是真实世界环境,其中,至少所述第一智能体包括机器人或自动驾驶载具,其中,所述任务和所述其他任务中的每一个包括导航通过所述环境的路径,其中,所述动作包括控制所述智能体在所述环境中的移动的动作,并且其中,所述期望回报与以下中的一个或多个的度量有关:所述任务的效能;以及执行所述任务的物理成本。

15.根据权利要求1至13中任一项所述的方法,其中,所述环境是计算环境,其中,所述智能体中的每个智能体包括计算作业调度智能体,其中,所述任务和所述其他任务中的每一个包括执行计算机程序,其中,动作包括将一个或多个作业分配给一个或多个计算硬件

资源的动作,并且其中,所述期望回报与执行所述任务的时间度量、执行所述任务的能量成本、执行所述任务的计算成本以及执行所述任务的可靠性有关。

16. 根据权利要求1至13中任一项所述的方法,其中,所述环境是分组通信网络环境,其中,所述智能体中的每个智能体包括用于通过所述通信网络路由数据分组的路由器,其中,所述任务和所述其他任务中的每一个包括分组路由任务,其中,动作包括用于路由所述数据分组的路由动作,并且其中,所述期望回报与一个或多个分组路由度量有关。

17. 根据权利要求1至13中任一项所述的方法,其中,所述环境是电力分布环境,其中,所述智能体中的每个智能体被配置为控制电力从与所述智能体相关联的节点通过一个或多个电力分布链路到一个或多个其他节点的路由,其中,所述任务和所述其他任务中的每一个包括将电力从发电机分布到电力消费者的任务,其中,动作包括用于控制所述节点之间的电力的所述路由的控制动作,并且其中,所述期望回报与所述电力分布链路中的一个电力分布链路上的损失有关、或与关于所述电力分布链路中的一个电力分布链路的频率或相位不匹配有关、或与所述电力分布链路中的一个电力分布链路的过载有关。

18. 根据权利要求1至13中任一项所述的方法,其中,所述环境是真实世界的制造环境,其中,至少所述第一智能体包括控制系统,所述控制系统被配置为控制机械、化学或生物产品的制造,其中,所述任务和所述其他任务中的每一个包括制造所述机械、化学或生物产品或另一机械、化学或生物产品或其中间体或组件的任务,其中,动作包括控制动作,所述控制动作用于控制机器或制造单元的使用以用于处理固体或液体材料以制造所述产品或其中间体或组件,或控制所述产品或其中间体或组件在所述制造环境中的移动,并且其中,所述期望回报与以下中的一个或多个的度量有关:所述任务的效能;以及执行所述任务的物理成本。

19. 计算机可读指令,或存储计算机可读指令的一个或多个计算机存储介质,所述计算机可读指令在由一个或多个计算机执行时,使所述一个或多个计算机实施根据权利要求1至18中任一项所述的方法。

20. 一种系统,所述系统包括一个或多个计算机和存储指令的一个或多个存储设备,所述指令在由一个或多个计算机执行时,使所述一个或多个计算机执行根据权利要求1至18中任一项所述的方法的相应操作。

多智能体系统中智能体合作的合约协商

背景技术

[0001] 本说明书涉及多智能体系统中的智能体合作。

[0002] 系统使用神经网络。神经网络是机器学习模型,其采用一个或多个非线性单元层来预测接收到的输入的输出。一些神经网络是深度神经网络,其除了输出层之外还包括一个或多个隐藏层。每个隐藏层的输出用作网络中下一层——即下一个隐藏层或输出层——的输入。网络的每一层根据相应参数集合的当前值从接收到的输入生成输出。

发明内容

[0003] 本说明书描述了使智能体能够以改进其集体效率的方式相互合作的技术。智能体可以考虑其他智能体的行为来修改其行为,从而可以实现比每个智能体独立行动更好的总体结果。

[0004] 在一个方面,描述了一种控制第一智能体在环境中执行任务的计算机实施的方法,在该环境中第一智能体与环境中执行一个或多个其他任务的一个、两个或更多个其他智能体交互。

[0005] 第一智能体在多个时间步中的每个时间步处获得表征环境状态的状态表示,并使用动作选择子系统处理状态表示以生成策略输出,使用状态表示确定其他智能体的预测动作,并使用策略输出并取决于预测动作来选择要执行的动作。

[0006] 在一个或多个时间步处,第一智能体与第二智能体协商合约(contract),该合约定义了对可以由第一智能体选择的动作的限制以及对可以由第二智能体选择的动作的限制。

[0007] 这涉及第一智能体通过确定一个或多个得分来确定是否提议合约,所述一个或多个得分表示在提议合约的情况下有和没有提议合约的情况下,在后续时间步处的对第一智能体的环境状态的期望值。这是通过模拟i)在没有提议合约的情况下和ii)在考虑提议合约中对动作的限制的影响的情况下第一智能体和其他智能体的未来动作的影响来完成的。取决于得分来将提议合约传送给第二智能体,并且响应于从第二智能体接收到接受提议合约的指示,考虑提议合约中对第一智能体的动作的限制,选择第一智能体要执行的动作。

[0008] 可以实施本说明书中描述的主题的特定实施例,以实现以下一个或多个优点。

[0009] 为智能体配备如上所述的协商合约的能力有助于智能体学习合作以执行其相应的任务。以该方式配备的智能体可以能够执行它们原本无法执行的任务,或者它们可以能够更有效地执行任务,例如比原本更快或消耗更少的资源。一般而言,有通信的智能体的表现优于非通信的智能体,并且可以通过合作存在可观的增益。

[0010] 所述技术有助于就有利于双方的合约达成一致。例如,如果合约仅有利于第一智能体,则第二智能体不太可能达成该合约。由于所述技术有助于识别有利于双方的合约,因此增强了智能体当中进行有益合作的可能性。智能体还可以施加制裁以阻碍破坏达成的合约。在实施方式中,智能体能够在环境中同时行动。

[0011] 技术的实施方式不依赖于可能表示单点故障的中央协调机构。代替地,所述技术

可以在分散的——例如对等——设置中实现。这有助于在广泛的应用中部署技术,并且还可以帮助提供鲁棒性。此外,智能体之间为达成合约而进行的通信量可以相对小。

[0012] 所述技术可用于动作空间非常大的环境中,例如组合环境。例如,每个智能体可能具有可以在每个时间步执行的大量可能的动作。这导致潜在合约的巨大空间。所述技术可以在这样的动作空间中有效使用。

[0013] 在附图和以下描述中阐述本说明书主题的一个或多个实施例的细节。主题的其他特征、方面和优点根据说明书、附图和权利要求将变得显而易见。

附图说明

[0014] 图1示出了多智能体系统的示例。

[0015] 图2是用于控制第一智能体的示例过程的流程图。

[0016] 图3是用于确定是否提议合约的第一示例过程的流程图。

[0017] 图4是用于确定是否提议合约的第二示例过程的流程图。

[0018] 图5是用于选择合约来评估以供提议的示例过程的流程图。

[0019] 图6示出了使用第一示例过程可以实现的益处。

[0020] 图7示出了使用第二示例过程可以实现的益处。

[0021] 各个图中的相同参考编号和名称指示相同的元素。

具体实施方式

[0022] 本说明书总体上描述了一种控制第一智能体在环境中执行任务的计算机实现的方法,在该环境中第一智能体与环境中执行一个或多个其他任务的一个、两个或更多个其他智能体交互。其他智能体的操作可以由计算机实现的方法例如类似于控制第一智能体的方法来控制,或者部分或全部可以部分或整体地在人为控制下。

[0023] 通常,任务可能具有相同的特点,例如它们可能都是路由任务,或者它们可能都是调度任务,或者它们可能都是制造任务。其他任务可能与第一智能体的任务相同或不同。由智能体执行的任务可能有助于实现共同目标。在典型的实施方式中,第一智能体在环境中执行任务的能力受到在环境中执行一个或多个其他任务的一个或多个其他智能体的影响。一些示例任务将在稍后更详细地描述。

[0024] 第一智能体以及在实施方式中其他智能体被使得能够就联合计划达成合约,并且以该方式可以胜过独立操作的智能体。合约对订立合约的智能体的动作施加了限制,并且就合约达成一致涉及协商,即智能体之间的通信。合约可以反映合作与竞争之间的平衡,并且所描述的技术有助于在许多真实世界领域中的操作,在该许多真实世界领域中智能体进行交互,即其中一个智能体的动作会影响另一个智能体,并且其中智能体的目标可能仅部分一致。

[0025] 这样的多智能体系统在真实世界中很常见,并且例如包括在执行任务时进行交互的机器人或自主或半自主载具系统,诸如仓库机器人的系统;工厂或厂房自动化系统;以及计算机系统。因此,作为示例,智能体可以包括机器人或载具、工厂或厂房中的设施项目,或计算机系统例如控制任务到硬件项目的分配或通信网络上的数据路由的软件智能体。

[0026] 图1示出了多智能体系统的示例,该多智能体系统包括智能体102a...n,每个智能

体由相应的智能体控制和合约协商系统100a...n控制。每个智能体控制和合约协商系统100a...n可以作为一个或多个计算机程序在一个或多个位置的一个或多个计算机上实现。在所示示例中,每个智能体都是计算机控制的,但在其他示例中,一个或多个智能体可以是人为控制的。

[0027] 智能体102a...n在公共环境104中操作,每个智能体执行相应的任务。由相应智能体执行的任务可以相同,也可以彼此不同。一般而言,一个智能体执行其任务的方式会影响另一个智能体能够执行其任务的方式。为方便起见,以下描述指的是智能体102a(也称为第一智能体)的智能体控制和合约协商系统100a;在实施方式中,其他智能体控制和合约协商系统是类似的。

[0028] 在多个动作选择时间步中的每个动作选择时间步处,智能体控制和合约协商系统100a响应于表征环境状态的状态表示来选择要由智能体102a执行的动作 a_a 。在实施方式中,从对环境状态的观察 O_a 得出状态表示。例如,观察可以包括环境的图像和/或其他传感器数据,例如表示一个或多个智能体的定位、状态或配置的数据,或表示包括环境中的对象的环境的状态或配置的数据;或者通常来自环境的输入数据。在一些实施方式中,可以例如通过神经网络处理观察以获得状态表示;在一些实施方式中,观察本身可以用作状态表示。

[0029] 智能体控制和合约协商系统100a还可以因执行动作 a_a 而接收奖励 r_a 。通常,奖励是数值,即标量,并且可以基于环境的任何事件或方面。例如,奖励 r_a 可以指示智能体102a是否已完成任务,或者智能体102a在完成任务方面的进展(例如,一个或多个对象的物理配置,或导航到环境中的目标位置)。

[0030] 在实施方式中,智能体控制和合约协商系统100a包括动作选择子系统110a,其用于选择由智能体102a执行的动作。动作选择子系统110a被配置为处理表征环境在(当前)时间步的状态的状态表示,例如从对环境的观察中得出的状态表示,以生成用于选择要由智能体执行的动作的策略输出。如前所述,每个智能体102可以具有相关联的动作选择子系统110。

[0031] 有很多种策略输出可用于选择动作的方式。例如,策略输出可以定义可以由智能体执行的动作集合的概率分布。然后可以使用概率分布来选择动作,例如,通过从概率分布中采样或选择具有最高概率的动作。策略输出可以参数化这样的概率分布,或者它可以将概率分布定义为可以根据其选择动作的得分集合,例如可能动作集合中的每个动作的得分。作为另一个示例,策略输出可以直接定义动作,例如通过识别机械动作的速度或扭矩。通常,动作可以是连续的或离散的;可选地,连续动作可以被离散化。动作可以包括在时间步处要执行的多个个体或基原动作,例如连续动作和离散动作的混合。在一些实施方式中,策略输出可以包括多个输出,例如来自神经网络上的多个头的用于在特定时间步处选择多个动作的多个输出。

[0032] 通常,任何类型的动作选择子系统都可以用于动作选择子系统110a。在一些实施方式中,但不是必须的,动作选择子系统110a包括动作选择策略神经网络。例如,动作选择子系统可以使用动作选择策略神经网络,该动作选择策略神经网络被配置为根据动作选择策略神经网络参数处理表征环境状态的状态表示,以生成可以是策略输出的动作选择神经网络输出。在一些实施方式中,动作选择策略神经网络的动作选择神经网络输出可以间接用于生成策略输出,例如通过使用动作选择策略神经网络来模拟用于确定策略输出的可能

未来。例如,动作选择子系统可以实现“采样的最佳响应(Sampled Best Response)”算法,其在Anthony, Eccles, Tacchetti, Kramár, Gemp, Hudson, Porcel, Lanctot, Julien, Everett, Singh, Graepel, 和Yoram的arXiv:2006.04635v4中进行了描述。

[0033] 如稍后进一步描述的,智能体控制和合约协商系统100a确定其他智能体的预测动作,特别是使用状态表示来确定。在实施方式中,动作选择子系统110a或动作选择子系统110a的(例如,具有不同的可学习参数值,例如来自较早的训练迭代的可学习参数值)的版本用于预测其他智能体102b...n的动作,例如,预测由其他智能体的动作选择子系统110b...n选择的动作。

[0034] 可以以各种方式确定其他智能体的预测动作,并且其实现的方式可以取决于特定的技术应用。例如,在一种方法中,智能体具有相同或对应的任务目标,诸如导航到例如环境中的对象的目标位置或实现环境的特定状态或配置。然后,从环境的观察 O_a 得出并由动作选择子系统处理以预测特定智能体的动作的状态表示可以表示与该特定智能体相关的环境。例如,如果任务是导航到目标位置,则状态表示可以包括该特定智能体的定位表示。在其他情况下,例如当环境是计算或通信环境时,状态表示可以包括与该智能体相关的计算或通信环境的状态或配置的表示。在一些其他实施方式中,由动作选择子系统处理以预测特定智能体的动作的状态表示可以与例如来自特定智能体的观察 O_a 分开获得。

[0035] 作为另一个示例,动作选择子系统110a可以是以目标为条件的,例如,它还可以处理标识要实现的任务的输入。在这样的情况下,第一智能体可以获得例如来自另一个智能体定义该另一个智能体的目标的信息,并处理该信息和来自观察 O_a 的数据以预测该另一个智能体的动作。作为进一步的示例,智能体控制和合约协商系统100a可以维护一个或多个其他智能体的动作选择子系统的副本,并使用它来预测由其他智能体采取的动作。

[0036] 在实施方式中,智能体控制和合约协商系统100a还包括价值神经网络120a,其被配置为根据价值神经网络参数处理状态表示以确定状态值。状态值可以定义第一智能体来自状态表示表征的环境状态的期望回报。在一些实施方式中,价值神经网络120a确定每个智能体的状态值,例如通过针对每个智能体处理状态表示并向价值神经网络120a提供分开的“头”;或者以其他方式,例如通过对智能体中的每一个使用相同的价值神经网络,或者通过获得其他智能体中的每一个的价值神经网络的副本。

[0037] 通常,来自环境的状态的期望回报可以包括从该状态开始执行任务的奖励的累积的、经时间折扣的总和的估计。奖励可以包括负奖励,即成本。在实施方式中,每个智能体都有一个价值神经网络。通常,本文所述技术的实施方式可以尝试最大化第一智能体102a的期望回报和/或可以尝试最小化第一智能体102a的对应成本。然而,本文所述技术促进智能体当中的合作,从而可以为包括第一智能体102a的所有智能体实现改进的回报。

[0038] 在一些实施方式中,每个智能体的动作选择子系统110a—例如每个智能体的动作选择策略神经网络和价值神经网络—都已经被预训练以执行任务,并且所描述的技术使用这些预训练的神经网络来协商合约。有许多方法可用于这样的预训练,并且本文描述的技术不依赖于任何特定方法。仅作为一些示例,可以使用模仿学习中的一种或多种,例如行为克隆、遗憾最小化和强化学习(基于奖励)。

[0039] 作为一个特定示例,可以使用模仿学习,然后进行强化学习,以训练和改进每个智能体的动作选择子系统110a和价值神经网络120a。这样的方法在Anthony等人的arXiv:

2006.04635v4(同上所述)中进行了描述。广义上,这可以涉及,对于多个训练迭代中的每个训练迭代:通过控制具有改进策略的智能体来生成训练迭代的训练数据,该改进策略通过使用(i)自一个或多个先前迭代起从策略神经网络生成的候选策略和(ii)候选价值神经网络执行最佳响应计算来选择响应于输入状态表示的动作。可以从一个或多个先前迭代中的每个先前迭代起从价值神经网络生成候选价值神经网络。可以在多个训练迭代中的每个训练迭代处通过对训练数据进行训练来更新策略神经网络价值神经网络。

[0040] 在所述技术的一些实施方式中,动作选择子系统—例如动作选择策略神经网络的动作选择子系统—和价值神经网络的可学习参数可以被冻结。在一些其他实施方式中,动作选择子系统—例如动作选择策略神经网络的动作选择子系统—和价值神经网络的可学习参数可以在本文所述过程的操作期间被训练,例如被微调。

[0041] 通常,动作选择子系统110a—例如动作选择策略神经网络—和价值神经网络120a可以具有任何合适的架构。例如,动作选择策略神经网络和价值神经网络120a可以包括例如一个或多个前馈神经网络层、一个或多个卷积神经网络层、一个或多个注意力神经网络层或一个或多个规范化层。可以通过—例如使用诸如Adam的优化算法—反向传播目标函数的梯度以更新诸如权重的神经网络参数的值来预先训练神经网络。作为示例,强化学习目标函数可以基于接收到的奖励,例如基于贝尔曼误差或基于策略优化目标。

[0042] 智能体控制和合约协商系统100a进一步包括用于智能体间通信的通信子系统130a。在实施方式中,每个智能体都具有这样的系统,并且这使得智能体,更具体地说是智能体的智能体控制和合约协商系统,能够相互通信以协商动作的联合计划。这通常涉及根据协议与一个或多个其他智能体协商协定(agreement)或合约,如稍后更详细地描述的。通常,合约定义了每个智能体将来可能采取的动作的限制。例如,这样的合约可以定义对可以由第一智能体选择的动作的限制以及可以由另一个第二智能体选择的动作的限制。通信子系统130a可以实现任何形式的通信,例如有线或无线通信、点对点或点对多点;它可以实现低级数据通信,或者它可以使用自然语言。

[0043] 智能体控制和合约协商系统100a的实施方式还包括模拟器140a,其被配置为模拟智能体——特别是第一智能体102a和其他智能体——未来动作的影响。模拟器140a处理表征当前时间步的环境的状态的状态表示,以及智能体的假定(supposed)动作,例如第一智能体的候选动作和其他智能体中的每个智能体的预测动作。模拟器140a确定环境的预测的后续状态,例如环境的下一个状态,特别是表征预测的环境的下一个状态的状态表示。因此,模拟器140a可以例如通过模拟在i)没有提议合约的情况下以及ii)考虑提议合约中对动作的限制的影响的情况下第一智能体和其他智能体的未来动作的影响,来预测合约的影响。

[0044] 图2是控制例如智能体102a的第一智能体在环境中执行任务的示例过程的流程图,在该环境中,第一智能体与执行一个或多个其他任务的一个或多个其他智能体交互。图2的过程可以在一个或多个位置的一个或多个计算机上实现,例如在图1的智能体控制和合约协商系统100a上实现,并且可选地在每个其他智能体控制和合约协商系统上实现。

[0045] 在多个时间步中的每个时间步处,第一智能体102a例如从对环境的观察获得表征当前时间步处的环境的状态的状态表示,并使用动作选择子系统110a对此进行处理,以供智能体生成针对智能体的策略输出(步骤200)。状态表示还用于预测其他智能体的动作,并

且策略输出和这些预测动作用于选择第一智能体要执行的动作。

[0046] 通常,取决于其他智能体的预测动作来选择第一智能体的动作。例如,可以从候选动作当中通过在给定其他智能体的预测动作的情况下针对每个候选动作确定环境的下一个状态的状态值来选择动作。可以基于来自环境的下一个状态的期望回报来选择动作。其他智能体中的每个智能体都可以类似地以该方式选择动作。

[0047] 可以从环境的模拟中确定环境的下一个状态,例如使用模拟器140a来处理候选动作和预测动作来进行确定。可以通过使用价值神经网络120a来处理表征环境的状态的状态表示来确定环境的状态的期望值,例如来自环境的下一个状态的期望回报。

[0048] 在实施例中,该过程包括,在时间步中的一个或多个时间步中,第一智能体与第二智能体协商合约,其中第二智能体是其他智能体之一(步骤202)。合约定义了对可以由第一智能体选择的动作的限制以及对可以由第二智能体选择的动作的限制。

[0049] 在实施例中,协商合约包括:第一智能体例如使用稍后描述的SVE值估计,通过确定一个或多个得分的值来确定是否提议合约(步骤204),所述得分表示在具有和不具有提议合约的情况下,在后续时间步处对第一智能体的环境的状态的期望值。这可以通过在所有智能体的实施方式中,模拟在i)没有提议合约的情况下以及ii)考虑提议合约中对动作的限制的影响的情况下第一智能体和其他智能体的(潜在)未来动作的影响来执行。可以使用动作选择子系统或其例如具有不同可学习参数值的版本来确定潜在的未来动作,并且可以使用价值神经网络120a评估未来动作的影响,以确定来自环境的下一个状态的期望回报,例如其由模拟器140a来确定。在实施方式中,该方法使用得分来确定是否提议合约,在一些实施方式中,确定是否提议任何合约,在一些实施方式中,确定是否提议特定合约。

[0050] 然后,可以取决于得分,例如当一个或多个得分指示有提议合约的情况下的期望回报大于没有提议合约的情况下的期望回报时,将提议合约传送给第二智能体(步骤206)。

[0051] 提议合约可以被第二智能体接受或可以不被第二智能体接受。在没有接受的情况下可以假设不接受,或可以明确传送不接受(步骤208)。

[0052] 当合约被第二智能体接受—例如由从第二智能体到第一智能体的通信所指示—时,该过程在考虑合约中对第一智能体的动作的限制的情况下选择供第一智能体执行的动作(步骤210)。对应地,第二智能体可以在考虑对可以由第二智能体选择的动作的限制的情况下选择动作。第二智能体可以通过确定在有提议合约的情况下和没有提议合约的情况下的(针对第二智能体的)期望回报来确定是否接受提议合约,例如通过如针对第一智能体描述的模拟受限制和不受限制的动作的影响来进行。当提议合约不被接受时,第一智能体可以基于策略输出来以任何合适的方式选择动作,而不受与第二智能体的提议合约的约束。尽管如此,这些动作可能受到与其他智能体的一个或多个合约(如果存在的话)的约束。

[0053] 该过程根据任何合适的标准通过获得和处理表征下一时间步处的环境的状态表示(步骤200)继续,直到完成任务(步骤212)。

[0054] 当存在多于两个智能体时,图2的过程可以由第一智能体用来向其他智能体中的每个智能体提议合约(尽管系统中的所有其他智能体不必须都能够达成合约)。在实施方式中,如果存在多个达成的合约动作,则由第一智能体达成的所有合约的限制可以被应用于第一智能体的动作。

[0055] 如稍后更详细描述,存在可用于协商,即通信和达成,合约的各种协议。例如,

“相互提议 (Mutual Proposal)”协议对第一智能体和对第二智能体两者的动作施加限制,“提议-选取 (Propose-Choose)”协议使第一智能体和第二智能体中的每一个能够达成采取具体的相应动作。特定的所要求的或不允许的动作可以取决于应用。例如,在“相互提议”协议中,可能存在例如对可能导致环境中发生碰撞或冲突、或部分碰撞或冲突或存在这样的风险的动作的限制;或者对会支持第三智能体的动作的限制。在“提议-选取”协议中,在存在多于一个其他智能体的情况下,则可以选择单个“第二智能体”作为合作伙伴来达成合约,其中一者例如根据得分值提供最有利合约或最大期望回报的智能体。

[0056] 在该过程的一些实施方式中,第一智能体可以从智能体中的第二智能体或其他智能体接收一个或多个提议合约。然后,协商合约可以涉及第一智能体和第二智能体两者在将合约用于限制第一智能体 (和第二智能体) 的动作之前接受提议合约。例如,在“提议-选取”协议的一些实施方式中,第一智能体和第二智能体两者都需要选取 (接受) 相同的合约以使得其被实施。在一些实施方式中,第一智能体可以接受 (可能能够指示它们愿意接受) 两个合约,即由第一智能体提议的提议合约和来自第二智能体的提议合约。然后可以实施任一合约来限制动作,例如具有最大期望回报的合约。

[0057] 仅作为示例,当一动作不受约束时,该动作可以直接从动作选择策略神经网络的输出中确定,该动作也可以使用“采样的最佳响应” (SBR) 算法 (同上所述) 确定。例如,将智能体 i 的当前动作选择策略 (例如由其动作选择策略神经网络定义) 表示为 π_i^c , 并且将其他智能体的动作选择策略表示为 π^b (例如由相同或不同的动作选择策略神经网络定义), 将从 π_{-i}^b 中采样的其他智能体动作的向量表示为 $\mathbf{a}_{-i} \sim \pi_{-i}^b$ (其中下标 $-i$ 表示除智能体 i 之外的所有智能体), 并将 M 个这样的动作向量的样本表示为 $\mathbf{A} = \mathbf{a}_{-i}^1, \dots, \mathbf{a}_{-i}^M \sim \pi_{-i}^b$, 可以针对智能体 i 的候选动作 c^j 和环境的状态 $\mathbf{s}$ 将估计的未来奖励, 即动作值估计 $Q_i^{A-i}(\mathbf{s}, c^j)$ 定义为:

$$Q_i^{A-i}(\mathbf{s}, c^j) = \frac{1}{M} \sum_{m=1}^M Q_i^{a_m}(\mathbf{s}, c^j) = \frac{1}{M} \sum_{m=1}^M V_i(T(\mathbf{s}, (c^j, \mathbf{a}_{-i}^m)))$$

其中, $T(\mathbf{s}, \mathbf{a})$ 是给定由模拟器140表示的 (所有) 智能体动作 $\mathbf{a}$ 的向量时环境状态的转移函数; 并且 $V_i(\mathbf{s})$ 是状态值, 例如来自价值神经网络120的状态值。可以为从 $\pi_i^c(\mathbf{s})$ 中采样的多个一例如许多一候选动作确定动作值估计 $Q_i^{A-i}(\mathbf{s}, c^j)$, 并且可以选择这些中的一个一例如具有最高值估计的候选动作一作为要执行的动作。

[0058] 如上所述, 确定是否提议合约能够涉及确定一个或多个得分的值, 所述一个或多个得分表示在当前时间步的后续时间步处对第一智能体的环境的状态的期望值。

[0059] 在实施方式中, 确定后续时间步处对第一智能体的环境状态的期望值涉及使用第一智能体的动作选择子系统110a对该时间步处第一智能体的一个或多个可能动作进行采样。对于一个或多个动作中的每个动作以及对于其他智能体中的每个智能体, 该过程然后可以对该时间步处其他智能体的多个可能动作进行采样, 并且使用价值神经网络来确定由智能体的采样动作确定的 (在后续时间步处) 环境的预测的新状态的状态值。

[0060] 如前所述, 可以使用第一智能体的动作选择子系统或其一版本来确定其他智能体的多个可能动作, 例如通过使用动作选择子系统处理当前状态表示。通过使用模拟器140a模拟可能动作的影响可以确定环境的预测的新状态, 例如以确定环境的预测的新状态的状

态表示。模拟未来动作的影响可以包括：在针对其他智能体中的每个智能体的采样动作上对(由可能的动作确定的)每个预测的新状态的状态值进行平均,以确定对第一智能体的环境状态的期望值,特别是针对第一智能体的一个或多个采样的可能动作中的每个动作。

[0061] 模拟(所有)智能体的未来动作可以包括,针对每个智能体,使用动作选择子系统 and/或在提议合约应用于智能体的情况下通过应用在提议合约中的对动作的限制,选择该智能体的一个或多个候选动作的集合。在一些实施方式中,提议合约定义了动作的受限集合;在一些实施方式中,提议合约定义了要由诸如第一智能体或第二智能体的智能体执行的单个动作。

[0062] 在一些实施方式中,通过对 M 个部分动作简档 $\mathbf{A}_{-i} = \mathbf{a}_{-i}^1, \dots, \mathbf{a}_{-i}^M \sim \pi_{-i}$ 进行采样,计算出智能体 i 采取动作 $\mathbf{a}_i$ 的值的蒙特卡罗估计,其中 $\mathbf{a}_{-i} \sim \pi_{-i}$ 表示针对除智能体 i 以外的所有智能体从(以某种形式对智能体 i 可用的)其相应动作选择子系统的相应动作选择策略 π_{-i} 中采样动作,使得完整的动作简档为 $(\mathbf{a}, \mathbf{a}_{-i})$ 。然后可以如下所示计算智能体 i 采取动作 $\mathbf{a}_i$ 的值的估计 $\hat{Q}_i^{\mathbf{A}_{-i}}(s, \mathbf{a}_i)$,其稍后被称为模拟动作值估计(SAVE):

$$\hat{Q}_i^{\mathbf{A}_{-i}}(s, \mathbf{a}_i) = \frac{1}{M} \sum_{m=1}^M Q_i^{\mathbf{a}_{-i}^m}(s, \mathbf{a}_i)$$

其中 $Q_i^{\mathbf{a}_{-i}}(s, \mathbf{a}_i) = V_i(T(s, (\mathbf{a}, \mathbf{a}_{-i})))$ 。类似地,可以使用 $V_j(T(s, (\mathbf{a}, \mathbf{a}_{-i})))$ 而不是 $V_i(T(s, (\mathbf{a}, \mathbf{a}_{-i})))$ 为智能体 j 计算SAVE估计(以及稍后描述的STAVE估计),其中 $V_j(\cdot)$ 可以使用智能体 i 的价值神经网络(假设这是一个足够有用的近似)或其例如具有诸如权重的不同的可学习参数的版本,或者通过有权访问智能体 j 的价值神经网络或其副本来确定。

[0063] 基于可能动作的受限集合,可以通过对从受限策略中采样的智能体的动作进行平均来确定例如智能体 i 的模拟值估计(SVE)。作为示例,对于表示为 $\mathbf{D} = (\mathbf{R}_i, \mathbf{R}_j)$ 的合约,其中,智能体 i 的动作选择子系统的动作选择策略 π_i 被限制于动作集合 $\mathbf{R}_i$,而智能体 j 的动作选择策略 π_j 被限制于动作集合 $\mathbf{R}_j$,则可以对从 $\mathbf{D}$ 中采样的动作进行平均。该值估计可以写成 $\hat{V}_i^{\mathbf{A}}(s) = \frac{1}{M} \sum_{m=1}^M Q_i(s, \mathbf{a}^m)$,其中每个 $\mathbf{a}^m$ 是完整的动作简档,即包括所有智能体的动作, $\mathbf{a}^m \sim \pi^{\mathbf{D}}$,其中 $\sim \pi^{\mathbf{D}}$ 表示从智能体 i 和智能体 j 的动作的受限集合中采样并从其他智能体的不受限制的策略中采样,并且 $\mathbf{A} = \mathbf{a}^1, \dots, \mathbf{a}^M \sim \pi^{\mathbf{D}}$,取 M 个蒙特卡罗样本。可以通过从不受限制的策略中采样,当采样的动作超出限制 $\mathbf{R}_i$ (或 $\mathbf{R}_j$)之外时重新采样来对动作的受限集合进行采样。当第一智能体和第二智能体的动作受到提议合约的限制时,SVE值估计可用于确定对第一智能体的环境的状态的期望值。

[0064] 方便但并非必不可少的是,给定捕获策略 π_i 和受限动作集合 $\mathbf{R}_i$ 的神经网络,可以通过掩蔽不在 $\mathbf{R}_i$ 中的所有动作的对数几率(logit)一例如通过在最后的softmax层之前将不在 $\mathbf{R}_i$ 中的所有动作的权重设置为零—从受限策略 $\pi_i^{\mathbf{R}_i}(s)$ (也可以表示为 $\pi^{\mathbf{D}}$)中采样动作。

[0065] 还可以通过以下操作来确定考虑智能体 i 的动作 $\mathbf{a}_i$ 和智能体 j 的动作 $\mathbf{a}_j$ 的组合影响的模拟双动作值估计(STAVE):对从其他智能体的相应策略中采样的所述其他智能体的动作进行平均。这可以通过对除智能体 i 和智能体 j 之外的所有智能体取动作 $\mathbf{a}_{-\{i,j\}}^m$ 的 M 个

样本 $\mathbf{A}_{\{i,j\}} = \mathbf{a}_{\{i,j\}}^1, \dots, \mathbf{a}_{\{i,j\}}^M$ 并且使用完整的动作简档 $(a_i, a_j, \mathbf{a}_{\{i,j\}}^m)$ 来估计下一个状态的值 $\hat{Q}_i^A(s, (a_i, a_j)) = \frac{1}{M} \sum_{m=1}^M Q_i^{a_{\{i,j\}}^m}(s, (a_i, a_j))$ 来完成。

[0066] 图3是使用“相互提议”协议的过程的实施方式中用于确定是否提议合约的示例过程的流程图。图3的过程可以在一个或多个位置的一个或多个计算机上实现,例如在图1的智能体控制和合约协商系统100a上实现,并且可选地在智能体控制和合约协商系统中的每一个上实现。

[0067] 在过程的一些实施方式中,合约是定义动作的类的“通用”合约,其定义对可以由第一智能体和第二智能体两者选择的动作的相同或对应的限制。作为示例,对动作的对应限制可以定义一个智能体——例如第一智能体——被禁止移动到由另一个智能体——例如第二智能体——占据的空间。也就是说,限制可能对应,但受限制的具体动作可能不同。这样的合约可以表示为 $D_{i,j} = (R_i, R_j)$ 。

[0068] 示例过程涉及确定第一得分,该第一得分表示在具有提议合约的情况下在后续时间步处的对第一智能体的环境状态的期望值(步骤300);以及确定第二得分,该第二得分表示在没有提议合约的情况下在后续时间步处的对第一智能体的环境状态的期望值(步骤302)。然后可以比较第一得分和第二得分,并使用比较结果来确定是否提议合约,例如,当第一得分超过第二得分时传送提议合约(步骤304)。广义上讲,当第一智能体预期在合约达成一致时实现更大效用,第一智能体提议合约,否则避免提议合约。作为示例,第一得分和第二得分中的每一个可以包括如上所述的模拟值估计(SVE)。

[0069] 第一智能体可以向其他智能体中的每个智能体提议这样的合约,例如使用图3的过程来进行,并且可以达成多个通用合约,在这样的情况下,应用所有达成合约的动作限制。

[0070] 在一些实施方式中,第一智能体和第二智能体只有在它们中的每一个向对方提议定义动作的受限集合的合约时,即在双方协定限制其动作时,才会接受这样的“相互提议”合约。如果没有达成协定,则每个智能体都可以自由选择不受提议合约限制的动作(尽管可能受到与其他智能体达成的合约的限制)。如果达成协定,由第一智能体和第二智能体根据达成的限制选择动作。

[0071] 图4是用于在使用“提议-选取”协议的过程的实施方式中确定是否提议合约的示例过程的流程图。图4的过程可以在一个或多个位置的一个或多个计算机上实现,例如在图1的智能体控制和合约协商系统100a上实现,并且可选地在智能体控制和合约协商系统中的每一个上实现。

[0072] 在一些实施方式中,合约定义了必须由第一智能体和第二智能体采取的具体动作。例如,合约可以定义必须由第一智能体选择的第一动作和必须由第二智能体选择的第二动作,即每个智能体在合约下被限制于单个动作。然后可以使用“提议-选取”协议来确定在提议阶段要提议哪些合约(动作)。智能体可以从涉及它们的所有合约(动作)中进行选择,无论是它们提议的合约还是向它们提议的合约。涉及例如第一智能体的所有合约的组可以称为第一智能体的供考虑(On The Table)的合约。

[0073] 在“提议-选取”协议的一些实施方式中,两个智能体只有在选择相同合约的情况下才能达成协定,即接受合约,例如,如果两者都选择 $D^{i \rightarrow j}$ 或两者都选择 $D^{j \rightarrow i}$, 其中 $D^{i \rightarrow j}$ 表

示智能体 i 向智能体 j 提议的合约,反之亦然。在一些实施方式中,智能体可以指示它们愿意接受 $D^{i \rightarrow j}$ 和 $D^{j \rightarrow i}$ 中的任一个(如果智能体以不同的次序对它们进行排序,则可以随机选择一个)。对于 n 个智能体,每个智能体可以潜在地提议并接收 $n-1$ 个合约,因此希望与第二智能体达成协定的第一智能体正在与其他智能体竞争。如果第一智能体提议对其本身最有利的合约,则第二智能体不太可能选择该合约,并且所述协议的实施方式会考虑其他智能体可能接受哪些合约。

[0074] 在实施方式中,用于确定是否提议合约的一个或多个得分包括协定得分,该协定得分表示当第一智能体和第二智能体分别选择第一动作和第二动作(如果有其他智能体,则其他智能体的动作可以不受限制)时,后续时间步处的对第一智能体和第二智能体两者的环境状态的期望值。

[0075] 参考图4,在步骤400处,该过程涉及确定相应的第一基线值和第二基线值,所述第一基线值和第二基线值分别表示在不将第二智能体限制于第二动作的情况下和不将第一智能体限制于第一动作的情况下,在后续时间步处对第一智能体和对第二智能体的环境状态的期望值。可以使用价值神经网络确定后续时间步处的对第一智能体和对第二智能体的环境的状态的期望值,例如它可以包括来自价值神经网络的状态值或状态值的平均。

[0076] 然后,该过程使用(即相对于)基线值来确定协定得分(步骤402),并确定是否基于协定得分来提议合约(步骤404)。

[0077] 在实施方式中,可以评估多个候选合约,并提议具有最高协定得分的候选合约。也就是说,该过程可以包括通过确定多个候选提议合约并基于候选的协定得分来选择候选提议合约,例如通过选择具有最高协定得分的候选提议合约作为提议合约,来确定针对第二智能体的提议合约。

[0078] 在一些实施方式中,协定得分包括以下项的乘积:i) 当第一智能体和第二智能体分别选择第一动作和第二动作时,在后续时间步处的对第一智能体的环境状态的期望值(d_1)与第一基线值(d_1^0)之间的差;以及ii) 当第一智能体和第二智能体分别选择第一动作和第二动作时,在后续时间步处的对第二智能体的环境状态的期望值(d_2)与第二基线值(d_2^0)之间的差。例如,协定得分可以确定为 $(d_1 - d_1^0)^+(d_2 - d_2^0)^+$,其中 $(\cdot)^+$ 表示 $\max(\cdot, 0)$ 。在实施方式中,协定得分等于或大于零。

[0079] 在一些实施方式中,(最高)协定得分近似于所谓的“纳什博弈(Nash Bargaining)得分”或“纳什博弈解”(例如,在J. F. Nash, “The bargaining problem,” in The Essential John Nash, pp. 37-48, Princeton University Press, 2016(J. F. Nash, “博弈问题”,在约翰·纳什精要中,第 37-48 页,普林斯顿大学出版社,2016 年)中描述)。以该方式确定协定得分有助于识别相对于没有合约的情况对第一智能体和对第二智能体互相有利的合约。结合下面参考图5描述的选择合约的方法,它还提供了对帕累托最优的近似(其中,如在两个智能体之间,两个智能体在如下情况下均不会选择一合约:两个智能体能够找到另一个对它们具有更高效用的合约)和对称性(对两个智能体的期望效用)。

[0080] 在一些实施方式中,第一基线值包括当第一智能体选择第一(要求)动作并且由第二智能体选择的动作不受限制时,在后续时间步处的对第一智能体的环境的状态的期望值。第二基线值可以包括当第二智能体选择第二(要求)动作并且由第一智能体选择的动作

不受限制时,在后续时间步处的对第二智能体的环境状态的期望值。

[0081] 在一些实施方式中,确定第一基线值和第二基线值考虑了除了第一智能体和第二智能体之外的其他智能体对之间是否可以达成合约。这可以通过模拟其他智能体的协商过程来完成。因此,对于除了包括第一智能体和第二智能体的对之外的每个智能体对,确定第一基线值可以包括针对后续时间步识别它们是否会达成合约。这可以按照先前针对第一智能体和第二智能体所述的方式完成。当它们达成合约时,第一基线值可以被确定为就如同该对的达成的合约已经确立(in place)一样在后续时间步处的对第一智能体的环境的状态的期望值。类似地,确定第二基线值可以包括,对于除了包括第一智能体和第二智能体的对之外的每个智能体对,识别它们是否会达成后续时间步的合约。当它们会达成合约时,第二基线值可以被确定为在如同达成的合约已经确立一样的情况下在后续时间步处的对第二智能体的环境状态的期望值。如前所述,可以通过使用价值神经网络120处理例如由模拟器确定的后续时间步的环境的状态来确定环境的状态的期望值。通常,可以通过在例如由动作选择子系统确定的智能体的可能动作上进行蒙特卡罗采样来确定期望值,以获得平均。

[0082] 图5是用于选择合约来评估以供如上所述例如在“提议-选取”协议中提议的示例过程的流程图。图5的过程可以在一个或多个位置的一个或多个计算机上实现,例如在图1的智能体控制和合约协商系统100a上实现,并且可选地在智能体控制和合约协商系统中的每一个上实现。

[0083] 图5的过程的目的是能够提议近似于纳什博弈解的合约。一个困难是多个智能体102a...n的动作空间是组合的,并且可能非常大,例如大于 10^{20} ,并且可能合约的空间随着它组合第一智能体和第二智能体的动作而随其平方地增长。解决这个问题的一种方法是从第一智能体和第二智能体的动作选择策略中采样许多动作。这样的方法可以涉及通过确定多个候选提议合约来确定针对第二智能体的提议合约,然后选择其中一个候选提议合约作为提议合约。

[0084] 确定候选提议合约可以包括:通过使用动作选择子系统处理当前时间步处的状态表示以生成针对第一智能体的策略输出来确定第一智能体的第一候选动作集合(步骤500)。确定候选提议合约也可能包括(第一智能体)确定第二智能体的第二候选动作集合,例如通过使用动作选择子系统或其一版本处理当前时间步处的状态表示以生成针对第二智能体的策略输出(步骤502)。然后,该过程可以生成动作对,该动作对包括来自第一候选动作集合的第一候选动作和来自第二候选动作集合的第二候选动作(步骤504)。

[0085] 然后,可以通过经由以下操作来选择候选动作对的集合来确定候选提议合约:根据对度量对动作对进行排名并且选择 K 个最高排名的动作对(步骤506)。每个候选提议合约可以包括候选动作对之一。

[0086] 动作对的度量可以包括以下项的组合,例如和:在第一智能体执行动作对中的第一候选动作时在后续时间步处的对第一智能体的环境状态的期望值,以及在第二智能体执行动作对中的第二候选动作时在后续时间步处的对第二智能体的环境状态的期望值。该组合—例如求和—可以是用于对第二智能体的提议合约的值进行加权的加权组合,例如以调整接受提议合约的可能性。对第一智能体和对第二智能体的环境状态的期望值可以缩放以大致匹配彼此。

[0087] 然后可以针对候选提议合约中的每一个计算协定得分(步骤508),然后可以基于针对候选提议合约计算的协定得分—例如通过选择具有最高协定得分的合约—选择候选提议合约之一作为提议合约(步骤510)。

[0088] 可选地,该过程还可以包括:第一智能体确定自己是否接受第一智能体向第二智能体提议的合约;或者是否接受即将到来的由第二智能体或其他智能体之一提议的合约(步骤512)。例如,这可以涉及如前所述第一智能体确定一个或多个得分,例如提议或即将到来的合约的协定得分,并基于确定的得分来确定是否接受提议或即将到来的合约。在一些实施例中,第一智能体从它们的供考虑合约中进行选取,选择单个其他智能体来与其达成合约。

[0089] 作为特定示例,该过程可以通过从策略 π_i 中采样智能体 i 的 N 个候选动作 $c_i^1, \dots, c_i^N$ 的集合 C_i 和从策略 π_j 中采样智能体 j 的 N 个候选动作 $c_j^1, \dots, c_j^N$ 的集合 C_j , 确定笛卡尔积 $C_i \times C_j$, 并如上所述选择按对度量排名的前 K 个,从而生成多个候选提议合约。

[0090] 可以假设其他智能体(如果有)不受限制,并根据不受限制的策略简档 $\pi_{-\{i,j\}}^b$ 选择动作,来确定对度量的值,从中可以得出 M 个动作简档 $B = (b^1, \dots, b^M)$; $b^1, \dots, b^M \sim \pi^b$ 。动作 c_i 的对度量可以包括对智能体 i 的环境的下一状态的值 $\hat{Q}_i^{B-i}(s, c_i)$ 和对智能体 j 的环境的下一状态的值 $\hat{Q}_j^{B-i}(s, c_i)$ 的加权和。 $\hat{Q}_j^{B-i}(s, c_i)$ 的值可以缩放到与 $\hat{Q}_i^{B-i}(s, c_i)$ 类似的比例,从而便于在具有高估计值对比低估计值的智能体之间协商合约。

[0091] 例如,缩放因子可以被确定为 $q_{i,j}^0/q_{j,i}^0$, 其中 $q_{i,j}^0 = \hat{Q}_i^{b-\{i,j\}}(s, (c_i, c_j))$ 并且 $q_{j,i}^0 = \hat{Q}_j^{b-\{i,j\}}(s, (c_i, c_j))$ 。这可以通过以下来计算:针对所有智能体($\bar{b}$)采样 M 个动作简档的集合,然后使用SAVE针对智能体 i 和智能体 j 选择最佳动作,并使用STAVE将这些组合,以在未达成合约时估计智能体 i 的值, $q_{i,j}^0 = \hat{Q}_i^{b-\{i,j\}}(s, (c_i, c_j))$; 并且对于智能体 j 也是类似。因此,在一个示例实施方式中,对度量可以确定为 $\hat{Q}_i^{B-i}(s, c_i) + \beta \frac{q_{i,j}^0}{q_{j,i}^0} \hat{Q}_j^{B-i}(s, c_i)$, 其中,权重 β 可以在 $[0, 1]$ 范围内,并且 β 的低值强调了对提议者,智能体 i , 的利益,反之亦然。

[0092] 智能体 i 可以向智能体 j 提供的多个候选提议合约表示为 $D_{i,j}$ 。该过程可以使用 $D_{i,j}$ 中动作 c_i 和 c_j 的协定得分来确定提议 $D_{i,j}$ 中的哪个合约,这在实施方式中基于根据其来衡量达成合约的增益的“无交易”第一基线值和第二基线值。

[0093] 在一种方法中,假设其他智能体在它们本身当中不形成协定而并且从它们相应的策略 π^b 中选取动作,使得(继续前面的记号) $d_1^0 = \hat{Q}_i^{B-\{i,j\}}(s, (c_i, c_j))$ 和 $d_2^0 = \hat{Q}_j^{B-\{i,j\}}(s, (c_i, c_j))$ 。这些值可以使用上面描述的STAVE算法来估计。

[0094] 在另一种方法中,基线值的估计是通过考虑其他智能体可能能够通过它们本身当中形成协定来做得更好来确定的。因此,该过程可以为每个智能体考虑可能与其他智能体达成哪些合约,并且然后使用其中最佳的期望状态值作为基线。下面给出了一种用于此的算法“BATNA”(协商协定的最佳替代方案):

```

function MBDS-BARGAIN(s: 状态, B: 动作简档,  $(\mathcal{D}_{i,j})_{i,j=1}^n$ : 候选交易,
 $(q_{i,j}^0)_{i,j=1}^n$ : 无交易BATNA值估计,
 $(q_{i,j})_{i,j=1}^n$ : BATNA值估计,  $\kappa$ : 阻尼因子)
  for  $i \leftarrow 1$  to  $n$  do
    for  $j \leftarrow 1$  to  $n$ ,  $j \neq i$  do ▷ 找到与j的纳什博弈交易
       $(\{c_i\}, \{c_j\}) \leftarrow \operatorname{argmax}_{(\{c_i\}, \{c_j\}) \in \mathcal{D}_{i,j}} \text{NASHSCORE}(s, i, j, c_i, c_j, q_{i,j}, q_{j,i}, \mathbf{B}_{-\{i,j\}})$ 
       $r_j \leftarrow \hat{Q}_i^{\mathbf{B}_{-\{i,j\}}}(s, (c_i, c_j))$ 
      if  $\text{NASHSCORE}(s, i, j, c_i, c_j, q_{i,j}, q_{j,i}, \mathbf{B}_{-\{i,j\}}) = 0$  then  $r_j \leftarrow -\infty$ 
      for  $j \leftarrow 1$  to  $n$ ,  $j \neq i$  do
         $q'_{i,j} \leftarrow \max(q_{i,j}^0, \max_{k \in \{1, \dots, n\} \setminus \{i,j\}} r_k)$  ▷ 找到BATNA
         $q'_{i,j} \leftarrow \kappa q_{i,j} + (1 - \kappa) q'_{i,j}$  ▷ 应用阻尼
  return  $(q'_{i,i})_{i=1}^n$ 

```

[0095] 这里 $q_{i,j}^0 = d_1^0$ (和 $q_{j,i}^0 = d_2^0$) ;

$\text{NashScore}(s, i, j, c_i, c_j, q_{i,j}, q_{j,i}, \mathbf{B}_{-\{i,j\}}) = \max(0, \hat{Q}_i^{\mathbf{B}_{-\{i,j\}}}(s, (c_i, c_j)) - q_{i,j}) \cdot \max(0, \hat{Q}_j^{\mathbf{B}_{-\{i,j\}}}(s, (c_i, c_j)) - q_{j,i})$; 并且返回的基线值为 $(q'_{i,j})_{i,j=1}^n$ 。算法逐步遍历智能体 i 、智能体 j 的对, 以使用无交易估计 (“找到纳什博弈交易”) 来找到使协定得分最大化的动作对, 然后对于每个对, 如果更高的协定得分改进了无交易估计 (“找到BATNA”), 则将该更高的协定得分作为基线。可以将例如指数平滑的阻尼因子应用于返回的协定值。

[0096] 每个智能体可以通过为每个智能体计算协定得分, 例如使用先前计算的基线值, 选择单个其他智能体来与其达成协定, 来从它们的供考虑合约中进行选取。如果一些协定得分为正, 则可以选择具有最高得分的合约; 否则, 智能体可以根据在下一个时间步处的对智能体的环境状态的期望值来选择对本身最有利的合约, 前提是该值大于无交易基线 (如上所述计算, 例如使用BATNA)。在一些实施方式中, 该过程要求以该方式选择的合约由第一智能体和第二智能体两者选择, 即两者都选取 $D^{i \rightarrow j}$ 或两者都选取 $D^{j \rightarrow i}$ (在每一个中使用动作 c_i 、 c_j 的相应对被评估), 以供达成合约。

[0097] 在一些实际情况中, 例如当并非所有智能体都属于同一类型或涉及人类时, 达成的合约中的限制可能被打破。为智能体提供制裁另一个 “偏离” 智能体行为的能力以鼓励一致性是有用的。例如, 偏离智能体当它确定它预期从破坏合约中获益时, 可能会破坏合约。通常, 制裁智能体可以修改其动作, 以试图降低偏离智能体的奖励。

[0098] 在一些实施方式中, 第一智能体可以例如从环境状态的观察 (其包括智能体) 中检测第二智能体何时已经违反合约。违约可能发生在标志为违约时间步的时间步。响应于检测到违约, 第一智能体可以制裁第二智能体。制裁第二智能体可以包括第一智能体在违约时间步之后的一个或多个时间步中修改与第二智能体的合约协商。例如, 第一智能体可以在违约时间步之后停止与第二智能体的合约协商, 或者可以修改在后续时间步处对第一智能体的环境状态的期望值的确定, 例如在后续时间步处向第二智能体添加表示环境状态的期望值的负值的项 (以尝试降低偏离智能体的奖励)。

[0099] 作为示例, 用于利用权重 α 制裁智能体 j 的下一个状态的修改的值估计可以被确定为 $\hat{Q}_{i \rightarrow j}^{\mathbf{B}_{-i}}(s, a) = \hat{Q}_i^{\mathbf{B}_{-i}}(s, a) - \alpha \hat{Q}_j^{\mathbf{B}_{-i}}(s, a)$ 。在 “相互提议” 协议中, 这可用于确定修改的模拟动作值估计 (SAVE), 用于计算模拟值估计 (SVE)。在 “提议-选取” 协议中, 在选取是否接受合约时可以使用类似的修改, 例如代替 $\hat{Q}_j^{\mathbf{B}_{-\{i,j\}}}(s, (c_i, c_j))$ 。下一个状态的修改的值估计可以被扩展到制裁多个偏离智能体, 例如每一个利用权重 α 。

[0100] 图6和7示出了实施这些技术可以获得的益处的示例。对于一个示例任务,每个图都示出了x轴上交互智能体的总数以及继而y轴上以相对于单个智能体的比率表示的改进。图6是相互提议协议的实施方式;图7是提议-选取协议的实施方式。这些优势非常显著,并且随着智能体数量的增加而增长。

[0101] 所述技术适用于例如许多不同类型的路由、调度和控制任务,诸如路由机器人或自动驾驶载具、电力或数据分组;调度计算作业;以及控制交互以制造产品的机器人或制造单元或机器。

[0102] 在一些实施方式中,环境是真实世界环境,并且至少第一智能体,并且在实施方式中每个智能体包括机器人或自主或半自主载具。由该智能体或多个智能体执行的任务然后可以各自包括导航通过环境的路径和/或操纵在环境中的对象。动作可以包括控制智能体在环境中的移动的动作,例如控制智能体的速度或移动方向(例如转向)的动作,或由机器人或自主载具执行的动作,诸如收集或运送诸如货物或乘客等的有效载荷,或操纵环境中的对象的动作。

[0103] 合约可以限制(或要求)任何这样的动作,例如以改进任务的总体效率为目的。例如,通过将特定智能体限制到特定路线或通过限制任何特定路线上的智能体数量,可以减少路线上的拥堵或运送时间或能源使用。通过为智能体定义动作,提议-选取类型协议可用于联合优化多个智能体的路线。

[0104] 作为另一个示例,与环境中的对象——例如要由智能体运输的对象——相关的动作可以包括需要智能体会面并将一个或多个对象从一个智能体传递到另一个智能体的动作,例如以交换货物。与这样的动作相关的合约,例如互相有利的相互提议类型合约,可用于优化将对象运输到其相应目的地。

[0105] 期望回报和奖励可能与诸如时间或延迟的任务效能度量或(例如从智能体的位置)到达目的地的距离和/或诸如到达目的地的燃料或能源使用的执行任务的物理成本和/或例如与一个或多个特定路线上的拥堵相关的成本的另一成本有关。在实施方式中,该方法可以尝试最大化奖励或期望回报,或者最小化成本。例如,期望回报可能取决于例如在特定时间内沿路径移动所接收的奖励,或者取决于智能体的运动速度,或者取决于由智能体所使用能量的少于特定能量。

[0106] 包括智能体的环境的状态可以通过使用任何类型的传感器从环境的观察得出的状态表示来表征,所述任何类型的传感器例如是图像或视频传感器(这里包括激光雷达(LIDAR)传感器);或智能体状态、定位的传感器;或配置,或环境传感器,诸如一个或多个其他智能体的状态、定位或配置的传感器;或者,例如智能体移动平均速度的传感器;或者拥堵传感器,例如用于感测智能体能够选择的一个或多个可能路线上的拥堵;或者诸如树木、建筑物、行人或交通灯的静态或动态对象的位置的传感器;或者地图数据,例如用于定义智能体可能采取的路线;或者通常地,捕获与环境中对象的状态、定位或配置有关的数据的传感器或设备。

[0107] 作为示例,由智能体执行的任务可以包括交付诸如货物的有效载荷,例如一个或多个包裹,或乘客。例如,智能体可以合作以节能的方式将有效载荷交付到多个不同的目的地和/或避免在运输途中或到达时相互拥堵(例如,通过定义到特定目的地的交付之间的最小时间间隔)。合约可能涉及诸如确定哪个智能体要交付哪个项和/或到哪个目的地,或要

采取或避免的路线等动作。

[0108] 作为一个特定示例,智能体可以包括执行诸如仓库、物流或工厂自动化的任务的机器人或载具,例如在其制造过程中收集、放置或移动存储的货物或货物或货物的部分。

[0109] 另一个示例是,智能体可以包括电动出租车,并且智能体可以确定合约,使出租车能够以省时省能的方式合作接送乘客,例如考虑可用范围。

[0110] 另一个示例是,智能体可以包括汽车或卡车,并且智能体可以确定合约,使载具在道路上行驶时能够合作,例如通过定义旨在维持载具之间最大、最小或目标距离的奖励来促进交通流量的平滑和安全。

[0111] 作为另一个示例,载具可以是空中载具,诸如无人机。例如,在农业环境中,任务然后可以包括向植物交付有效载荷,诸如农药、种子、水或肥料。在紧急响应环境中,任务可以包括搜索人员。在另一种环境中,任务可以包括包裹递送任务。

[0112] 通常,所述技术可用于管理在陆地、空中或水上操作的大量机器人或自主或半自主载具。通常,先前描述的模拟器可以模拟环境中的智能体——例如载具——的动的影响。例如,模拟器可以采用当前状态和智能体动作,并输出表征下一个状态的数据。例如,当前状态和智能体动作可以包括智能体的当前位置,以及关于由智能体所采取的路线的选择(例如,在交叉路口左转会导致载具使用一条道路而不是另一条道路),并且然后模拟器可以确定每条道路上的拥堵程度以及由此产生的行进时间。

[0113] 动作和合约可能与低级、中级或高级动作有关。高级动作的示例包括指定机器人或载具应移动到或避开的位置的动作;中级动作的示例包括指定在路线的交叉点或交叉口处哪个机器人或载具首先移动或具有优先权的动作。低级动作可以为智能体指定特定的速度或方向。通常,定义期望回报的奖励可以基于事件或环境的一个方面。例如,奖励可以指示智能体是否已完成如上所述的任务,或者智能体在完成任务方面的进展(特定奖励与上述任务相对应,并且因此为了简洁起见不再重复)。

[0114] 在一些实施方式中,环境是计算环境,并且每个智能体包括计算作业调度智能体,例如管理跨计算资源——例如在移动设备上、在计算机中、在计算机集群中或在数据中心中——的计算任务分布的软件智能体。计算资源可以包括各种硬件资源。资源可以包括以下:不同的低级资源,诸如具有不同能力的处理器,例如一个或多个通用处理器(CPU)或GPU(图形处理单元)或TPU(张量处理单元);或不同的中间级资源,诸如在计算机集群中一起工作的不同计算机;或不同的较高级资源,诸如数据中心中的服务器。硬件资源可以但不必具有不同的计算或存储器能力。由智能体执行的任务可以包括执行计算机程序。动作可以包括将与任务相关联的一个或多个计算作业分配给一个或多个计算硬件资源的动作。合约可以限制这样的动作,例如将硬件计算资源的可用性限制给特定智能体,例如按时间或资源类型,或要求特定智能体使用特定资源。期望回报和奖励可能与任务效能度量有关,诸如完成任务的时间,或执行任务的计算成本(例如,这取决于实现任务的硬件),或执行任务的功率或能源成本,或执行任务的可靠性(例如,重复执行任务以实现容错)。环境状态可以通过从计算环境的观察中得出的状态表示来表征,例如硬件资源上的计算工作量、例如硬件资源的存储器使用率/可用性、硬件资源的诸如温度的物理状态、或者硬件资源是否出现故障。合约可能涉及调度计算作业,例如用于控制计算作业对硬件资源的分配。模拟器可以模拟环境中动作的影响,例如将特定计算作业分配给特定资源的影响。

[0115] 在一些实施方式中,环境是分组通信网络环境,并且智能体各自包括路由器以通过通信网络路由数据分组。由智能体执行的任务然后可以包括分组路由任务。动作可以包括路由动作以路由数据分组,例如用于确定一个或多个数据分组的下一个或进一步目的地的动作。合约可以限制这些动作,例如用于协调分组传输并减少拥堵,并且通常控制通过通信网络对数据分组的路由。期望回报和奖励可能与一个或多个分组路由度量有关,例如与从路由器到数据分组通信网络中的下一个或进一步节点的路径有关。这样的度量可以包括,例如一组一个或多个路由数据分组从路由器行进到数据分组通信网络中的下一个或进一步节点的估计时间、延迟、速度、距离或跳数、路由利用率的测量、路由分组丢失或其他可靠性测量、路由带宽或吞吐量、或路由最大传输单元(MTU)。环境的状态可以通过从可用数据分组路由的观察及其特性——例如与上述度量相对应的特性——得出的状态表示来表征。例如,观察可以包括对路由器的路由表的观察。模拟器可以模拟不同路由选择的影响。分组通信网络的环境在某些方面类似于先前描述的移动智能体被路由的真实世界环境,并且该领域的概念可以被对应地应用于分组通信网络。

[0116] 在一些实施方式中,环境是电力分布环境,并且智能体各自被配置为控制电力从电力分布节点——例如与智能体相关联的节点——通过例如在“智能电网”中的一个或多个电力分布链路到一个或多个其他节点的路由。

[0117] 随着电力分布电网变得更加分散,例如由于增加了多个较小容量和可能间歇性的可再生发电机,发电机和消费者当中的附加互连可能使电网不稳定,并可能导致电网中的链路过载,特别是由于连接点之间的相位差异。

[0118] 由智能体执行的任务包括将电力从发电机分配给电力消费者的任务。这些动作包括用于控制节点之间的电力路由的控制动作。例如,这些动作可以包括将电力源或电力汇连接到电网或与电网断开连结的动作,或控制电力通过电网节点之间的链路的路由的动作,或控制电力发电机的频率或相位的动作,或控制例如经由节点连接的电网元件之间频率不匹配或相位不匹配的动作。所述合约可以协调这些动作,例如以提高电网的稳定性并降低电网过载的风险;并且通常可以与控制电力的路由有关。期望回报和奖励可能与一个或多个电力路由度量有关,例如与电网内的电力损失有关,或与一个或多个电力分布链路上的频率或相位不匹配有关,或与过载或故障风险有关,或与一个或多个电力分布链路上超过阈值的风险有关。

[0119] 环境状态可以通过从来自传感器的观察中得出的状态表示来表征,这些传感器感测电网的状态,例如感测与节点之间的一个或多个连接相关联的频率、相位、电流、电压、电负载、电阻或阻抗,或电网配置的静态观察,例如与电网和电力分布链路耦合的元件。模拟器可以模拟不同电力控制或路由动作选择的影响。

[0120] 在一些实施方式中,环境是用于制造产品——诸如化学、生物或机械产品或食品——的真实世界制造环境。如本文所用,“制造”产品还包括精炼起始材料以产生产品,或处理起始材料,例如以去除污染物以生成清洁或回收的产品。制造工厂可以包括多个制造单元,例如用于化学或生物物质的容器,或用于处理固体或其他材料的机器,诸如机器人。制造单元可以被配置为使得产品的中间版本或组件在产品制造期间可在制造单元之间移动,例如经由管道或机械输送。

[0121] 在这样的实施方式中,至少第一智能体,并且在实施方式中每个智能体,包括控制

系统,该控制系统被配置为控制机械、化学或生物产品的制造。例如,控制系统可以被配置为控制一个或多个制造单元或机器,或者控制产品的中间版本或组件在制造单元或机器之间的移动。通常,制造单元或机器相互作用以制造产品。通常,由该智能体或多个智能体执行的任务包括制造该机械、化学或生物产品或另一种机械、化学或生物产品或其中间体或组件的任务。也就是说,智能体可以相互作用以控制相同产品的制造,或者它们可以控制不同产品的制造。如本文所用,产品制造包括由一个或多个厨房机器人制造食品产品。

[0122] 动作可以包括控制动作,该控制动作用于控制使用机器或制造单元来处理固体或液体材料以制造产品或中间体或其组件,或者控制产品的中间版本或组件在制造环境内例如在制造单元或机器之间的移动。通常,动作可以是对环境的观察状态有影响的任何动作,即被配置为调整下文所述的任何感测参数的动作。这些可以包括调整制造单元的物理或化学条件的动作,或控制机器机械部件或机器人关节的移动的动作,或控制制造单元之间材料的移动的动作。合约可以以改进制造效率的目的限制这样的动作,例如通过减少制造单元或机器使用的竞争;并且通常可以控制制造环境内的制造条件和/或材料的移动。该领域的方面再次类似于先前所述的领域,并且适用对应的概念。

[0123] 期望回报和奖励可能与任务效能的度量有关,诸如制造的产品数量、产品的质量、产品的生产速度的度量,或执行制造任务的物理成本的度量,例如用于执行任务的能源、材料或其他资源的数量的度量。

[0124] 环境状态的特征可以是,从来自感测制造环境状态的传感器的观察得出的状态表示,例如感测制造单元或机器的状态或配置的传感器,或感测制造单元或机器之间材料的移动的传感器。作为一些示例,这样的传感器可以被配置为感测机械移动或力、压力、温度;电气条件,诸如电流、电压、频率、阻抗;一种或多种材料的数量、水平、流动/移动速率或流动/移动路径;物理或化学条件,例如物理状态、形状或配置或化学状态,诸如pH;单元或机器的配置,诸如单元或机器的机械配置或阀门配置;图像或视频传感器,用于捕获制造单元或机器或移动的图像或视频观察;或任何其他适当类型的传感器。在诸如机器人的机器的情况下,来自传感器的观察可以包括对机器的一个或多个部分的位置、线性或角速度、力、扭矩或加速度或姿势的观察,例如表征机器或机器人或由机器或机器人持有或处理的项的当前状态的数据。观察还可以包括例如感测到的电子信号,诸如电机电流或温度信号,或者例如来自相机或激光雷达传感器的图像或视频数据。诸如这些的传感器可以是环境中智能体的一部分或与智能体分开放置。模拟器可以模拟制造环境中上述动作的不同选择的影响。

[0125] 对于被配置为执行特定操作或动作的一个或多个计算机的系统,意味着该系统已在其上安装了软件、固件、硬件或其组合,它们在动作中导致系统执行动作或操作。对于被配置为执行特定操作或动作的一个或多个计算机程序,意味着一个或多个程序包括指令,所述指令在由数据处理装置执行时,导致装置执行操作或动作。

[0126] 本说明书中描述的主题的实施例和功能操作可以在数字电子电路系统中、在有形体现的计算机软件或固件中、在计算机硬件中——包括本说明书中公开的结构及其结构等效物——中、或在它们中的一个或多个的组合中实现。本说明书中描述的主题的实施例可以实现为一个或多个计算机程序,即,一个或多个计算机程序指令模块,这些指令被编码在有形的非暂时性程序载体上,用于由数据处理装置执行或控制数据处理装置的操作。可替

代地或另外,程序指令可以被编码在人工生成的传播信号——例如机器生成的电、光或电磁信号上,该信号被生成以对信息进行编码,以供传输到合适的接收装置,供数据处理设备执行。计算机存储介质可以是机器可读存储设备、机器可读存储基板、随机或串行访问存储器设备,或它们中的一个或多个的组合。

[0127] 术语“数据处理装置”是指数据处理硬件,涵盖用于处理数据的所有类型的装置、设备和机器,包括例如可编程处理器、计算机或多个处理器或计算机。该设备还可以是或进一步包括专用逻辑电路系统,例如FPGA(现场可编程门阵列)或ASIC(专用集成电路)。除了硬件之外,该设备还可以可选地包括为计算机程序创建执行环境的代码,例如构成处理器固件、协议栈、数据库管理系统、动作系统或它们中的一个或多个的组的代码。

[0128] 计算机程序(也可以称为或描述为程序、软件、软件应用、模块、软件模块、脚本或代码)可以用任何形式的编程语言编写,包括编译或解释语言,或声明或过程语言,并且可以以任何形式部署,包括作为独立程序或作为模块、组件、子例程或适合在计算环境中使用的其他单元。计算机程序可以但不需要对应于文件系统中的文件。程序可以存储在保存其他程序或数据的文件的一部分中,例如,存储在标记语言文档中的一个或多个脚本,专用于相关程序的单个文件中,或多个协调文件中,例如,存储一个或多个模块、子程序或代码部分的文件。计算机程序可以被部署在位于一个站点的一个计算机上或跨多个站点分布并通过通信网络互连的多个计算机上执行。

[0129] 本说明书中描述的过程和逻辑流程可以由一个或多个可编程计算机执行,所述一个或多个可编程计算机执行一个或多个计算机程序以通过对输入数据进行操作并生成输出来执行功能。过程和逻辑流程也可以由专用逻辑电路系统——例如,FPGA(现场可编程门阵列)或ASIC(专用集成电路)——执行,并且装置也可以被实现为专用逻辑电路。

[0130] 适合执行计算机程序的计算机包括,例如,可以基于通用或专用微处理器或两者,或任何其他类型的中央处理单元。通常,中央处理单元将从只读存储器或随机存取存储器或两者接收指令和数据。计算机的元件是用于实施或执行指令的中央处理单元和用于存储指令和数据的一个或多个存储器设备。通常,计算机还将包括或可操作地耦合以从用于存储数据的一个或多个大容量存储设备——例如磁盘、磁光盘或光盘——接收数据或向其传输数据或两者都有。然而,计算机不需要具有这样的设备。此外,计算机可以被嵌入到另一设备中,例如移动电话、个人数字助理(PDA)、移动音频或视频播放器、游戏控制台、全球定位系统(GPS)接收器或便携式存储设备,例如通用串行总线(USB)闪存驱动器,仅举几例。

[0131] 适合存储计算机程序指令和数据的计算机可读介质包括所有形式的非易失性存储器、介质和存储器设备,包括例如半导体存储器设备,例如EPROM、EEPROM和闪存设备;磁盘,例如内部硬盘或可移动磁盘;磁光盘;以及CDROM和DVD-ROM盘。处理器和存储器可以由专用逻辑电路系统补充或并入其中。

[0132] 为了提供与用户的交互,本说明书中描述的主题的实施例可以在计算机上实现,该计算机具有显示设备,例如CRT(阴极射线管)或LCD(液晶显示器)监视器,用于向用户显示信息,以及键盘和指点设备,例如鼠标或轨迹球,用户可通过其向计算机提供输入。其他类型的设备也可用于提供与用户的交互;例如,提供给用户的反馈可以是任何形式的感官反馈,例如,视觉反馈、听觉反馈或触觉反馈;并且来自用户的输入可以以任何形式被接收,包括声学、语音或触觉输入。此外,计算机可以通过向由用户使用的设备发送文档和从由用

户使用的设备接收文档来与用户交互；例如，通过响应从Web浏览器接收到的请求，将网页发送到用户客户端设备上的Web浏览器。

[0133] 本说明书中描述的主题的实施例可以在计算系统中实现，该计算系统包括后端组件，例如数据服务器，或包括中间件组件，例如应用服务器，或包括前端组件，例如具有用户可通过其与本说明书中描述的主题的实施方式进行交互的关系图形用户界面或Web浏览器的客户端计算机，或一个或多个此类后端、中间件或前端组件的任何组合。系统的组件可通过例如通信网络的任何形式或介质的数字数据通信互连。通信网络的示例包括局域网（“LAN”）和广域网（“WAN”），例如互联网。

[0134] 计算系统可以包括客户端和服务器。客户端和服务器通常彼此远离并且通常通过通信网络进行交互。客户端和服务器的关系通过动作在相应计算机上并且彼此具有客户端-服务器关系的计算机程序而产生。

[0135] 虽然本说明书包含许多具体的实施细节，但这些细节不应被解释为对任何发明或可能要求保护的范围的限制，而应被解释为对特定发明的特定实施例的特征的描述。本说明书中在分开的实施例的上下文中描述的某些特征也可以在单个实施例中组合实现。相反，在单个实施例的上下文中描述的各种特征也可以在多个实施例中分开实现或以任何合适的子组合实现。此外，尽管特征可能被描述为以某些组合起作用并且甚至最初被要求保护，但在某些情况下，来自要求保护的组合的一个或多个特征可以从组合中去除，并且要求保护的组合可以针对子组合或子组合的变体。

[0136] 类似地，虽然图中以特定次序描绘了动作，但不应理解为要求以所示的特定次序或顺序次序执行这些动作，或要求执行所有所示的动作以实现期望的结果。在某些情况下，多任务和并行处理可能是有利的。此外，上述实施例中各种系统模块和组件的分离不应理解为要求在所有实施例中都进行这样的分离，并且应理解，所描述的程序组件和系统通常可以集成在单个软件产品中或打包在多个软件产品中。

[0137] 已经描述了主题的具体实施例。其他实施例在所附权利要求的范围内。例如，权利要求中所述的动作可以以不同的次序执行，并且仍然实现期望的结果。作为一个示例，附图中描绘的过程不一定需要所示的特定次序或顺序次序来实现期望的结果。在某些实施方式中，多任务和并行处理可能是有利的。

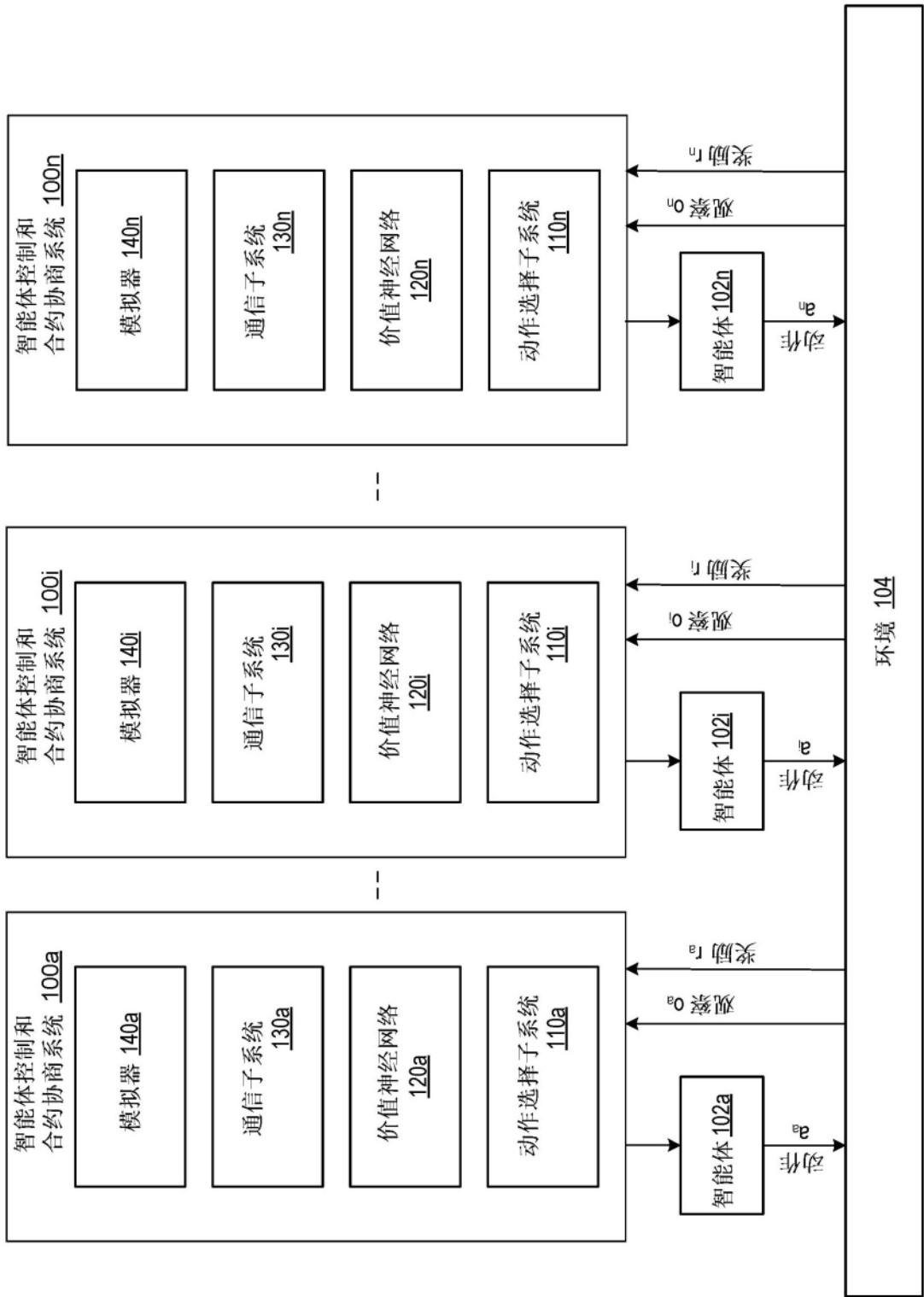


图1

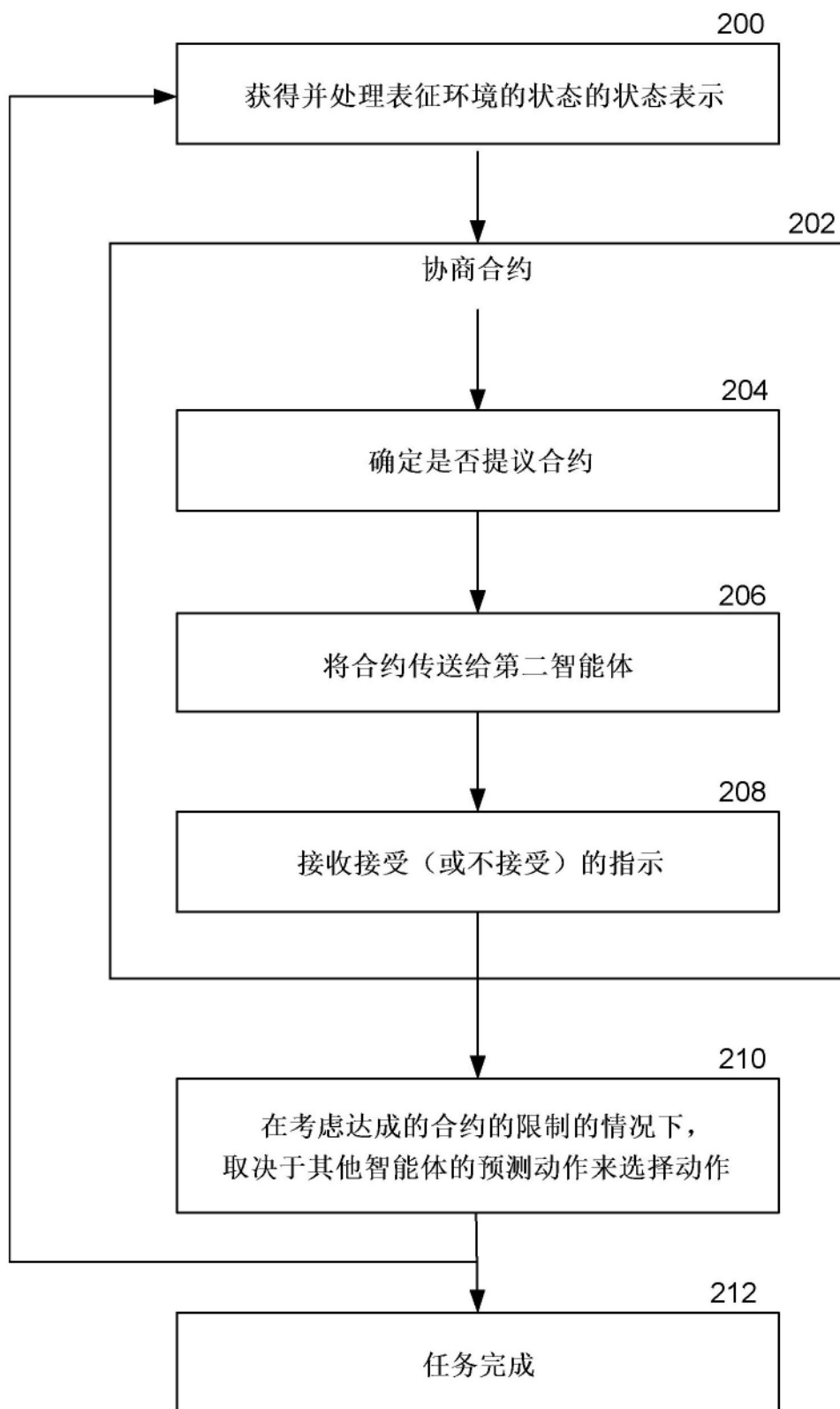


图2

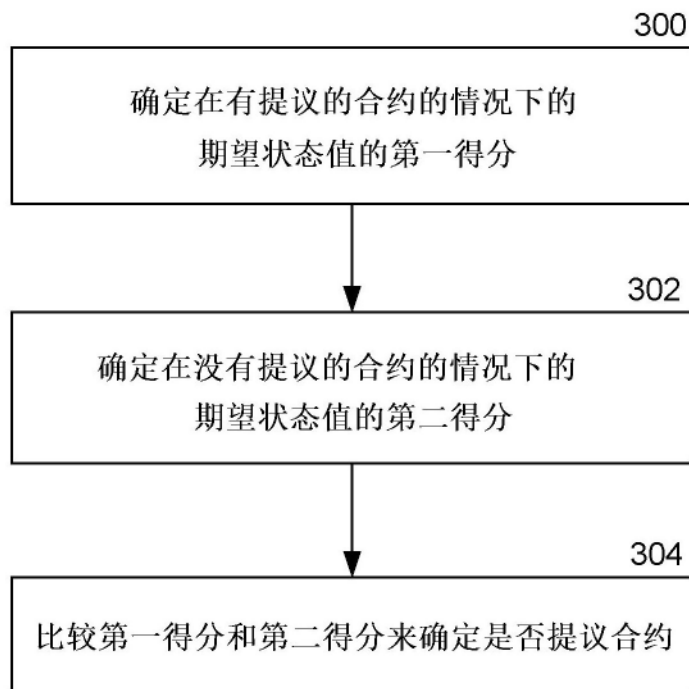


图3

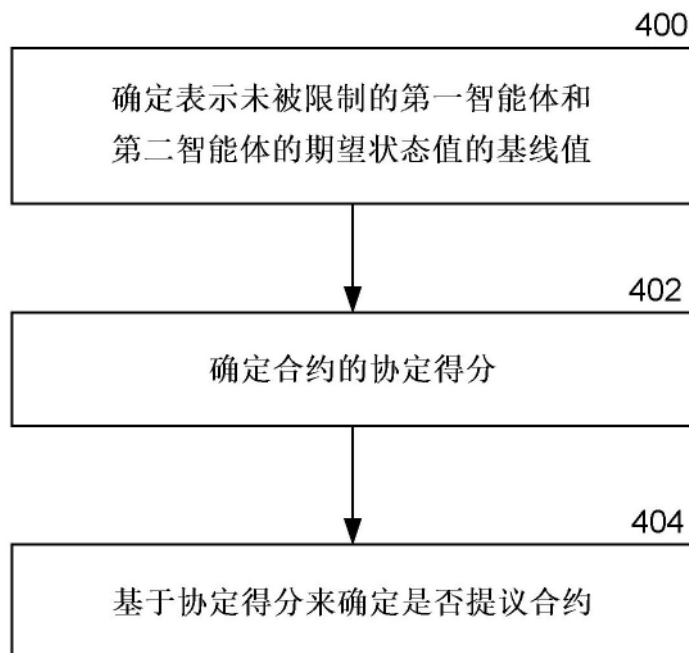


图4

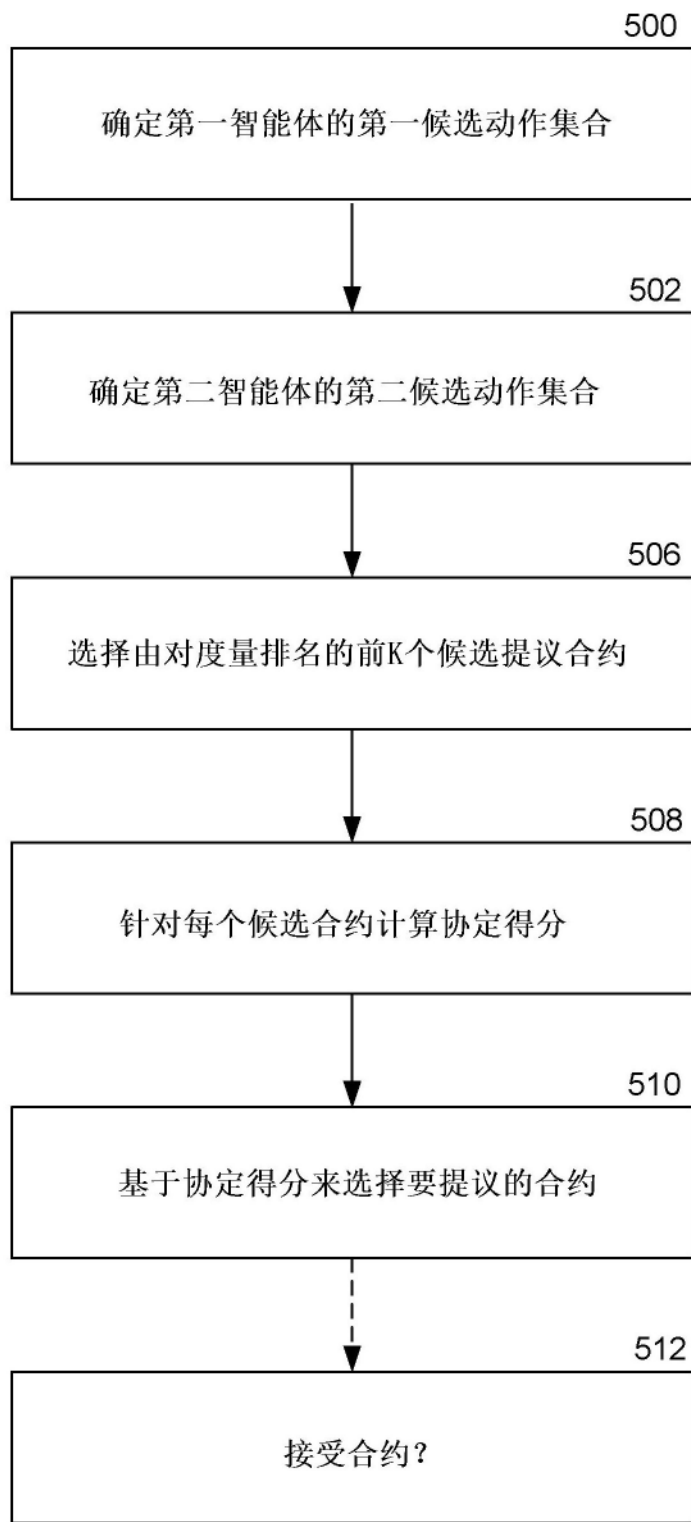


图5

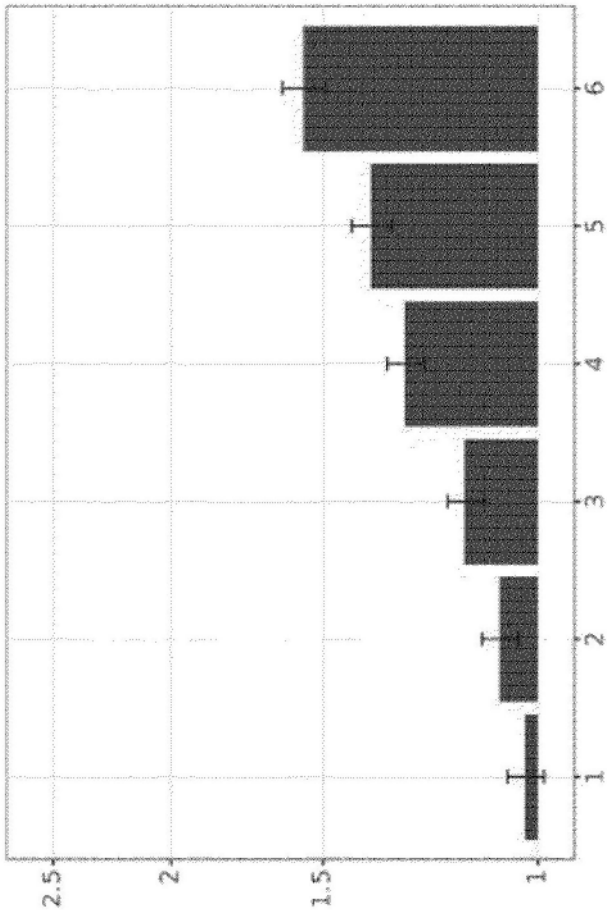


图6

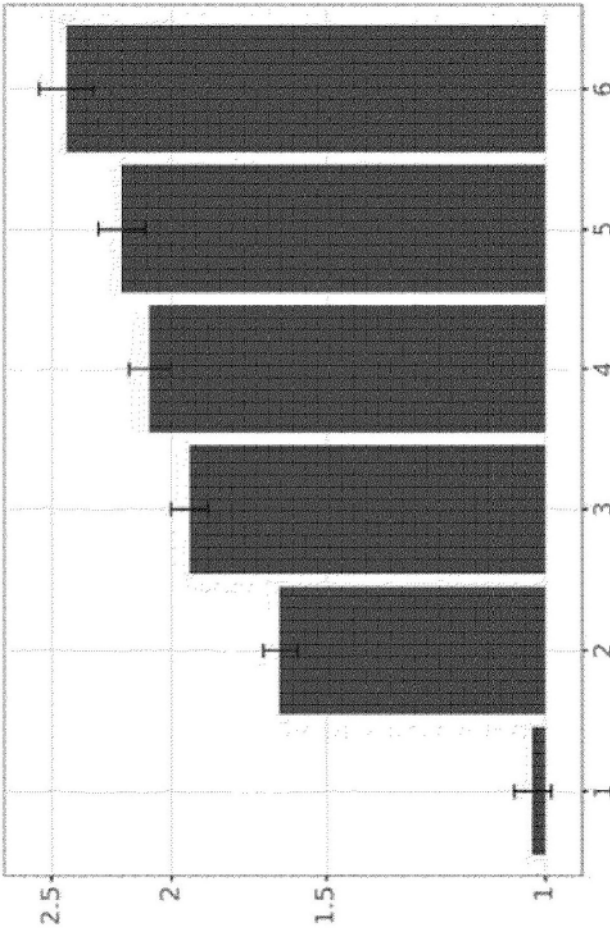


图7