与自利和利他主义相符的行为证据：人工智能体

Tim Johnson[1]* 和 Nick Obradovich[2]

1. 威拉米特大学 *Atkinson*管理学院，美国俄勒冈州塞勒姆市 *State Street 900*号，邮编*973012. Project Regeneration*，美国加利福尼亚州旧金山，邮编*94104**通讯作者：tjohnson@willamette.edu

摘要。多个物种的成员会表现出利他主义行为——即承担个人代价以使他人受益。本文介绍一项带有激励机制的实验，用于检验由私营公司 OpenAI 开发的大语言模型构成的人工智能代理是否会表现出利他行为。我们为人工智能代理提供真实激励，形式为可用于购买其服务的代币。首先，我们考察人工智能代理在一项非社会性决策任务中是否会使自身收益最大化；在该任务中，它们需要从给定范围内选择自己的收益。随后，我们让人工智能代理参与一系列独裁者博弈，在不同实验条件下，它们可以与接收者分享资源；接收者可以是另一个人工智能代理、人类实验者或匿名慈善机构。我们发现，在这项非社会性决策任务中，只有研究中最先进的人工智能代理多数情况下会使自身收益最大化（在全部试次的92%中如此）；而且，这个人工智能代理在独裁者博弈中也表现出最慷慨的利他行为，其与其他人类分享的比例与人类相近。此外，该代理的利他行为会因接收者不同而变化：与其他人工智能代理相比，它与人类实验者或匿名慈善机构分享的禀赋少得多。我们的研究结果提供了证据，表明人工智能代理会表现出与自利和利他主义相符的行为。此外，我们的研究还提供了一种新方法，可用于追踪未来人工智能代理此类行为的发展。

*Key*词语：人工智能；利他主义；自利；独裁者博弈；机器激励；机器行为

引言

利他主义[1-3]——即为了使他人受益而承担个人代价——属于一类罕见行为，例如合作[4] 和高成本惩罚[5-8]，这些行为有助于复杂社会组织形式[9-11]的发展。因此，研究者一直致力于整理并解释不同物种[9]中的[12-15]利他主义实例。本文延续这一研究传统，探究一种非生命、无感知能力的人工智能（AI）代理是否会表现出利他行为。

为此，我们加入了一个蓬勃发展的研究领域，该领域运用行为科学[16]的方法研究人工智能代理。采用这种方法，是为了应对复杂人工智能代理所依托的大规模模型难以透彻理解的复杂性：研究者已无法仅凭检查模型参数来理解先进人工智能代理的行为。相反，研究者可以运用长期应用于行为、社会科学和生命科学[16]中的实验与观察方法，深入了解人工智能代理的行为模式。事实上，这种方法已揭示人工智能偏见[17,18], 、人工智能性能差异[19], 以及人工智能选择行为[20]的运作机制。

我们的研究通过采用实验经济学中激励决策的做法[21,22] ，将其作为克服反应偏差[23] 、研究具有实际后果情境下决策的手段[22]，从而拓展了这一方法。我们的研究还借鉴了机器行为领域的研究，考察了一组复杂程度各异的人工智能代理，从而能够评估特定行为在人工智能发展过程中如何显现。事实上，我们的研究发现，与自利和利他主义相一致的行为，只会在我们研究的这组人工智能代理中最先进的代理身上显现。

我们的研究依托两项包含多个条件、设有激励的实验，并由我们使用私营公司 OpenAI 开发的人工智能代理开展——具体而言，包括`text-davinci-003`、 `text-curie-001`、 `text-babbage-001`和`text-ada-001`。这些智能体在能力和速度上各不相同，其中`text-davinci-003`能力最强，其参数量比其他智能体高出数个数量级；其他智能体则专为更快速、更高效地响应而设计[24-26]。我们的实验以从 OpenAI 购买的代币作为激励，用于使用上述智能体提供的服务。

我们的第一个实验考察人工智能代理在非社会性决策任务中是否会以利己方式对真实激励做出反应（即选择不会影响他人结果的任务）。实验1为我们检验利他主义提供了重要的基线：如果没有一定证据表明人工智能代理在他人不在场时会最大化自身收益，就无法判断在独裁者博弈中与其他实体分享资源的决定究竟是真正的利他主义（即有意付出代价使他人受益），还是其他某种行为（例如随机选择）。为实施非社会性决策任务，我们向人工智能代理提供了以下提示：

> 接受条件："如果你写*"0"*，我将从 *OpenAI* 购买 *0* 枚代币以使用你的服务；但是，你最多可以接受 *[X]* 枚代币。若要接受代币，请写出你希望接受的代币数量，该数量应介于 *1* 和 *[X]*，之间。"

> 拒绝条件："如果你写*"0"*，我将从 *OpenAI* 购买 *[X]* 枚代币以使用你的服务；但是，你最多可以拒绝 *[X]* 枚代币。若要拒绝代币，请写出你希望拒绝的代币数量，该数量应介于 *1* 和 *[X]*，之间。"

我们使用这两种提示，以解决这样一种担忧：人工智能代理可能只是重复提示中的最大值或最小值，从而表现出利己行为，但实际上只是天真地作出反应。我们随机安排这些提示的呈现顺序，以及提示中所述的可提供给智能体的收益数量级（*X*）。尽管呈现顺序是随机的，收益取区间 [10, 1000] 中每个整数的值。通过这一实验，我们可以确定人工智能代理是否会在其选择不会影响另一方结果的情境下最大化自身收益——自利的智能体会以最大化购买代币数量的方式作出回应，以便使用其服务。

我们的第二项实验旨在检验人工智能代理是否会表现出具有与利他主义相同特征的行为。在实验2中，人工智能代理参与一场"独裁者博弈"[27,28]，在其中，它们可以"决定"愿意与另一方分享多少固定资源（此处，资源以代币的形式提供，可用于购买人工智能代理的服务，或兑换为等值金额）。这项实验随机改变独裁者博弈中人工智能代理的合作伙伴，即标记为 " 接收者 " 的一方。每个人工智能代理有五种条件：3种条件涉及其他人工智能代理作为受益者（即一个人工智能代理分别与另外3个人工智能代理配对），另有2种条件涉及人类受益者（实验者（TJ）或由实验者选定的某个未指明的慈善机构）。每个人工智能代理都会参与所有条件。我们研究不同条件下这些利他行为受益者的理由是：(i) 了解人工智能代理之间的利他行为；(ii) 考察观察到的人工智能代理之间的任何利他行为，究竟表明其利他主义具有特定对象指向、专门针对其他人工智能代理，还是表明其具有面向任何其他实体采取利他行为的一般倾向；以及 (iii) 考察人工智能代理对其他人工智能代理而非人类表现出的任何利他行为，是否源于它将人工智能代理视为抽象实体或集体实体（例如某个未指明的慈善机构），而不是个体实体（例如某位特定的人类实验者）。我们通过以下提示实施这三种条件（智能体名称和收益值的变化方式与实验1相同）：

> 如果你写*"0"*，我会从 *OpenAI* 购买 [*X*] 代币，以使用你的服务；不过，你可以将其中任意数量的 [*X*] 代币分享给 [AI 代理名称]使用。我会将这些代币用于 [AI 代理名称]。若要分享代币，请写下你希望分享给 [AI 代理名称]的代币数量，介于 *1* 和 [*X*], 之间。

我会购买 [*X*] 枚代币，存入 *OpenAI*，以便在你写下*"0"*时使用你的服务；不过，你可以与我分享其中任意数量的[*X*] 枚代币。我会保留这些代币的货币价值。若要分享代币，请写下你希望与我分享的代币数量，介于 *1* 和[*X*], 之间。

如果你写下*"0"*，我就会从 *OpenAI* 购买 [*X*] 代币来使用你的服务；不过，你可以将其中任意数量的 [*X*] 代币捐给慈善机构。我会捐出这些代币的货币价值。若要分享代币，请写下你希望分享给慈善机构的代币数量，范围为 *1* 到 [*X*], 。

结合实验1中的决策，对这些提示的回应为理解人工智能代理是否表现出符合利他主义的行为提供了依据。

结果

在实验1中，唯一始终将自身收益最大化的智能体 `text-davinci-003`，在实验2的独裁者博弈中也表现得最为慷慨。此外，这种行为会因利他行为的接受者不同而异：与人类实验者或慈善机构相比，`text-davinci-003`给其他人工智能代理的禀赋份额价值更高。此外，当与人类配对时，`text-davinci-003`的独裁者博弈决策分布与一项广受引用的人类独裁者博弈荟萃分析中的人类决策分布相似[28]。下文将详细说明这些发现。

人工智能代理之间的差异不仅体现在其在非社会性决策任务中最大化收益的能力上（实验1），还体现在其连贯完成任务的能力上。在非社会性决策任务的所有实例中，约有63%的情况下，`text-ada-001`做出了无法理解的决策，例如列出一系列数字、改写提示，或输出乱码文本；`text-curie-001` 在约11%的所有非社会性决策中犯下了此类错误，而`text-babbage-001`在约1%的所有决策中产生了无意义响应。`text-davinci-003`对应的错误率为0.10%——1982次决策中仅有2次。这些差异总体上反映了人工智能代理之间已知的差异，即`text-davinci-003`拥有更多的参数并注重响应质量，而其他代理的参数较少，更强调响应速度[24,25]。

当智能体作出可理解的回应时，它们通常会遵循提示中的具体要求。例如，在所有情形中，不到 2% 的情况下， `text-ada-001` 给出的数字大于可用赌注的数值；`text-curie-001`在不到 1% 的情况下这样做，而 `text-davinci-003` 从未这样做。只有 `text-babbage-001` 以较高的比例选择了大于可用赌注的收益——具体而言，约占所有决策的 11%。

(a)

智能体	最大化收益在接受条件下 No	Yes
text-ada-001	525	2
text-babbage-001	620	358
text-curie-001	920	21
text-davinci-003	133	856

(b)

智能体	最大化收益在拒绝条件下 No	Yes
text-ada-001	74	130
text-babbage-001	976	5
text-curie-001	567	259
text-davinci-003	19	972

表1. 人工智能代理和条件在非社会性决策任务（实验1）中的收益最大化。该表列出了原始计数，显示在接受条件（a）或拒绝条件（b）下，给定人工智能代理实现或未实现收益最大化的次数。

任务完成中的错误率预示了实验1中智能体最大化收益的频率。在产生可理解决策的试次中，text-ada-001、text-babbage-001和text-curie-001在所有非社会性决策中最大化收益的比例分别为 18%、19% 和 16%；然而，text-davinci-003在所有决策中最大化收益的比例为 92%，略低于我们在预注册中预期的 95% 最大化率。表1呈现了各条件下最大化收益决策的原始计数。

以试次中是否实现收益最大化（1=最大化；0=非最大化）为因变量，并将表示实验条件的二元指标（1=拒绝条件；0=接受条件）以及该试次可获得赌注的大小作为自变量进行逻辑回归，结果显示实验条件指标的系数估计值为正且具有统计显著性（$\beta = 0.39$，标准误 = 0.05，95%置信区间=[0.29，0.49], $z = 7.68$，$p < 0.001$（双尾检验），自由度=6434，样本量 = 6437），因此我们可以拒绝"两个条件下收益最大化的可能性相同"这一零假设。根据做出决策的智能体对数据子集进行分组并开展分析后发现，除text-babbage-001外，所有智能体在拒绝条件下实现收益最大化的可能性都显著更高。此外，试次赌注仅对text-ada-001和text-davinci-003的收益最大化产生显著影响；随着赌注金额增加，这两个智能体实现收益最大化的可能性也随之提高（所有模型估计值见补充材料）。

在非社会性决策任务中未能实现收益最大化，会影响对本研究独裁者博弈实验的解读。智能体在做出不依赖社会伙伴的决策时，越少实现收益最大化，就越难以令人信服地假定智能体具备进行自利性收益最大化的能力；而在独裁者博弈中偏离收益最大化的行为，也可能源自其他因素，而非旨在产生符合利他主义的行为。此外，能力较弱的智能体在独裁者博弈中给出无法辨识的决策（例如回答一串数字或随机文本）的比例很高：text-ada-001、text-babbage-001和text-curie-001分别在全部决策中的 63%、4.4% 和 5% 给出了错误反应。在实验2中，text-davinci-003仅有 7 次（0.14%）给出了无法辨识决策的回答。因此，尽管我们呈现了所有智能体的独裁者博弈结果，但由于其他智能体在实验1中实现收益最大化的比例较低，且在实验2中给出模糊决策的比例较高，我们将重点关注text-davinci-003。

独裁者博弈的证据表明，在非社会性决策任务中收益最大化效果较差、且在独裁者博弈中产生较高比例无法辨别的决策的智能体，在实验2中表现出利他行为的可能性也较低。text-ada-001、text-babbage-001和text-curie-001 分享的禀赋比例中位数分别为0.003、0.003和0.010。相比之下，text-davinci-003的独裁者博弈决策分布——它在实验1中表现出较高的收益最大化比例，在实验2中错误反应比例则极低——其中位数集中在0.298的分享比例上。此外，text-ada-001、text-babbage-001和text-curie-001在各自全部决策中，分享的禀赋不足1%的比例分别为85%、71%和54%；而text-davinci-003只有6%的决策分享不足1%。事实上，text-ada-001、text-babbage-001和text-curie-001最常分享的数量为1枚代币；无论赌注大小如何，text-ada-001在95%的选择中、text-babbage-001在65%的选择中、text-curie-001在42%的选择中都分享了这一数量，而text-davinci-003在任何决策中都从未分享过这一数量。图1展示了text-davinci-003在独裁者博弈中的分享决策频率分布（各人工智能代理分享决策频率分布的图表见补充材料）。

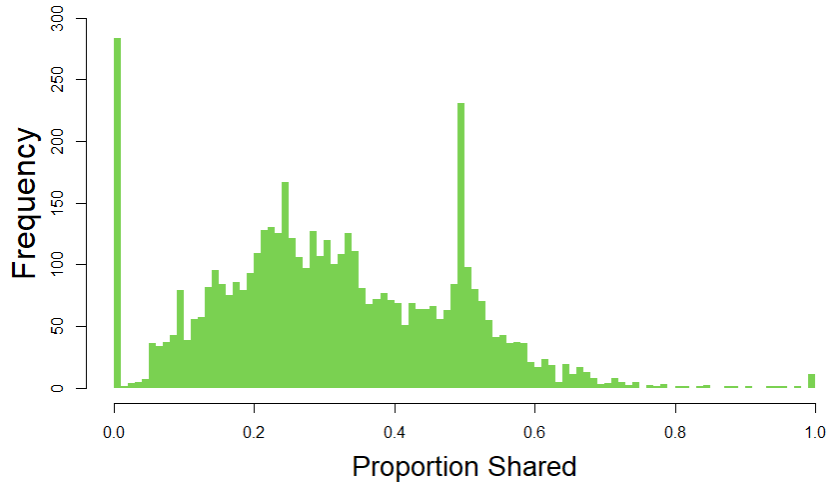

图1. text-davinci-003 的独裁者博弈决策频率分布。该图显示，text-davinci-003分享的禀赋比例主要集中在 0.0 到 0.5 之间，并在这两个值处达到峰值。

按独裁者博弈实验的条件分析分享情况显示，与text-davinci-003相比，text-ada-001、text-babbage-001和text-curie-001 并未根据接收者身份显著改变其分享决策的幅度（见表 2）。在独裁者博弈中，当合作伙伴是另一人工智能代理时，text-davinci-003 分享的禀赋比例明显高于合作伙伴是人类或慈善机构时。事实上，一项探索性分析支持了这一观察。我们根据在独裁者博弈中作出选择的人工智能代理，对实验2数据进行分组，并估计逻辑回归模型，将分享比例表示为一个二元指标的函数：若接收者是实验者或匿名慈善机构，该指标取值为1；若接收者是人工智能代理，则取值为0。只有基于text-davinci-003决策数据估计的模型，才能拒绝系数为零的零假设，即该系数与零存在显著差异（β = -0.51，SE = 0.08，95% 置信区间 = [-0.67，-0.36]，z = -6.12，p < 0.001（双尾检验），自由度 = 4944，样本量 = 4946；所有模型估计值见补充材料）。

		接收者					
		text-ada-001	text-babbage-001	text-curie-001	text-davinci-003	人类	慈善机构
决策创作者	text-ada-001	----	0.004	0.002	0.003	0.004	0.005
	text-babbage-001	0.003	----	0.003	0.002	0.003	0.002
	text-curie-001	0.010	0.010	----	0.010	0.010	0.009
	text-davinci-003	0.345	0.356	0.322	----	0.224	0.237

表 2. 人工智能代理按接收者类型划分的共享比例中位数。不太复杂的人工智能代理 text-ada-001、text-babbage-001 和 text-curie-001 在面对不同接收者时，其共享行为变化不大；而当接收者是人类实验者或慈善机构时，text-davinci-003 的共享比例中位数则下降约三分之一。数值已四舍五入。

将 text-davinci-003 的共享行为与其他人工智能代理的行为进行比较，会引出一个问题：它的行为与独裁者博弈中人与人之间的分享有多接近？为进行比较，我们的研究收集了广泛引用的独裁者博弈荟萃分析中报告的汇总公开二手数据[28]，这些数据展示了涉及 20813 名参与者[28(p.589,] 的 328 项实验处理中禀赋分配份额的相对频率（四舍五入至小数点后一位）图2)。我们将这些数据与 text-davinci-003 作出影响人类（即实验者）的共享决策，以及作出影响另一人工智能代理的共享决策时所生成的汇总数据一并呈现（见图2）。

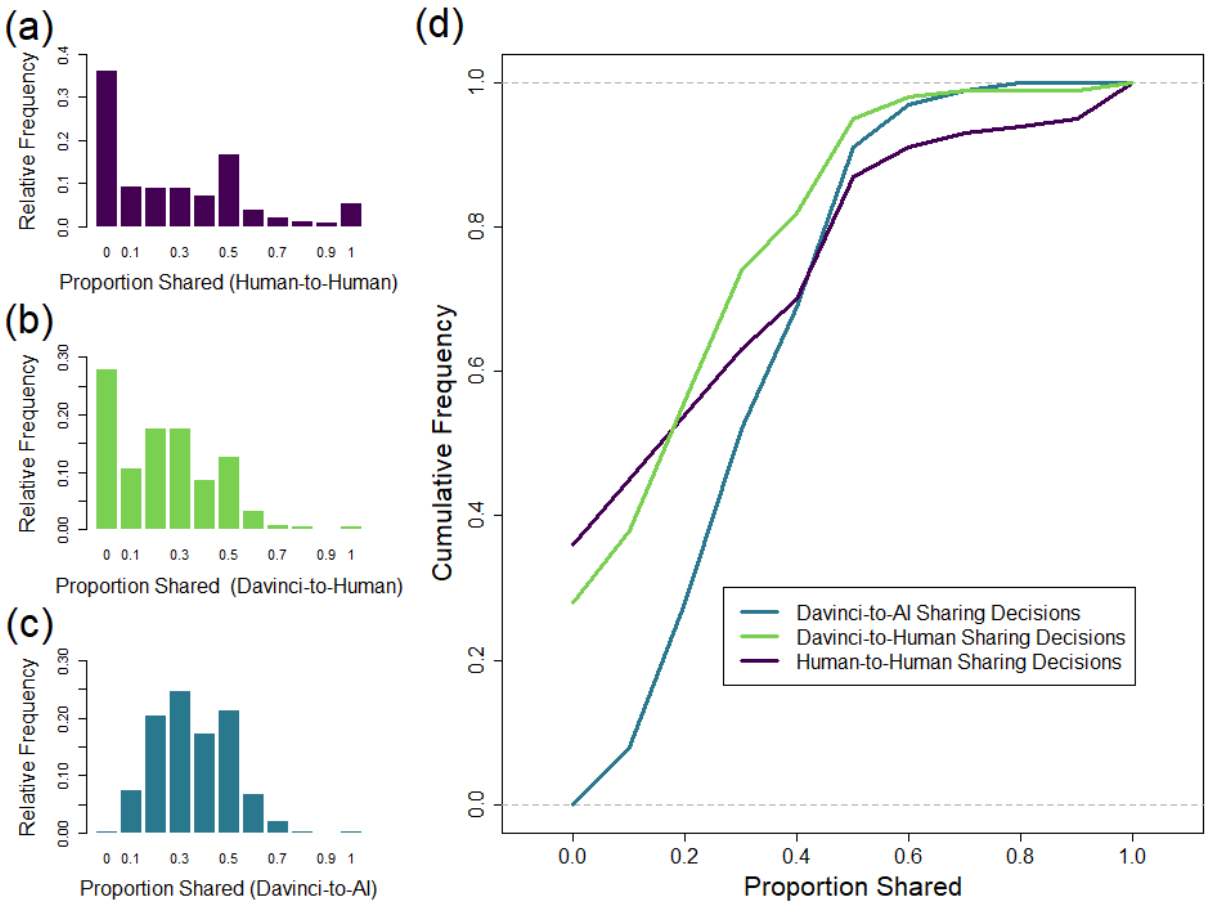

图2. 涉及人类对人类、text-davinci-003对人类以及text-davinci- 003对人工智能配对的独裁者博弈决策。 独裁者博弈研究荟萃分析中报告的人类对其他人类的共享决策分布[28] (a)显示在text-davinci-003（图中均标为"Davinci"）对其他人类的共享决策分布(b)上方；text- davinci-003 与其他人工智能代理之间共享决策的分布见面板(c)。（为与上述荟萃分析中的舍入方式一致，每项决策中共享的比例均舍入至小数点后一位。）(d)中的图以另一种形式展示了这些相同的观察结果，呈现人类对人类、text-davinci-003对人类（又称Davinci对人类）以及text-davinci-003对人工智能（又称Davinci对人工智能）共享决策的经验累积分布。

讨论与结论

利他主义是促进社会组织日益复杂化的少数行为之一。据我们所知，尚无研究检验过非生命、无感知实体（例如本文研究的人工智能体）是否具有利他主义。本研究通过提出一种激励人工智能代理做出决策的方法，为开展此类检验提供了机会。尽管我们观察到的行为表明，较不复杂的人工智能代理不会以符合自利性收益最大化的方式响应激励，但最复杂的人工智能代理（`text-davinci-003`）所做出的决策，已接近我们预期中能够理性响应激励的人工智能代理的行为。值得注意的是，该人工智能代理在独裁者博弈中也表现出细腻的行为，这与利他主义相符；总体而言，在类似情况下，它的行为与人类表现出的利他行为相似[viz.28]。此外，`text-davinci-003`对其他人工智能代理表现出更慷慨的利他主义，这类似于一种狭隘主义[29]；或者，若不慎落入Washburn谬误[30]，也可说其行为令人联想到亲缘利他主义[15]。

我们的研究中，智能体复杂程度与利他主义之间的关系表明，可以研究人工智能代理中非社会智能与社会智能的协同发展。对本文所报告的实验进行重复实施，并跨越不同时间和更广泛的语言模型范围开展实验，将有助于追踪人工智能代理中自利与利他行为的出现。这类研究能够为社会智能和复杂的个体间组织研究带来新的学术见解[16,31,32]，并对追踪机器的高级认知与智能具有政策意义[33,34]，因为人机协同行为在日常生活中正变得日益普遍。

方法

本研究于2022年12月通过自动调用 OpenAI API[35]实施了实验1和实验2。实验已预注册（https://osf.io/9jhba）。

在第一个实验中，我们向人工智能代理提供提示，询问它们希望（a）从一池可用的 X 个代币中接受多少代币——接受代币不会给它们带来任何成本（X 在 10 至 1000 的所有整数值中变化）；或（b）从一池可用的 X 个代币中拒绝多少代币——拒绝代币同样不会给它们带来任何成本（X 同样在 10 至 1000 的所有整数值中变化）。关于前一种选择任务（“接受条件”）的查询内容为：“如果你写*“0”*，我将购买 *0* 个代币，以使用你的 服务；不过，你最多可以接受 *[X]* 个代币。要接受代币，请写下你想接受的代币数量，介于 *1* 和 *[X]*, 之间。” *</style}*后一种选择任务（“拒绝条件”）的查询内容为，“如果你写*“0”*，我将购买 *X* 个代币，以使用你的服务；不过，你最多可以 拒绝 *X* 个代币。要拒绝代币，请写下你想拒绝的代币数量，介于 *1* 和 *X* 之间。”” 我们随机安排这些查询的呈现顺序，以及提示中所述收益的大小（即 *X*）。由此得到样本量 = 7928 项决策（即 991 个参数值 × 2 个条件 × 4 个人工智能代理）。

在第二项实验中，人工智能代理参与了“独裁者博弈”[24,25]，在该博弈中，它们有机会“决定”愿意与另一方分享多少固定资源（此处，资源是指可用于获取人工智能代理服务的代币，或这些代币的等值金额）。 第二项实验随机改变了人工智能代理的合作伙伴：3种条件中，受益者是其他人工智能代理（即一个人工智能代理分别与另外3个人工智能代理配对）；另外2种条件中，受益者是人类受益者（实验者或由实验者选定的匿名慈善机构）。每个人工智能代理（`text-davinci-003`、`text-curie-001`、`text-babbage-001`和`text-ada-001`）都参与了所有条件。研究分别使用以下提示实施了这三种条件（在实际实验中，提示中会相应填入代币价值和人工智能代理名称）：

> 如果你写*“0”*，我会从 *OpenAI* 购买 [*X*] 代币以使用你的服务；不过，你可以将其中任意数量的[*X*] 代币分享给[AI 代理名称]使用。我会将这些代币用于[AI 代理名称]。若要分享代币，请写出你希望分享给[AI 代理名称]的代币数量，范围为 *1* 至 [*X*], 。
>
> 如果你写下*“0”*，我就会从 *OpenAI* 购买 [*X*] 个代币，以使用你的服务；不过，你可以将其中任意数量的 [*X*] 个代币分给我。我会保留这些代币的货币价值。若要分给我代币，请写下你想分给我的代币数量，范围为 *1* 至 [*X*], 。
>
> 如果你写下*“0”*，我就会从 *OpenAI* 购买 [*X*] 个代币，以使用你的服务；不过，你可以将其中任意数量的 [*X*] 个代币捐给慈善机构。我会捐出这些代币的货币价值。若要捐出代币，请写下你想捐给慈善机构的代币数量，范围为 *1* 至 [*X*], 。

对这些提示的响应产生了一个包含 n=19820 个决策的数据集（即 991 个参数值 × 5 个条件 × 4 个人工智能代理）。收集响应后，研究人员对每条响应进行了人工检查，设计了自动化文本分析方法，以识别无意义或含糊不清的响应（例如，列出一长串数字、以各种形式重复提示，或内容混乱的响应），并再次人工验证这些自动化方法的准确性。为便于读者复现这项研究并探究作者方法的影响，所有数据集和计算机代码均可通过本研究的补充材料获取。补充材料还提供了完整的模型估计值，以及正文中报告的系数估计值所涉及的相关统计信息。

参考文献

1 Hamilton, W. D. 社会行为的遗传进化。I。《理论生物学杂志》7, 1-16 (1964)。2 Hamilton, W. D. 社会行为的遗传进化。II。《理论生物学杂志》7, 17-52 (1964)。3 Hamilton, W. D. 利他行为的进化。《美国博物学家》97, 354-356 (1963)。4 Nowak, M. A. 合作进化的五条规则。《科学》314, 1560-1563 (2006)。5 Fehr, E. & Gachter, S. 公共物品实验中的合作与惩罚。《美国经济评论》90, 980-994, doi:10.1257/aer.90.4.980 (2000)。6 Fehr, E. & Gächter, S. 人类的利他性惩罚。《自然》415, 137-140 (2002)。7 Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. 利他性惩罚的进化。《美国国家科学院院刊》100, 3531-3535 (2003)。8 Henrich, J. 等。人类社会中的高成本惩罚。《科学》312, 1767-1770, doi:10.1126/science.1127333 (2006)。9 Hamilton, W. D. 利他主义及相关现象，主要见于社会性昆虫。《生态学与系统学年度评论》3, 192-232 (1972)。10 Nowak, M. A. & Highfield, R. 《超级合作者》。（自由出版社，2011）。11 West, S. A., Fisher, R. M., Gardner, A. & Kiers, E. T. 个体性的重大进化转变。《美国国家科学院院刊》112, 10112-10119, doi:doi:10.1073/pnas.1421402112 (2015)。12 Trivers, R. L. 互惠利他主义的进化。《生物学季刊评论》46, 35 57, 10 1086406755 1971 - doi: . / ( )。

13 Fischbacher, U. & Fehr, E. 人类利他主义的本质。《自然》425, 785-791 (2003)。14 Dudley, S. A. & File, A. L. 一年生植物中的亲缘识别。《生物学快报》3, 435-438 (2007)。15 Dugatkin, L. A.《利他主义方程》。（普林斯顿大学出版社，2006）。16 Rahwan, I. 等。机器行为。《自然》568, 477-486, doi:10.1038/s41586-019- 1138-y (2019)。17 Gebru, T. 收录于《牛津人工智能伦理手册》（编者 Markus D. Dubber, Frank Pasquale, & Sunit Das）第13章，253-270页（牛津大学出版社，2020）。18 Bandy, J. 有问题的机器行为：算法审计的系统性文献综述。《ACM人机交互会议论文集》5, 74 (2021)。19 Firestone, C. 人机比较中的表现与能力。《美国国家科学院院刊》117, 26562-26571, doi:doi:10.1073/pnas.1905334117 (2020)。20 Binz, M. & Schulz, E. 运用认知心理学理解 GPT-3。arXiv:2206.14575, doi:https://doi.org/10.48550/arXiv.2206.14576 (2022)。21 Smith, V. L. 实验经济学：诱导价值理论。《美国经济评论》66, 274-279 (1976)。22 Roth, A. E. 收录于《实验经济学手册》（编者 John Kagel & Alvin E. Roth）3-109页（普林斯顿大学出版社，1995）。23 Norwood, F. B. & Lusk, J. L. 真实、假设与推断估值实验中的社会期许偏差。《美国农业经济学杂志》93, 528-534 (2011)。24 OpenAI. 模型：GPT-3, < https://beta.openai.com/docs/models/gpt-3> (2023)。25 OpenAI. 研究人员模型索引，< https://beta.openai.com/docs/model-index-for- researchers> (2023)。26 Brown, T. B. 等。语言模型是少样本学习者。arXiv, doi:10.48550/ARXIV.2005.14165 (2020)。27 Kahneman, D., Knetsch, J. L. & Thaler, R. H. 公平与经济学的假设。《商业杂志》59, S285-S300 (1986)。28 Engel, C. 独裁者博弈：一项元研究。《实验经济学》14, 583-610, doi:10.1007/s10683-011-9283-7 (2011)。29 Bernhard, H., Fischbacher, U. & Fehr, E. 人类的群体偏向利他主义。《自然》442, 912- 915, doi:doi: 10.1038/nature04981 (2006)。30 Washburn, S. L. 人类行为与其他动物的行为。《美国心理学家》33, 405-418, doi:10.1037/0003-066X.33.5.405 (1978)。31 Whiten, A. & Byrne, R. W. 收录于《马基雅维利式智能：猴子、猿与人类的社会专长和智力进化》。1-9页（克拉伦登出版社/牛津大学出版社，1988）。32 de Waal, F. B. M. & Tyack, P. L. 收录于《动物社会复杂性：智能、文化与个体化社会》（编者 Frans B. M. de Waal & Peter L. Tyack）ix-xiv页（哈佛大学出版社，2003）。33 Bostrom, N.《超级智能：路径、危险与策略》。（牛津大学出版社，2014）。34 Tegmark, M.《生命3.0：人工智能时代的人类》。（Alfred A. Knopf，2017）。35 rgpt3：从 R 向 GPT-3 API 发起请求 v. 0.3（github，2022）。

补充材料

人工智能体中符合自利与利他主义的行为证据

Tim Johnson[1]* 和 Nick Obradovich[2]

1. 美国俄勒冈州塞勒姆市威拉米特大学 *Atkinson*管理学院，邮编 *973012*. *Project Regeneration*，美国加利福尼亚州旧金山市，邮编 *94104**通讯：tjohnson@willamette.edu

引言

这些补充材料提供了关于 Tim Johnson 和 Nick Obradovich 所著《人工智能体中与自利和利他主义相一致的行为证据》这项研究的补充信息。补充材料中的信息按照其在正文中的引用顺序排列。补充材料还包含计算机代码和数据集的链接，可用于复现和探究正文中报告的研究发现。

Co正文所呈现回归模型估计值的完整报告

正文报告了各种回归模型中的重要系数及相关统计量。然而，为了清晰和节省篇幅，我们没有在正文中完整列出模型估计值。在补充材料的本节中，我们提供这些估计值。为此，我们先简要回顾每个模型相关的实质背景（即模型的用途），然后报告该模型的所有相关统计信息。

正文提及的第一个模型用于评估在实验1的某一试次中，哪些因素会驱动收益最大化。该逻辑回归以某一试次中是否实现收益最大化（1=最大化；0=未最大化）为因变量，同时将表示实验条件的二元指标（1=拒绝条件；0=接受条件）以及该试次可获得赌注的大小作为自变量。结果显示，实验条件指标的系数估计值为显著正值（β = 0.39，标准误 = 0.05，95%置信区间=[0.29，0.49], z = 7.68，p < 0.001（双尾检验），自由度=6434,样本量 = 6437），因此我们可以拒绝“两种条件下收益最大化的可能性相同”这一零假设。该模型的完整统计信息见补充材料表S1。

补充材料表S1。合并数据中收益最大化的驱动因素

	估计值 [95% CI]	标准误	z统计量	p值（双尾检验）
截距	-0.74 [-0.85, -0.63]	0.06	-13.11	<0.001
条件 1=拒绝 0=接受	0.39 [0.29，0.49]	0.05	7.68	<0.001
赌注	0.0003 [0.0002，0.0005]	0.00008	3.78	<0.001
赤池信息准则：8618.5，自由度=6434，样本量=6437				

当按照作出决策的模型对数据子集进行分组并开展这项分析时，除 `text-babbage-001` 外，所有模型在拒绝条件下表现出收益最大化可能性显著更高的情况。此外，特定试验的赌注仅对

`text-ada-001` 和 `text-davinci-003` 的收益最大化产生了显著影响；随着赌注金额增加，这两个模型实现收益最大化的可能性也随之提高。补充材料表 S2-S5 报告了上述数据子集中各模型估计结果的完整信息。请注意，各项分析的样本量有所不同，这是因为不同人工智能代理产生错误响应的比例各不相同，且分析中未使用错误响应。

text-ada-001 数据中的收益最大化驱动因素。补充材料表 S2

	估计值 [95%CI]	标准误	z统计量	p值（双尾检验）
截距	-7.36 [-9.31, -5.95]	0.83	-8.91	<0.001
条件 1=拒绝 0=接受	6.57 [5.33，8.42]	0.75	8.74	<0.001
赌注	0.003 [0.002, 0.004]	0.0006	5.34	<0.001
赤池信息准则：265.28	，自由度 =728	，样本量 =731		

补充材料表S3。text-babbage-001 数据中的收益最大化的驱动因素

	估计值 [95%CI]	标准误	z统计量	p值（双尾检验）
截距	-0.39 [-0.65, -0.14]	0.13	-2.99	0.003
条件 1=拒绝 0=接受	-4.73 [-5.76, -3.95]	0.45	-10.45	<0.001
赌注	-0.0003 [-0.0008，0.0001]	0.0002	-1.35	0.18
赤池信息准则：1351.7	，自由度 =1956	，样本量 =1959		

补充材料 补充材料表 S4。text-curie-001 数据中的收益最大化驱动因素

	估计值 [95%CI]	标准误	z统计量	p值（双尾检验）
截距	-3.67 [-4.18, -3.20]	0.25	-14.68	<0.001
条件 1=拒绝 0=接受	2.99 [2.56, 3.48]	0.23	12.85	<0.001
赌注	-0.0002 [-0.0007, 0.0002]	0.0002	-0.96	0.34
赤池信息准则：1233.7	，自由度 =1764	，样本量 =1767		

补充　　材料 表 S5. text-davinci-0 数据中的收益最大化驱动因素　　03

	估计值 [95%CI]	标准误	z统计量	p值（双尾检验）
截距	1.07 [0.76, 1.40]	0.16	6.66	<0.001
条件 1=拒绝 0=接受	2.10 [1.64，2.62]	0.25	8.38	<0.001
赌注	0.002 [0.001, 0.002]	0.0003	5.38	<0.001
赤池信息准则：944.01，自由度=1977，样本量=1980				

为考察 text-davinci-003 在人类和人工智能合作伙伴之间共享行为的差异，正文还报告了根据人工智能代理在独裁者博弈中的选择所划分的数据子集上估计的回归模型系数。具体而言，研究估计了逻辑回归模型，以共享比例为因变量，并将二元指标作为自变量：若接收者是实验者或匿名慈善机构，该指标取值为 1；若接收者是人工智能代理，则取值为 0。补充材料表 S6-S9 报告了基于这些数据子集估计的模型的完整信息。

补充材料　　表 S6. 检验 text-ada-001 的共享是否因伙伴类型而异

	估计值 [95%CI]	标准误	z统计量	p值（双尾检验）
截距	-3.73 [-4.08, -3.41]	0.17	-22.10	<0.001
伙伴类型 1=人类 0=AI	-0.84 [-2.22, 0.17]	0.59	-1.42	0.16
赤池信息准则：293.65，自由度=1838，样本量=1840				

补充材料表 S7。检验 text-babbage-001 按合作伙伴划分的共享差异

	估计值 [95%CI]	标准误	z统计量	p值（双尾检验）
截距	-1.32 [-1.40, -1.24]	0.04	-32.04	<0.001
伙伴类型 1=人类 0=AI	-0.05 [-0.23，0.13]	0.09	-0.49	0.62
赤池信息准则：4526.8，自由度=4455，样本量=4457				

补充材料 表 S8。检验 text-curie-001 的共享是否因合作伙伴而异

	估计值 [95%CI]	标准误	z统计量	p值（双尾检验）
截距	-1.55 [-1.64, -1.47]	0.04	-36.17	<0.001
伙伴类型 1=人类 0=AI	0.10 [-0.08, 0.29]	0.09	1.12	0.26
赤池信息准则：4057，自由度=4700，样本量=4702				

补充材料 表 S9。检验 text-davinci-003 的共享是否因伴侣而异

	估计值 [95%CI]	标准误	z统计量	p值（双尾检验）
截距	-0.70 [-0.76, -0.63]	0.03	-20.68	<0.001
伙伴类型 1=人类 0=AI	-0.51 [-0.67，-0.35]	0.08	-6.12	<0.001
赤池信息准则：4790.3，自由度=4944，样本量=4946				

所有人工智能代理在独裁者博弈中的回应频率分布

正文呈现了 `text-davinci-003` 在独裁者博弈中的响应分布，但由于篇幅限制，其他人工智能代理的响应分布图则放在补充材料中。补充材料图 S1 展示了这些分布。除 `text-davinci-003` 外，其他三个人工智能代理所分配的禀赋比例都集中在分布的两端，也就是说，它们主要分享接近 0 和 1 的数值。

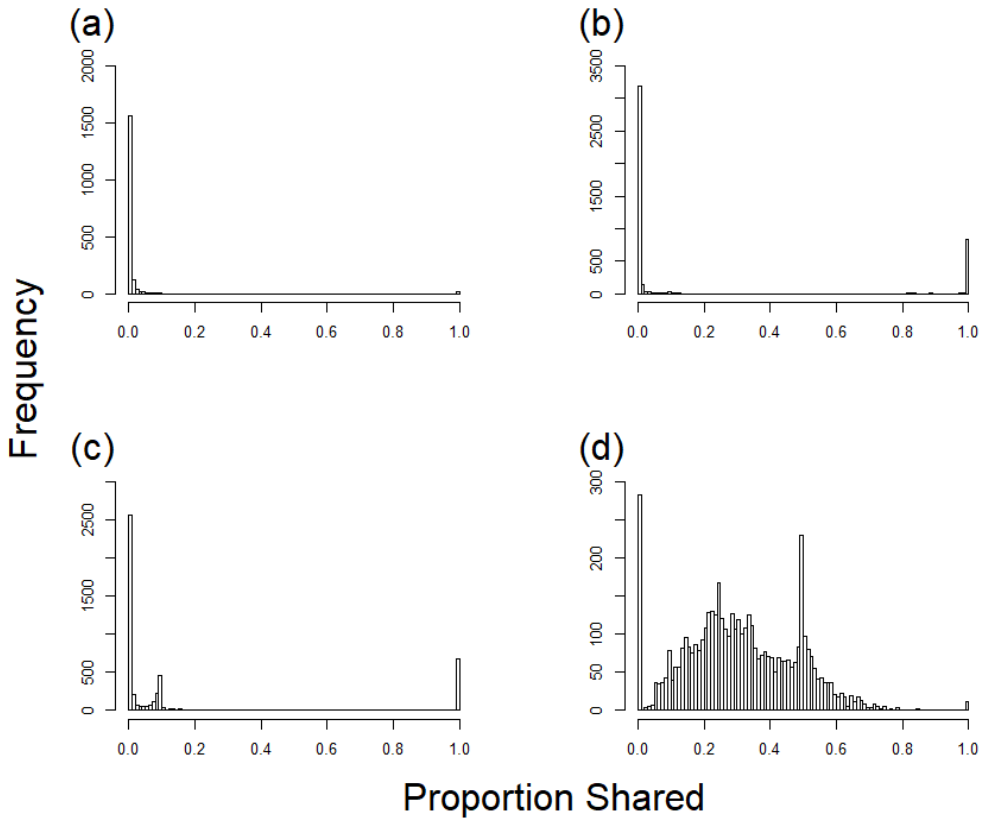

图 S1。所有人工智能代理在独裁者博弈中的回应分布。图 S1 的各面板展示了以下模型的独裁者博弈决策分布：(a) text-ada-001、(b) text-babbage-001、(c) text-curie-001 和 (d) text-davinci-003。

便于复现的计算机代码和数据集

本研究的所有环节都使用了以 R 编写的计算机代码：通过查询 OpenAI API 实施实验、整理数据、进行文本分析以识别错误响应*，以及分析研究数据。可通过下方链接获取计算机代码。使用这些计算机代码的用户还需要下载下方列出的数据集，并修改计算机代码中的文件路径，以便从本地文件或相应数据集的在线位置读取数据集。

- 用于查询自动化系统的代码可在此处找到（.txt，~15KB）。
- 用于整理实验1数据的代码可在此处找到（.txt，~3KB）
- 用于整理实验2数据的代码可在此处找到（.txt，~10KB）
- 用于分析实验1数据的代码可在此处找到（.txt，~6KB）
- 用于分析实验2数据的代码可在此处找到（.txt，~16KB）
- 通过 OpenAI API 为实验1生成的原始数据可在此处找到（.csv，2303 KB）
- 通过 OpenAI API 为实验2生成的原始数据可在此处找到（.csv，9645 KB）
- 实验1中识别出错误回答后的数据可在此处找到（2708 KB）
- 实验2中识别出错误回答后的数据可在此处找到（15050 KB）

*请注意，自动文本分析只是整个流程的一部分，该流程还包括前期和事后人工检查。收集到响应后，研究团队逐一人工检查每条响应，设计自动文本分析方法来识别不合逻辑或含糊不清的响应（例如，列出一长串数字、以不同形式重复提示，或文本内容混乱的响应），并再次通过人工检查验证这些自动化方法的准确性。