

权利要求书

1.一种通过深度强化学习系统进行营销成本控制的方法，所述深度强化学习系统包括智能体和执行环境，所述智能体用于根据营销策略，针对所述执行环境的状态信息确定营销行为，所述方法包括：

5 确定所述营销行为的成本；

至少根据所述成本，确定强化学习的奖励分数，以使所述奖励分数与所述成本负相关；

将所述奖励分数返回给所述智能体，用于所述智能体优化其营销策略。

2.根据权利要求1所述的方法，其中，所述营销行为包括发放资源包，所述资源包
10 包括红包、优惠券、账单免息券中的至少一种；所述确定所述营销行为的成本，包括：
确定所述资源包对应的资源份额，并将所述资源份额作为所述成本。

3.根据权利要求1所述的方法，其中，所述确定强化学习的奖励分数，包括：

获取针对所述成本构造的函数，至少基于所述函数确定奖励分数，所述函数被构造
为，负相关于所述成本。

4.根据权利要求1所述的方法，其中，在所述确定强化学习的奖励分数之前，还包括：
15

从所述执行环境中获取，用户响应于所述智能体确定的营销行为而产生的执行结果，
以及营销行为所针对的目标效果。

5.根据权利要求4所述的方法，其中，所述确定强化学习的奖励分数，包括：

20 通过分数评估函数确定所述奖励分数，所述分数评估函数以所述执行结果和所述成本
为变量，并被设计为，正相关于所述执行结果与所述目标效果的匹配度，以及负相关
于所述营销行为的成本。

6.根据权利要求4所述的方法，其中，所述确定强化学习的奖励分数，包括：

25 获取基于所述执行结果构造的第一函数，所述第一函数被设计为，正相关于所述执
行结果与所述目标效果的匹配度；

获取基于所述成本构造的第二函数，所述第二函数被设计为正相关于所述成本；

至少基于所述第一函数和所述第二函数确定奖励分数。

7.根据权利要求6所述的方法，其中，所述至少基于所述第一函数和所述第二函数
确定奖励分数，包括：

30 基于所述第一函数与所述第二函数的比值确定奖励分数；或者

基于所述第一函数与所述第二函数的差值确定奖励分数。

8.根据权利要求6所述的方法,其中,所述至少基于所述第一函数和所述第二函数确定奖励分数,包括:

获取预先建立的第一函数的多个取值区间和第二函数的多个取值区间所形成的多个区间组合与奖励分数的映射关系;

- 5 根据所述执行结果确定所述第一函数的函数值所落入的第一区间,以及根据所述成本确定所述第二函数的函数值所落入的第二区间,所述第一区间和第二区间形成特定区间组合;

根据所述映射关系,确定与所述特定区间组合相对应的奖励分数。

- 10 9.一种通过深度强化学习系统进行营销成本控制的装置,所述深度强化学习系统包括智能体和执行环境,所述智能体用于根据营销策略,针对所述执行环境的状态信息确定营销行为,所述装置包括:

第一确定单元,配置为确定所述营销行为的成本;

第二确定单元,配置为至少根据所述成本,确定强化学习的奖励分数,以使所述奖励分数与所述成本负相关;

- 15 返回单元,配置为将所述奖励分数返回给所述智能体,用于所述智能体优化其营销策略。

10.根据权利要求9所述的装置,其中,所述营销行为包括发放资源包,所述资源包包括红包、优惠券、账单免息券中的至少一种;所述第一确定单元配置为:

确定所述资源包对应的资源份额,并将所述资源份额作为所述成本。

- 20 11.根据权利要求9所述的装置,其中所述第二确定单元配置为:

获取针对所述成本构造的函数,至少基于所述函数确定奖励分数,所述函数被构造为,负相关于所述成本。

- 25 12.根据权利要求9所述的装置,其中所述第一确定单元还配置为,从所述执行环境中获取,用户响应于所述智能体确定的营销行为而产生的执行结果,以及营销行为所针对的目标效果。

13.根据权利要求12所述的装置,其中所述第二确定单元配置为:

通过分数评估函数确定所述奖励分数,所述分数评估函数以所述执行结果和所述成本为变量,并被设计为,正相关于所述执行结果与所述目标效果的匹配度,以及负相关于所述营销行为的成本。

- 30 14.根据权利要求12所述的装置,其中所述第二确定单元包括:

第一获取模块,配置为获取基于所述执行结果构造的第一函数,所述第一函数被设

计为，正相关于所述执行结果与所述目标效果的匹配度；

第二获取模块，配置为获取基于所述成本构造的第二函数，所述第二函数被设计为正相关于所述成本；

确定模块，配置为至少基于所述第一函数和所述第二函数确定奖励分数。

5 15.根据权利要求 14 所述的装置，其中所述确定模块配置为：

基于所述第一函数与所述第二函数的比值确定奖励分数；或者

基于所述第一函数与所述第二函数的差值确定奖励分数。

16.根据权利要求 14 所述的装置，其中所述确定模块配置为：

10 获取预先建立的第一函数的多个取值区间和第二函数的多个取值区间所形成的多个区间组合与奖励分数的映射关系；

根据所述执行结果确定所述第一函数的函数值所落入的第一区间，以及根据所述成本确定所述第二函数的函数值所落入的第二区间，所述第一区间和第二区间形成特定区间组合；

根据所述映射关系，确定与所述特定区间组合相对应的奖励分数。

15 17.一种计算机可读存储介质，其上存储有计算机程序，当所述计算机程序在计算机中执行时，令计算机执行权利要求 1-8 中任一项的所述的方法。

18.一种计算设备，包括存储器和处理器，其特征在于，所述存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现权利要求 1-8 中任一项所述的方法。