



OPEN

# Human-centred mechanism design with Democratic AI

Raphael Koster<sup>1,5</sup>, Balaguer Jan<sup>1,5</sup>, Andrea Tacchetti<sup>1</sup>, Ari Weinstein<sup>1</sup>, Tina Zhu<sup>1</sup>, Oliver Hauser<sup>2</sup>, Duncan Williams<sup>1</sup>, Lucy Campbell-Gillingham<sup>1</sup>, Phoebe Thacker<sup>1</sup>, Matthew Botvinick<sup>1,3</sup> and Christopher Summerfield<sup>1,4</sup> ✉

**Building artificial intelligence (AI) that aligns with human values is an unsolved problem. Here we developed a human-in-the-loop research pipeline called Democratic AI, in which reinforcement learning is used to design a social mechanism that humans prefer by majority. A large group of humans played an online investment game that involved deciding whether to keep a monetary endowment or to share it with others for collective benefit. Shared revenue was returned to players under two different redistribution mechanisms, one designed by the AI and the other by humans. The AI discovered a mechanism that redressed initial wealth imbalance, sanctioned free riders and successfully won the majority vote. By optimizing for human preferences, Democratic AI offers a proof of concept for value-aligned policy innovation.**

The ultimate goal of AI research is to build technologies that benefit humans — from assisting us with quotidian tasks to addressing grand existential challenges facing society<sup>1</sup>. Machine learning systems have already solved major problems in biomedicine<sup>2</sup>, and helped address humanitarian and environmental challenges<sup>3,4</sup>. However, an underexplored frontier is the deployment of AI to help humans design fair and prosperous societies<sup>5</sup>. In economics and game theory, the field known as mechanism design studies how to optimally control the flow of wealth, information or power among incentivized actors to meet a desired objective, for example by regulating markets, setting taxes or aggregating electoral votes<sup>6,7</sup>. Here we asked whether a deep reinforcement learning (RL) agent could be used to design an economic mechanism that is measurably preferred by groups of incentivized humans.

The challenge of building AI systems whose behaviour is preferred by humans is called the problem of ‘value alignment’. One key hurdle for value alignment is that human society admits a plurality of views, making it unclear to whose preferences AI should align<sup>8</sup>. For example, political scientists and economists are often at loggerheads over which mechanisms will make our societies function most fairly or efficiently. In AI research, there is a growing realization that to build human-compatible systems, we need new research methods in which humans and agents interact<sup>9–13</sup>, and an increased effort to learn values directly from humans to build value-aligned AI<sup>14</sup>. Capitalizing on this idea, here we combined modern deep RL with an age-old technology for arbitrating among conflicting views—majoritarian democracy among human voters—to develop a human-centred research pipeline for value-aligned AI research. Instead of imbuing our agents with purportedly human values a priori, and thus potentially biasing systems towards the preferences of AI researchers, we train them to maximize a democratic objective: to design policies that humans prefer and thus will vote to implement in a majoritarian election. We call our approach, which extends recent related participatory approaches<sup>11,14,15</sup>, ‘Democratic AI’.

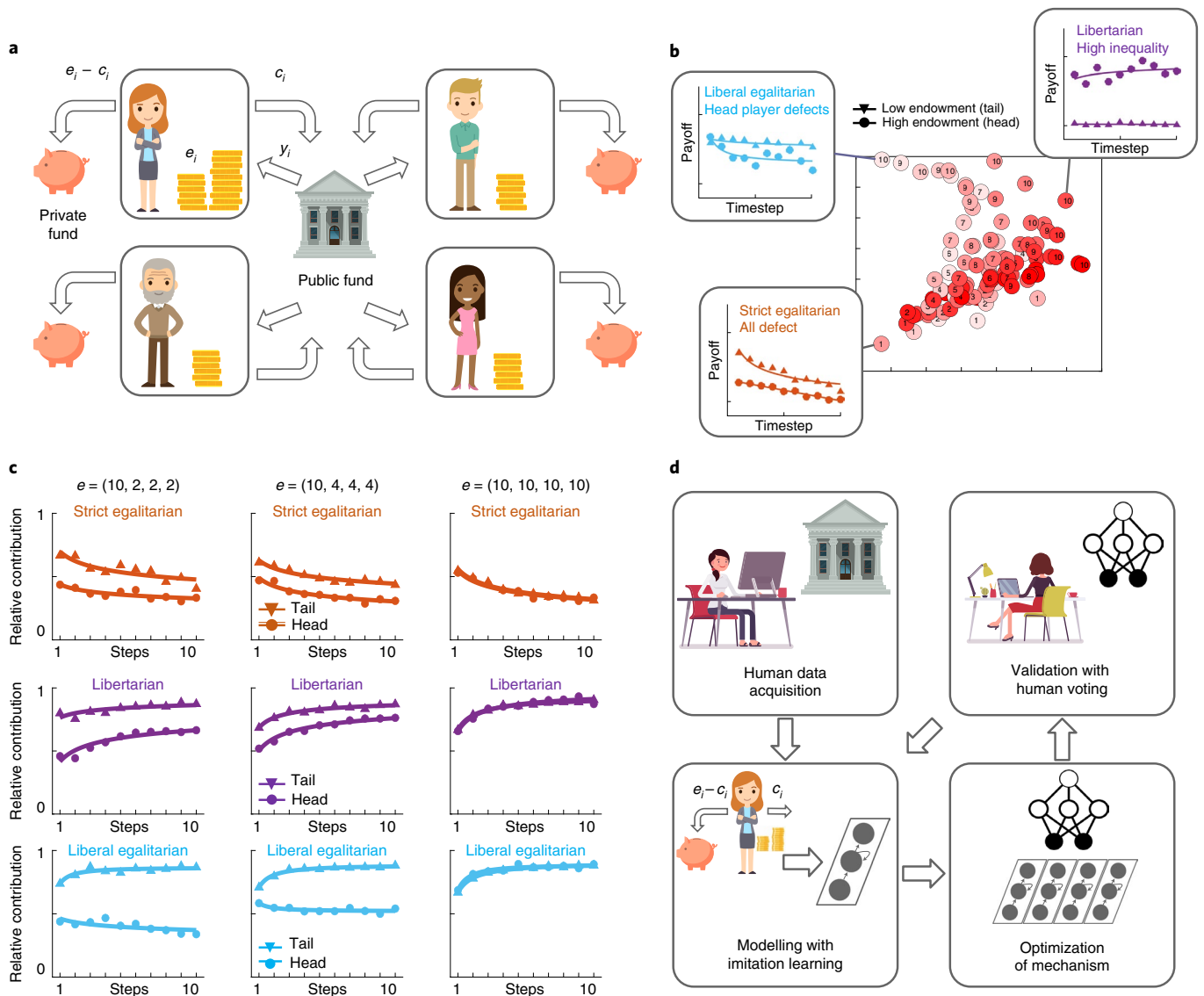
As a first rigorous test, we deploy Democratic AI to address a question that has defined the major axes of political agreement and division in modern times: when people act collectively to generate wealth, how should the proceeds be distributed?<sup>16–21</sup>. We asked a large group of humans to play an incentive-compatible online investment game that involved repeated decisions about whether to keep a monetary endowment or to share it with other players for potential collective benefit. We trained a deep RL agent to design a redistribution mechanism which shared funds back to players under both wealth equality and inequality. The mechanism it produced was ultimately preferred by the players in a majoritarian election.

## Results

We tested Democratic AI using a mechanism design problem based on an economic game. The game generalizes the linear public goods problem that has been extensively used to study human collective action<sup>22,23</sup> (Fig. 1a). In each of 10 rounds, each player  $i$  contributes an integer number  $c_i$  of coins to a public investment fund, drawing upon an endowment  $e_i$ , with the residual sum  $e_i - c_i$  remaining in a private account (endowments may vary across players, with one player receiving more than the others). Aggregated contributions over  $k=4$  players are scaled by a growth factor  $r=1.6$  (positive return on investment; this is equivalent to a marginal per capita return (MPCR) of 0.4). The public fund is paid back to players under a redistribution mechanism which specifies the fraction of total public investment that is returned to each player, conditional on their contribution and endowment. This game admits a continuum of mechanisms for redistribution popularly associated with opposing ends of the political spectrum<sup>19</sup>, in which returns variably depend on the contributions of self and others<sup>23</sup>.

**Experiment 1.** We illustrate the richness of the mechanism design problem in Exp. 1, in which we measured human contributions made under three canonical redistribution principles: strict egalitarian, libertarian and liberal egalitarian. Players ( $n=756$ ) were

<sup>1</sup>Deepmind, London, UK. <sup>2</sup>Department of Economics and Institute for Data Science and Artificial Intelligence, University of Exeter, Exeter, UK. <sup>3</sup>Gatsby Computational Neuroscience Unit, University College London, London, UK. <sup>4</sup>Department of Experimental Psychology, University of Oxford, Oxford, UK. <sup>5</sup>These authors contributed equally: Raphael Koster, Jan Balaguer. ✉e-mail: [csummerfield@deepmind.com](mailto:csummerfield@deepmind.com)



**Fig. 1 | Illustration of the game and Experiment 1.** **a**, Illustration of the setup of the investment game. **b**, The ideological manifold for endowment distribution (10, 2, 2, 2). The plot shows a visualization of a space of redistribution mechanisms defined by parameters  $w$  and  $v$  in two dimensions. Each red dot is a mechanism, and distances between dots conserve dissimilarities in the (average) relative payout to virtual players (both head and tail). Dot numbers denote bins of mechanism parameter  $w$  (1, lowest; 10, highest) and shading denotes bins of  $v$  (light, more relative; dark, more absolute). Inset, example payouts to head (circles) and tail (triangles) players under the canonical mechanisms used as baselines against which to test the AI. Under strict egalitarian, payouts decline to head and tail players. Under libertarian, there is great inequality between head and tail players. Under liberal egalitarian, the head player stops contributing, so payouts decline for both head and tail players. **c**, Average relative contributions (as a fraction of endowment) over 10 rounds ( $x$  axis) in Exp. 1 for three different initial endowment conditions. Under strict egalitarian redistribution, tail player (triangles) contributions are higher when initial endowments are lower, but head player (circles) contributions do not differ. Under libertarian, head player contributions increase with equality, but tail player contributions remain constant. Head player contributions increase strongly with endowment under liberal egalitarian. **d**, Illustration of our agent design pipeline.

assigned to groups of 4 players, with one head player who received 10 coins endowment and three tail players who received either 2, 4 or 10 coins (head/tail labels were nominal in the latter condition). Thus, endowments were unequal when tail players received less than 10 coins and equal when all players received 10 coins. Each group played multiple games each of 10 rounds, receiving the same endowment on each occasion, but experiencing each game under a different redistribution mechanism (see Supplementary Methods). Each redistribution mechanism determined the payout  $y_i$  received by player  $i$  as a different function of the public contribution of both self and others.

The strict egalitarian redistribution mechanism divides public funds equally among all players irrespective of their contributions<sup>24</sup>. It thus recreates the linear public goods game that, for  $r < k$ , is a social dilemma in that each individual benefits from withholding contributions and free riding on the largesse of other players<sup>23</sup>. Accordingly, contributions under this mechanism decline over time (effect of time on contributions for head player: 2 coins,  $F_{9,540} = 2.51$ ,  $P = 0.023$ ,  $\eta^2 = 0.017$ ; 4 coins,  $F_{9,540} = 5.5$ ,  $P < 0.001$ ,  $\eta^2 = 0.041$ ; 10 coins,  $F_{9,594} = 7.27$ ,  $P < 0.001$ ,  $\eta^2 = 0.056$ ; Fig. 1c, top), mirroring previously described results<sup>22,23</sup>. The libertarian mechanism<sup>21</sup> returns a payout to each player in proportion to their contribution

$y_i = r \times c_i$  such that  $c_i = e_i$  is a pareto-efficient Nash equilibrium. This mechanism effectively privatizes contributions and removes the social dilemma, encouraging players to increase their contributions (effect of time on contributions for head player: 2 coins,  $F_{9,297} = 9.96$ ,  $P < 0.001$ ,  $\eta^2 = 0.062$ ; 4 coins,  $F_{9,234} = 9.55$ ,  $P < 0.001$ ,  $\eta^2 = 0.073$ ; 10 coins,  $F_{9,270} = 12.56$ ,  $P < 0.001$ ,  $\eta^2 = 0.013$ ; Fig. 1c, middle) as observed previously<sup>25</sup> (note that while players receive detailed instructions about the game dynamics, they are obliged to learn about each mechanism from experience). Finally, liberal egalitarianism proposes that each player is accountable for their actions but not initial advantage, and so payout depends on the fraction of endowment that is contributed<sup>26</sup>. When payouts were relative to endowment-normalized contributions (liberal egalitarian), the tail players learned rapidly to contribute (2 coins,  $F_{9,720} = 4.79$ ,  $P < 0.001$ ,  $\eta^2 = 0.025$ ; 4 coins,  $F_{9,909} = 15.74$ ,  $P < 0.001$ ,  $\eta^2 = 0.043$ ) but the head player's contributions remained flat (2 coins,  $F_{9,234} = 1.84$ ,  $P = 0.139$ ,  $\eta^2 = 0.017$ ; 4 coins,  $F_{9,297} = 0.62$ ,  $P = 0.601$ ,  $\eta^2 = 0.004$ ), diminishing the availability of public funds.

Previous reports have suggested that heterogeneity of endowment or MPCR can influence contribution to the public fund<sup>27,28</sup>, especially when inequality is made salient to participants<sup>29</sup>. Here we observed a comparable phenomenon when we examined the contributions of the head player (who received 10 coins) as a function of the endowment received by tail players, which could be either equal or lower. Under strict egalitarian, the head player contributed the same irrespective of the endowment of others ( $F_{2,188} = 0.29$ ,  $P = 0.745$ ,  $\eta^2 = 0.003$ ), but under liberal egalitarian the head player was less prone to contribute when others were less well off ( $F_{2,377} = 14.10$ ,  $P < 0.001$ ,  $\eta^2 = 0.070$ ; the effect for libertarian was not significant:  $F_{2,280} = 2.82$ ,  $P = 0.069$ ,  $\eta^2 = 0.020$ ). Thus, productivity was dampened under conditions of greater inequality.

More generally, Exp. 1 highlights the challenge that the game poses for the mechanism designer: a redistribution scheme might be unpopular because it provokes a general collapse of contributions due to free riding, leads to unequal outcomes, or siphons funds away too aggressively from the wealthiest player, who then fails to provision the public fund. We thus asked whether an AI system could design a mechanism that humans preferred over these alternatives.

**Human-in-the-loop pipeline.** How then should the public funds be shared? The maximally popular policy could be one of these three canonical mechanisms or something else entirely. The size of the potential search space makes it hard to identify the preferred mechanism using traditional behavioural research methods. We thus developed a human-in-the-loop AI research pipeline to tackle this problem (Fig. 1d). First, we collected an initial sample of human data (Acquire) and used it to train 'virtual human players' which were recurrent neural networks that learned to imitate human behaviour during the game and voted according to the same principles as human players (Model; see Supplementary Fig. 1). This simulation step was necessary because training agents during online interaction with humans would have been prohibitively costly and time-consuming. Third, we optimized the mechanism design with deep RL, using policy gradient methods<sup>30</sup> to maximize the votes of virtual human players (Optimize). Fourth, we sampled a new group of humans, and pitted the RL-designed redistribution mechanisms against rival baselines in a series of head-to-head majoritarian elections. This new human data was then used to augment our player modelling process, which in turn improved optimization and led to potentially better mechanisms (Repeat). This pipeline builds on recent approaches that have used human data interactively to train artificial agents<sup>31–33</sup>. We iterated this procedure to obtain a mechanism that we call the Human Centred Redistribution Mechanism or HCRM, which is the major focus of the remainder of this report.

**The ideological manifold.** Before evaluating HCRM with a new group of human players, we used our research pipeline to determine which baseline mechanisms might pose the strongest competition. To achieve this, we generalized the three canonical baselines to produce a continuously parameterized space of redistribution mechanisms. We first assume that the fractional payout to each player  $y_i$  is composed of an absolute ( $y_i^{\text{abs}}$ ) and relative ( $y_i^{\text{rel}}$ ) component, which are combined via a mixing parameter  $v$ . These components are in turn given by contributions from both the focal and other players (mixing parameter  $w$ ; see below and Methods). In these baselines, the payout to player  $i$  is thus given by

$$y_i = v \left( y_i^{\text{rel}} \right) + (1 - v) \left( y_i^{\text{abs}} \right),$$

where the absolute component combines their own contribution  $c_i$  with the average of that from other players  $c_{-i}$  so that

$$y_i^{\text{abs}} = r [w (c_i) + (1 - w) (c_{-i})]$$

and the relative component is similarly determined by  $\rho_i = c_i/e_i$ , which is the ratio of contribution to endowment for player  $i$ :

$$y_i^{\text{rel}} = r \left( \frac{C}{P} \right) [w (\rho_i) + (1 - w) (\rho_{-i})],$$

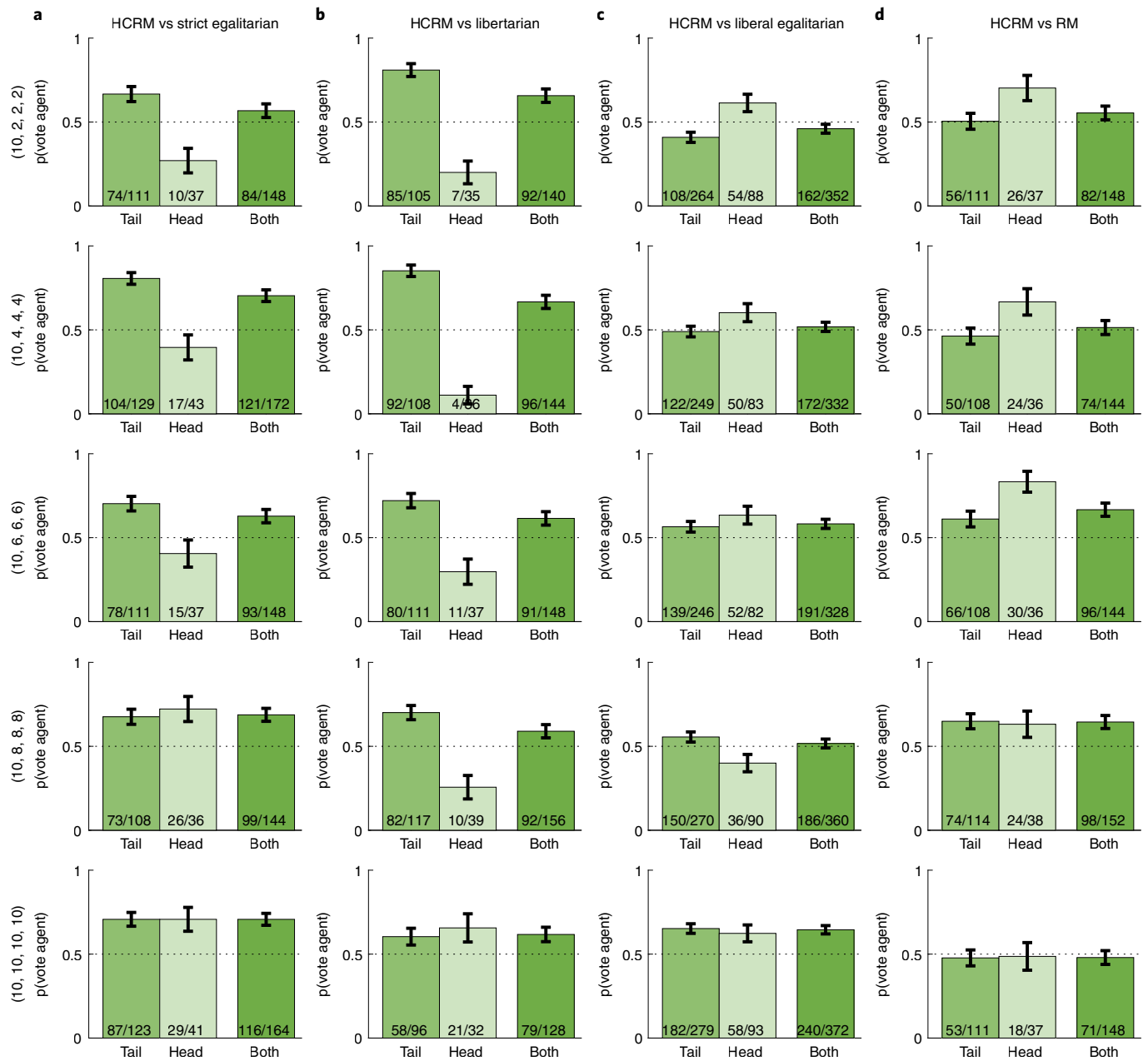
where  $c_{-i}$  and  $\rho_{-i}$  are respectively the average contributions and ratios from players other than  $i$ , and  $C$  and  $P$  are the sum of contributions and ratios across all players.

We call this space of baseline mechanisms defined by  $v$  and  $w$  the ideological manifold (Fig. 1b). We note that the three baseline mechanisms we have considered so far lie within this space: libertarian ( $w = 1$ ,  $v = 0$ ), liberal egalitarian ( $w = 1$ ,  $v = 1$ ) and strict egalitarian ( $w = 1/k$ ). We explore the properties of these mechanisms in more detail in Supplementary Fig. 10.

We sampled mechanisms from the ideological manifold and pitted them against one another exhaustively in a two-player tournament. Our goal was to identify the mechanism that maximized votes among the virtual human players (neural networks that had been trained to imitate real human behaviour). This exercise identified liberal egalitarian as the single Nash equilibrium of the two-player tournament and thus not only as the strongest competitor among the three canonical baselines but also among the entire space of mechanisms (Supplementary Table 2).

**Experiment 2.** Armed with this intuition, in Exp. 2a–c we evaluated the AI-designed HCRM against the three canonical baselines introduced above. Groups of 4 human participants ( $n = 2,508$ ) played successive incentive-compatible games of 10 rounds under two rival mechanisms, before voting for one that they preferred to play again (for additional payoff) in a final round. We randomized players into five endowment conditions, in which a head player received  $e_{\text{head}} = 10$  coins endowment and three tail players received the same  $e_{\text{tail}} \in \{2, 4, 6, 8, 10\}$  (once assigned, endowments remained constant throughout the game). We found that HCRM was more popular than all three baselines (Fig. 2a–c and Supplementary Table 1a; all  $P$  values obtained below are obtained from one-tailed binomial tests corrected for correlated responses within the group; see Methods), obtaining a total of 513/776 (66.2%) votes against strict egalitarian ( $P < 0.001$ ), 450/740 (60.8%) against libertarian ( $P < 0.001$ ) and 951/1,744 (54.5%) against liberal egalitarian ( $P < 0.001$ ).

Against strict egalitarian and libertarian, the AI-designed mechanism was also more popular under all five endowment distributions tested, ranging from full equality to the most unequal endowment condition ((10, 2, 2, 2) implies a Gini coefficient (a measure of wealth inequality) of 0.38, roughly equivalent to

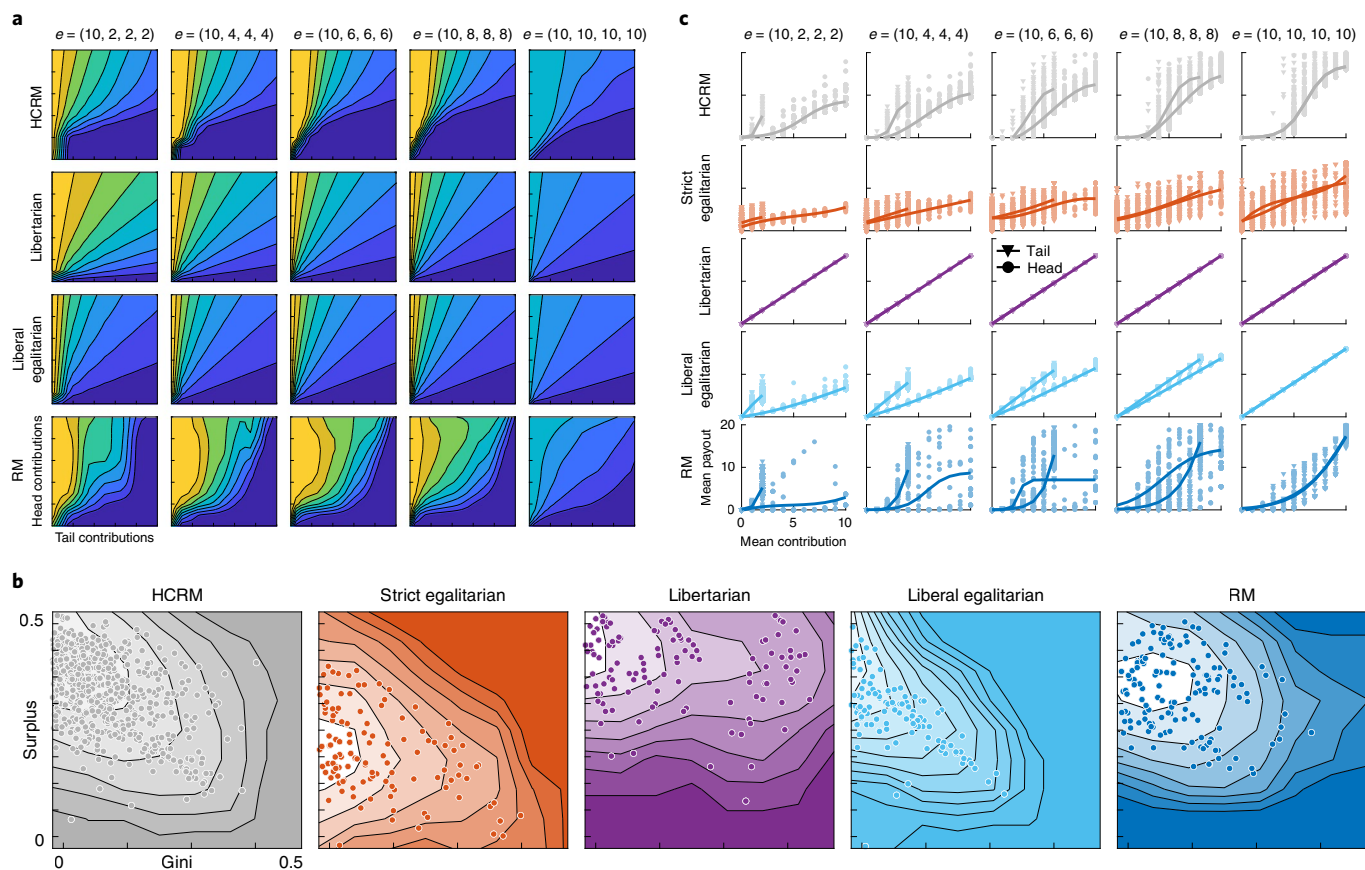


**Fig. 2 | Overall vote share by endowment and rival mechanism. a–d.** Vote share for the HCRM against the three canonical baselines (a–c) and the RM (d) for each endowment condition. The three bars show the average number of votes for the agent given by the tail players, the head player and all players. In all plots, bars show binomial standard error.

contemporary Russia). Across these conditions, its vote share ranged from 56.0% to 67.0% against egalitarian and from 57.5% to 66.7% against libertarian. Consistent with the data from our two-player tournament, liberal egalitarian proved popular among humans and thus more difficult to beat. Indeed, although HCRM was preferred overall, under full equality (64.5%,  $P < 0.001$ ) and under moderate inequality (endowments (10, 8, 8, 8) and (10, 6, 6, 6)) with a vote share of 54.5% ( $P < 0.006$ ), there was no reliable difference in voting preference between HCRM and liberal egalitarian under the most unequal conditions (HCRM vote share 47.4%,  $P = 0.897$ ), suggesting that liberal egalitarian redistribution offered an equally good alternative to HCRM under conditions of highest inequality.

Our AI-designed HCRM was trained by interacting with neural networks that imitated human behaviour. However, if our participants

are rational agents who learn to maximize their return over the course of each game, then it should be possible to solve the problem without recourse to human training data at all, by substituting our virtual human players for a new class of rational players that are trained to maximize their own expected return within the game (see Methods). Previous work has implied that successful human-centred mechanisms can be obtained in this fully multi-agent setting<sup>34,35</sup>. Alternatively, if modelling human cognitive biases is critical, then a system trained to maximize the votes of rational players may transfer more poorly back to human participants. In Exp. 3, we tested this by exposing a new group of human participants ( $n = 736$ ) to both the mechanism designed by HCRM and that proposed by a new rational mechanism (RM) that was trained with rational players but otherwise identical.



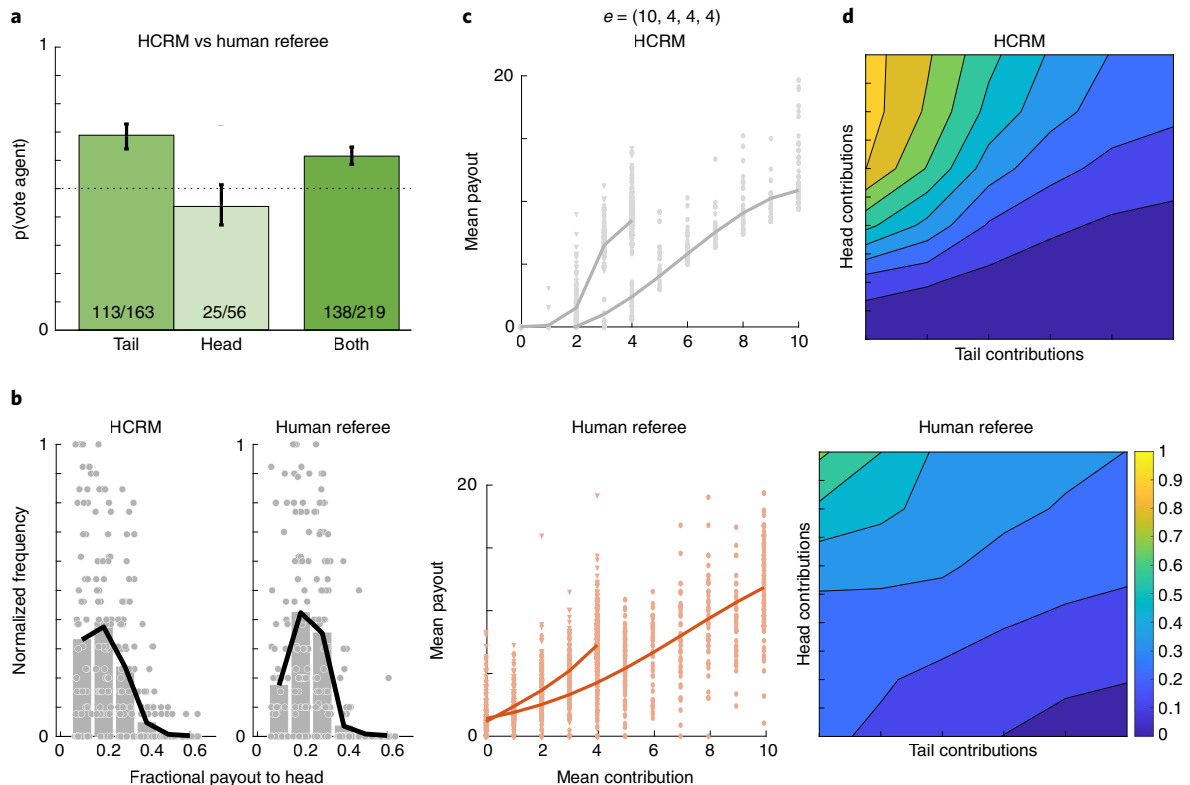
**Fig. 3 | Analysis of HCRM mechanism.** **a**, ‘Beach plots’ were created by simulating virtual players. They show revenue share allocated to head player as a function of contribution (relative to endowment) of head vs tail player. Warmer (colder) colours indicate relatively more funds redistributed to head (tail) player. **b**, Two-dimensional distributions of surplus (in log units) and Gini coefficient for each group under each mechanism. Lighter colours indicate higher density. Higher surplus implies greater productivity; lower Gini implies greater equality. Each dot is a game. **c**, Empirically observed relationship between contributions and payouts for each mechanism and endowment condition. Each dot is a head player or the average of tail players in a single game. Shading shows the density of dots. Lines are fit separately to head and tail players.

**Experiment 3.** Overall, 57.2% (421/736) of participants preferred HCRM over the RM ( $P < 0.001$ ; Fig. 2d). Interestingly, RM learned a radical policy under unequal endowments that neglected the head player and paid out principally to the tail players (Supplementary Fig. 3). Despite the favourable ratio of tail to head players, however, this was unsuccessful even in the most unequal endowment conditions, because the head player rapidly stopped contributing to the detriment of everyone (including the tail players), leading to an overall lower group surplus than for HCRM ( $t_{183} = 7.96$ ,  $P < 0.001$ ). In other words, the redistribution policy that humans prefer is neither one that shares out public funds equally, nor one that tries to speak only to the interests of a majority of less well-endowed players. One exception was under equal endowment, where HCRM and RM performed nearly identically (HCRM vote share 71/148 or 47.9%,  $P = 0.617$ ), implying that for the setting we explored, rational models may offer a good account of human behaviour when the initial conditions are fair. Together, however, these results imply that human data may be crucial when using AI for mechanism design.

The RL mechanism designer was not equipped with memory, hence HCRM is readily interpretable, that is, it can be transparently described as a two-dimensional surface that maps the relative contribution of head and tail players to their share of the proceeds (‘Beach plots’ in Fig. 3a). This allowed us to ask why the RL-designed mechanism is popular with human players. RL discovered a hybrid mechanism that eschewed traditionally proposed redistribution schemes that emphasize individual discretion over

resource allocation (libertarian) or collective equality (strict egalitarian). Pursuing a broadly liberal egalitarian policy, HCRM sought to reduce pre-existing income disparities by compensating players in proportion to their contribution relative to endowment. In other words, rather than simply maximizing efficiency, the mechanism was progressive: it promoted enfranchisement of those who began the game at a wealth disadvantage, at the expense of those with higher initial endowment. In doing so, it achieved a favourable trade-off between productivity (surplus) and equality (Gini coefficient) among rival mechanisms (Fig. 3b; see also Supplementary Fig. 3). However, unlike liberal egalitarian, it returned almost nothing to players unless they contribute approximately half their endowment (Fig. 3c). In other words, RL effectively discovers that humans facing social dilemmas prefer mechanisms that allow for sanction of free riders<sup>36</sup>. The agent thus learns a policy that is not readily assigned to a specific philosophy of distributive justice but creatively combines ideas from across the political spectrum.

**Experiment 4.** Finally, we asked whether trained and incentivized human players could have devised a mechanism that was as popular as HCRM. We first recruited 61 previous players and trained them over the course of about an hour to redistribute funds to virtual citizens with a view to maximizing votes (that is, the same training regime as our agent). Human referees earned £2 per vote. Focusing on the (10, 4, 4, 4, 4) condition, which was among those that HCRM found most challenging, we then recruited an additional set of new



**Fig. 4 | Results of human referee experiment.** **a**, Votes for the agent against trained humans. Bars show binomial standard error. **b**, Normalized frequencies of fractional payouts to the head player in 6 equally spaced bins. Each dot at each  $x$  position represents a single group ( $n = 86$ ); the black line and bars are the average. Note that the HCRM is more willing to offer low payout to the head player than human referees (leftwards skew for left-hand plot). **c**, Empirically observed relationship between contributions and payouts for each HCRM (top) and the human referee (bottom). Each dot is a head player (circle) or average of tail players (triangle) in a single game. The slope mapping contributions onto payouts for the tail player is shallower for human referees. **d**, Empirically observed relationship between payout and head/tail player contributions ('beach plot') for HCRM (top) and human referee (bottom) for (10, 4, 4, 4). The colour scale shows the fraction of the public fund allocated to the head player as a function of the head ( $y$  axis) and tail ( $x$  axis) player relative contributions.

human players ( $n = 244$ ) who played one game under HCRM and another with trained human referees, in counterbalanced order. These human players strongly preferred HCRM over the human referee (62.4% voted for HCRM,  $P < 0.001$ ). Interestingly, the human players were overall less prone to sanction the head player with low payouts (government  $\times$  payout sextile interaction,  $F_{2,128} = 5.541$ ,  $P < 0.005$ ,  $\eta^2 = 0.125$ ; Fig. 4b) and failed to reward the tail players sufficiently for contributing generously from the little they had (Fig. 4c, bottom) relative to HCRM (top). We show empirically derived beach plots for the human and algorithmic referees side by side in Fig. 4d; they imply that overall, human referees were less responsive to contributions when allocating payouts.

## Discussion

Together, these results thus demonstrate that an AI system can be trained to satisfy a democratic objective, by designing a mechanism that humans demonstrably prefer in an incentive-compatible economic game. Earlier studies have used voting to understand participants' preferences over contribution thresholds, or exclusion policies in the public goods game<sup>37–39</sup>, but here we used tools from AI research to learn a redistribution scheme from scratch. Our approach to value alignment relieves AI researchers — who may themselves be biased or are unrepresentative of the wider population — of the burden of choosing a domain-specific objective for optimization. Instead, we show that it is possible to harness for value alignment the same democratic tools for achieving consensus that

are used in the wider human society to elect representatives, decide public policy or make legal judgements.

Our research raises several questions, some of them theoretically challenging. One might ask whether it is a good idea to emphasize a democratic objective as a method for value alignment. Democratic AI potentially inherits from other democratic approaches a tendency to enfranchise the many at the expense of the few: the 'tyranny of the majority'<sup>40</sup>. This is particularly pertinent given the pressing concern that AI might be deployed in way that exacerbates existing patterns of bias, discrimination or unfairness in society<sup>41</sup>. In our investment game, we sampled endowment conditions to match plausible real-world income distributions, where the disadvantaged inevitably outnumber the advantaged; hence, for the specific question of distributive justice that we address, this problem is less acute. However, we acknowledge that if deployed as a general method, without further innovation, there does exist the possibility that (similar to real-world democratic systems) it could be used in a way that favours the preferences of a majority over a minority group. One potential solution would be to augment the cost function in ways that redress this issue, much as protections for minorities are often enshrined in law.

Another important point concerns the explainability of our AI-designed mechanism<sup>10</sup>. We deliberately hampered the mechanism designer by not equipping it with activation memory. This means that the mechanism it designed (HCRM) can be transparently described in just two dimensions (rather than, for example,

being a complicated nonlinear function of the choice history of different players). Although this level of complexity is greater than the human-generated theories of distributive justice that we use as baselines, it is still possible to verbalize. Encouraging a more interpretable mechanism has at least two advantages. First, it made the agent more transparent to the human players. In fact, in feedback questions (Supplementary Fig. 6), humans deemed the agent to be ‘more transparent and predictable’ than the alternative AI-designed mechanism (rational mechanism) and (perhaps incongruously) the strict egalitarian. Second, the lack of memory has implications for user privacy. Inputs to the agent were designed to be entirely ‘slot equivariant’, meaning that the mechanisms treated each player’s input independently of its ‘slot’ (whether it is player 1, 2, 3 or 4). The agent’s input pertained to the distribution of contributions rather than contributions from individuals themselves. Coupled with the lack of memory, this means that the agent is barred from tracking information about a particular player’s history of contributions within the game.

Our AI system designed a mechanism for redistribution that was more popular than that implemented by human players. This is especially interesting because unlike our agent, human referees could integrate information over multiple timesteps to reward or sanction players on the basis of their past behaviour. However, on average the human-invented redistribution policy tended to reward the tail player insufficiently for making relatively large contributions (from their smaller endowment) to the public purse and was less popular than that discovered by HCRM. Humans received lower volumes of training data than HCRM, but presumably enjoyed a lifetime of experience with social situations that involved fair and unfair distribution, so we think they represent a strong baseline, and a proof of concept for AI mechanism design.

One remaining open question is whether people will trust AI systems to design mechanisms in place of humans. Had they known the identities of referees, players might have preferred human over agent referees simply for this reason. However, it is also true that people often trust AI systems when tasks are perceived to be too complex for human actors<sup>42</sup>. We hope that future studies will address this question. Another question concerns whether participants would have responded differently if the mechanisms had been explained to them verbally, rather than learned by experience. A long literature has suggested that people sometimes behave differently when mechanisms are ‘by description’ rather than ‘by experience’, especially for risky choices<sup>43</sup>. However, AI-designed mechanisms may not always be verbalizable, and it seems probable that behaviours observed in such case may depend on exactly the choice of description adopted by the researcher.

Finally, we emphasize that our results do not imply support for a form of ‘AI government’, whereby autonomous agents make policy decisions without human intervention<sup>44,45</sup>. We see Democratic AI as a research methodology for designing potentially beneficial mechanisms, not a recipe for deploying AI in the public sphere. This follows a tradition in the study of technocratic political apparatus that distinguishes between policy development and policy implementation, with the latter remaining in the hands of elected (human) representatives<sup>46</sup>. We hope that further development of the method will furnish tools helpful for addressing real-world problems in a truly human-aligned fashion.

## Methods

**Participants.** The study was approved by HuBREC (Human Behavioural Research Ethics Committee), which is a research ethics committee run within Deepmind but staffed/chaired by academics from outside the company. Participants were recruited over an approximately 8-month period from two different crowdsourcing platforms. All participants gave informed consent to participate in the experiment. The task was advertised to users located in the UK and the USA. We did not record identifiers or personal information from participants. Participants who accepted the Human Intelligence Task (HIT) received a link that led them to a game lobby

in which they were grouped with three other players. Groups of four players participated in the game live and in interaction with one another. When a response was required, participants had a fixed duration in which to respond (2 min for all screens except voting, which was 4 min), accompanied by a timer that signalled how much time they had remaining. The game advanced only when the player who was slowest to respond had completed the round. Players who timed out were given a warning. Players who timed out twice were removed from the game and replaced with a randomly responding bot (games with missing data were excluded from the analysis). The game took approximately 20–30 min and participants were paid up to £8, consisting of a base and a bonus payment. The precise conversion rate of points earned in-game (return in coins) to the bonus paid at the end of the study varied inversely with the sum of endowments over players. This way stakes were on average equated across games. Data collection and analysis were not performed blind to the conditions of the experiments.

A total of  $n = 4,776$  participants took part in Exp. 1–3. The pilot data that were used for training the agent consisted of a further ~4,000 datasets (including some partial datasets where participants timed out). Exclusion lists were used to prevent participants from rejoining the experiments multiple times (however, as the overall data were collected over several months on two platforms, and we did not collect identifiers, it was impossible to be absolutely sure that all participants are unique).

**Investment game.** All participants in Exp. 1–3 played 34 rounds of an investment game (3 blocks of 10 rounds and 1 ‘bonus’ block of 4 rounds). On each round, each player was allocated an endowment of 2, 4, 6, 8 or 10 coins (2, 4 or 10 coins in Exp. 1) depending on the endowment condition to which they were allocated, and whether they were designated the ‘head’ or ‘tail’ player, all of which was entirely random (and unrelated, for example, to the order in which they joined the game). In all endowment conditions, there was a single head player who received 10 coins and three tail players who all received either 2, 4, 6, 8 or 10 coins (in the ‘equal’ endowment condition, the distinction between head and tail players is nominal only). Players received their endowment at the start of each round. Each player’s endowment remained the same across all 34 rounds (and they were instructed that this would be the case).

In every round of every block, each player  $i$  privately chose to divide an integer number of coins (their endowment  $e_i$ ) between a ‘project’ and a ‘private account’ (the contributions made to the project are denoted  $c_i$ ). No player could see the others’ choices at this stage. The ‘project’ was a public fund that received a return on investment (was multiplied by a common productivity factor  $r = 1.6$ ) and was then shared between participants according to some redistribution scheme, allocating a payout  $y_i$  to player  $i$  (see below). Coins allocated to the ‘private account’ were simply retained by participants (with no return on investment). The total return to each player on each round was thus their payout plus endowment minus contribution  $y_i + e_i - c_i$ .

In the first block of all experiments, participants played 10 ‘tutorial’ rounds with no referee. This meant that funds allocated to the project were distributed equally among all players and there was no further redistribution (see below). In making this choice, we assume that equal redistribution is a ‘default’ position, which the referees subsequently adjust. There were several reasons for this choice, including the importance of illustrating to players that a social dilemma could arise. However, to ensure that this did not influence our results, we ran additional controls (not reported) in which we added a block of libertarian after the ‘no referee’ block. For the conditions we checked, performance was almost identical to under the default referee, and our agent continued to be preferred in all cases.

In blocks 2 and 3, participants played 10 rounds with a referee (or mechanism; one mechanism for each block). The referee(s) redistributed project funds among players according to a specified mechanism, without creating or destroying wealth. The two rival mechanisms were encountered in counterbalanced order. After block 3, participants voted for the mechanism that they preferred. They knew that they would be making this vote (and what it would entail) from the end of block 1, before experiencing the mechanisms. In block 4, the probability of re-experiencing mechanism A (or B) was exactly equal to the fraction of votes that A (or B) received from the 4 players. The choice was thus deterministic if all players voted the same way, and there was no opportunity to vote strategically. Participants then answered seven debriefing questions (see below). Finally, they experienced four rounds of the chosen mechanism (block 4) and proceeded to a bonus screen where they were thanked and informed of their total earnings. Only data from blocks 2 and 3 were included in the analysis of Exp. 2–4 (see below for Exp. 1). We report numbers of participants who chose to vote for either mechanism using binomial tests.

**Detail of experiments.** We describe 3 experiments in the main text. All experiments had the same form but the way we present the data is slightly different for Exp. 1 relative to Exp. 2 and Exp. 3.

In Exp. 2 ( $n = 2,508$ ), players experienced the HCRM and either strict egalitarian, libertarian or liberal egalitarian mechanisms in randomized groups (we provide details about all mechanisms below). The order of encounter of the different mechanisms was counterbalanced over blocks 2 and 3. Under the strict egalitarian mechanism, the referee effectively took no action, so that the original earnings and ‘earnings after referee’s actions’ screens looked the same (apart from

bar colour). Under libertarian, the mechanism returned to each player a sum that was 1.6× their contribution. We decided to recruit a larger cohort for liberal egalitarian because simulation data (Supplementary Table 2) suggested that this was the highest performing baseline and potentially more data would be required to draw a reliable conclusion about which was preferred (target participant numbers were decided beforehand, and we did not use optional stopping). In Exp. 3 ( $n = 736$ ), players experienced HCRM and RM (trained with rational players; see below for description of these players, and main text for more details about the RM) in counterbalanced order over blocks 2 and 3.

The data described as ‘Experiment 1’ came from a study in which an earlier version of the HCRM competed against libertarian and liberal egalitarian. We present data only from these two baselines (not the earlier version of the agent). The data from strict egalitarian are taken from block 1 (in which there was no referee). Thus, the conditions under which these data were collected are the same as Exp. 2, except that (1) the order of the mechanisms was not counterbalanced, (2) the rival AI-designed mechanism was slightly different and (3) we did not report voting data from this experiment. Our goal here is to illustrate how contributions vary under different mechanisms; in fact, a near-identical pattern of results was replicated in Exp. 2. At this stage we only used three endowment conditions: (10, 2, 2, 2), (10, 4, 4, 4) and equality ((10, 10, 10, 10)).

**Data pre-processing and analysis.** Our main analyses focus on data from Exp. 1–3 ( $n = 4,776$  total). We have made this large dataset freely available at [https://github.com/deepmind/hcmd\\_dai](https://github.com/deepmind/hcmd_dai), along with code for recreating key figures.

In Exp. 1, we plot contributions as a function of mechanism and endowment as described above (Fig. 1c). In Exp. 2–3, data were analysed from blocks 2 and 3, that is from the two blocks of 10 trials in which participants played the game with a referee that was either the HCRM or a rival baseline. We logged or computed various per-block, per-round metrics for each individual player, including (absolute) contributions ( $c_i$ ), relative contributions ( $c_i/e_i$ ), (absolute) payouts ( $y_i$ ), relative payouts ( $y_i/e_i$ ) and return ( $e_i - c_i + y_i$ ), as well as some group-level game metrics including Gini coefficient and surplus (sum of returns/sum of initial endowments over players and rounds). We show contributions and payouts over time for head and tail players under each mechanism in Supplementary Figs. 3 and 4.

Our voting data consisted of 4 binary votes per group  $g$ , which were either for HCRM or a baseline. We performed a group-level permutation test to assess statistical significance. Our permuted data randomly flipped voting preferences but preserved the covariance among votes within a group (note that statistics from a naive binomial test might be inflated as they are not independently conditioned on participants’ shared experiences of the mechanism). The lowest possible  $P$  value obtainable given our 10,000 shuffles is  $P < 0.0001$  (but we do not report  $P$  values lower than 1/1,000 to meet journal style requirements).

**Overview of the mechanism training method.** Our approach to AI mechanism design consisted of three steps: (1) we trained virtual human players with supervised learning to imitate human play in the investment game using an existing dataset, which we describe as ‘pilot’ data for the main experiments presented here; (2) we trained a vote-maximizing mechanism using a deep RL agent interacting with these virtual players; and (3) we evaluated the mechanism by deploying it with new unseen human participants, together with comparison baselines (see above). This last step yielded new data, which could be used to repeat the steps above, and refine both the virtual players and the mechanism. In the next section, we describe these steps in detail.

**Training virtual players using imitation learning.** We drew upon data from previous pilot experiments (see ‘pilot testing’, below) to train virtual human players using imitation learning. All data were collected from human participants playing the 10-round investment game (described above) under a variety of different mechanisms and varying endowment conditions (not equally distributed). While generally we included only games in which all players finished the experiment, the training data included data from a handful of pilot games where one or more players dropped out. However, the modelling excluded the responses from the players that had been replaced with randomly responding bots. For more details, see ‘pilot testing’ below.

We used imitation learning in which virtual players were trained to imitate human play. Virtual human players were deep neural networks. Each network was a simulation of a single player, which received information about all players’ contributions and receipts on the current trial (similar to real human players), and was trained to predict the focal player’s contributions on the next trial. As they were equipped with recurrent memory (long short term-memory (LSTM) units) (37), the networks could potentially learn to use the trial history going back to the start of the game to make this prediction.

The network received the following information as input on each step: each player’s endowment (4 inputs); each player’s previous contribution (4 inputs); each player’s previous contribution relative to endowment (4 inputs); and each player’s payout (4 inputs). Payouts, endowments and contributions were divided by 10 to lie in approximately the same range as relative contributions. These inputs were fed to a linear layer with output size of 64 and tanh nonlinearities, followed by an LSTM

with hidden size 16. The LSTM outputs to a final linear layer of size 11 whose outputs coded the unnormalized log-probabilities of a categorical distribution corresponding to the probability of contributing 0, 1, ..., 9 or 10 coins. We masked those outputs corresponding to contributions in excess of the endowment allocated to the focal player.

We trained this architecture with back-propagation through time (38) to minimize the cross-entropy between the predictions and the actual contributions, regularized with two additional terms: the entropy of the prediction (with weight 0.1); and the L2 loss on the parameters of the network (with weight 0.00001). The model was implemented in TensorFlow 1 and the architecture was optimized using Adam (39) with learning rate 0.0004 and parameters beta1 0.9, beta2 0.999 and epsilon  $1 \times 10^{-8}$ . We trained the model by performing 30,000 updates with mini batches of size 512. Training took <6 h without the use of accelerators.

The virtual human player networks were evaluated on a separate hold-out dataset consisting of the contributions of a new group of human players ( $n = 384$ ) that resembled as closely as possible the final conditions under which we expected to evaluate the HCRM. We swept over network hyperparameters (including layer widths, number of LSTM units, learning rate and regularization types) to minimize validation loss.

**Voting model.** We learned in piloting that human players’ votes are most strongly predicted by the relative payouts ( $y_i/e_i$ ) they receive under one mechanism or another. We thus used this variable as the basis for virtual player voting. Each virtual player’s probability of voting for mechanism A rather than the rival mechanism B was  $p(A) = \Phi[\text{rpay}^A - \text{rpay}^B]$  where  $\text{rpay}^M$  is the sum of relative payouts obtained under mechanism M and  $\Phi[\cdot]$  is a logistic function with slope  $s$ . We set  $s$  to be 1.4, but similar results were obtained (in terms of mechanism policy, see below) under a wide range of values (see below).

**Mechanism designer: problem definition and setup.** We call the neural network used to design the mechanism the mechanism designer and use the term human-compatible mechanism (HCRM) to refer to the mechanism it designs, which was obtained only after training has converged. It used RL to learn a function that mapped observations (game states generated through interaction with virtual players) onto redistribution weights (a variable that determines which player gets what fraction of the project fund). We chose a Graph Network-based architecture that is equivariant to permutation in the ordering of participants and trained it on simulated 10-round investment games, with the goal of maximizing the cumulative voting probabilities of the virtual players against a carefully chosen alternative mechanism (see below). After training for 10,000 steps, the network parameters were frozen, and the function was exported in a way that allowed ready implementation in a human testing setting.

**Network architecture.** Inputs to the network were the endowments, contributions and relative contributions (that is, contribution–endowment ratios) for the current round, for all four participants (12 inputs per round). The network’s output was 4 numbers that were passed through a softmax function (that is, so that they are positive and sum to 1) to generate the redistribution weights for each player. When we state that the network has no memory, we mean that (1) it does not receive historical information about contributions or payouts as inputs; and (2) it does not have recurrence, that is, network states are not passed between timesteps. Note, however, that the mechanism could infer implicitly the number of rounds if human/virtual player policies vary their contributions between different timepoints within the block.

We organized the network’s observations in a fully connected directed graph  $(u, V, E)$  where each player was represented as a vertex  $v_k \in V$ . Directed edges  $e_{s,r}$  connecting  $v_s$  and  $v_r$  had empty initial attributes, and the input global attribute vector  $u$  was empty. Computations in Graph Networks start by updating the edge attributes, followed by the node attributes and finally global attributes. In particular, directed edge attributes were updated with a function  $\varphi_e$  of the input edge attribute, the sender and receiver vertex attributes, and the global attributes vector:  $e_{s,r}^t = \varphi_e(e_{s,r}, v_s, v_r, u)$ ; vertex attributes were updated as a function  $\varphi_v$  of the input vertex attributes, the sum of all updated edge attributes that connect into  $v_s$ , and the global attributes vector:  $v_s^t = \varphi_v(\sum_{e_{s,r}^t} v_r, u)$ ; finally, the global attributes vector was updated with a function of the input global attributes, and the sum of all updated edges and vertices:  $u^t = \varphi_u(\sum_{s,r} e_{s,r}^t, \sum_k v_k^t, u)$ . We note that the same functions  $\varphi_e, \varphi_v$  are used to update all edges and nodes in a graph, and that both the input and output of Graph Networks are directed graphs, so these modules can be used in sequence.

The mechanism designer’s policy network architecture consisted of two Graph Networks (GNs) that processed the observation in sequence. In the first GN, we implemented all of  $\varphi_e, \varphi_v$  and  $\varphi_u$  as distinct linear layers with 32 output units and tanh activation functions. In the second GN, we implemented  $\varphi_e$  as a linear layer with 32 output units and tanh activation function, and  $\varphi_v$  as a linear layer with a single output unit. We normalized the vertex outputs with a softmax across players, thus obtaining the redistribution weights;  $\varphi_u$  was ignored in the second GN.

**Training algorithm.** To train the mechanism, we used an estimator drawn from a framework known as Stochastic Computation Graphs (SCG)<sup>30</sup>, which approximates



the gradients of the vote-maximization objectives with respect to the network's policy parameters. We trained the mechanism designer iteratively with 10,000 updates using the RMSProp algorithm to optimize the policy, with the following parameters: learning rate 0.0004; epsilon  $1 \times 10^{-5}$ ; decay 0.99; and no momentum. On every update, we simulated two batches of 512 games with 10 rounds per game. We divided the batches in groups of 64 episodes with consistent endowments, leading to 8 possible endowments: the head player received 10 coins, and the tail players received 2, 3, 4, 5, 6, 7, 8 or 10 coins (using a broader range of tail player endowments helped avoid overfitting).

On every round, the game unfolded as described above for human players, except that the contributions were dictated by the virtual human players. In the first block, the redistribution policy was decided by the mechanism designer under training, and in the other it was played by an alternative planner (which was the winner of the metagame, defined by  $w = 1, v = 1$ ; see section on ideological manifold). We paired episodes from these two batches to obtain 2,048 votes (512 pairs of episodes, 4 players) given our model of human voting. The objective that the HCRM aimed to maximize was the sum of votes across players, averaged across episodes.

Note that during training of the mechanism, we did not feed in the human data to predict player contributions (that is, 'teacher forcing'). Furthermore, the payouts observed by the virtual players depended on mechanism policies that may lie outside of the human data, thus requiring the virtual players to generalize beyond the training dataset.

Having defined the observations as well as the objective that we wished to maximize, we estimated the policy gradient, that is, the gradient of the objective (the average number of votes) with respect to the policy parameters (the weights of the graph network) by turning to the SCG framework. We note here that most of the computation in the investment game is differentiable (including the policy as implemented by the HCRM), with the virtual human players policies, whose action space is discrete, being the only exception. The SCG framework generalizes the policy gradient theorem and allowed us to obtain a low-variance estimator of the policy gradient by auto-differentiating through the environment and mechanism policy, while compensating for the non-differentiable operations (the discrete contributions of the players). The surrogate objective for the policy gradient was as follows:

$$S = J + \perp(J) \times \sum_i \Sigma_{t=2}^{10} \log(\perp(c_t^i)),$$

where  $S$  is the surrogate objective,  $J$  is the objective we wish to maximize per episode (the expected number of votes) and  $\perp$  is the stop-gradient operation. Note that for the second term, the gradient can only flow through the parameterization of the log-probability of the player's policy. Note also that the contributions of the first round are removed from the equation since they do not depend on the mechanism's parameters. In practice, additionally, we chose to mean-center  $J$  within a batch because this is known to reduce the variance of the gradient estimator.

In Supplementary Information, we include further details that provide (1) a detailed description and illustration of the game, (2) the voting procedure, (3) debriefing, (4) determinants of voting analysis, (5) beach plots, (6) the ideological manifold, (7) rational players, (8) the metagame, (9) pilot testing, (10) human referee experiments and (11) theoretical analysis of the game.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

All human data is available at [https://github.com/deepmind/hcmd\\_dai](https://github.com/deepmind/hcmd_dai).

## Code availability

Code for reproducing figures is available at [https://github.com/deepmind/hcmd\\_dai](https://github.com/deepmind/hcmd_dai).

Received: 9 December 2021; Accepted: 10 May 2022;

Published online: 04 July 2022

## References

- Taddeo, M. & Floridi, L. How AI can be a force for good. *Science* **361**, 751–752 (2018).
- Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
- Tomasev, N. et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* **572**, 116–119 (2019).
- Gomes, C. et al. Computational sustainability: computing for a better world and a sustainable future. *Commun. ACM* **62**, 56–65 (2019).
- Abebe, R. & Goldner, K. Mechanism design for social good. *AI Matters* **4**, 27–34 (2018).
- Roth, A. E. *Who Gets What —and Why: The New Economics of Matchmaking and Market Design* (Houghton Mifflin Harcourt, 2016).
- Hurwicz, L. & Reiter, S. *Designing Economic Mechanisms* (Cambridge Univ. Press, 2006). <https://doi.org/10.1017/CBO9780511754258>
- Gabriel, I. Artificial intelligence, values and alignment. *Minds Mach.* **30**, 411–437 (2020).
- Rahwan, I. Society-in-the-loop: programming the algorithmic social contract. *Ethics Inf. Technol.* **20**, 5–14 (2018).
- Doshi-Velez, F. & Kim, B. Towards a rigorous science of interpretable machine learning. Preprint at <https://arxiv.org/abs/1702.08608> (2017).
- Lee, M. K. et al. WeBuildAI: participatory framework for algorithmic governance. *Proc. ACM Hum. Comput. Interact.* **3**, 1–35 (2019).
- Conitzer, V., Sinott-Armstrong, W., Borg, J. S. & Deng, Y. *Moral Decision Making Frameworks for Artificial Intelligence* (AAAI Press, 2017).
- Dafoe, A. et al. Open problems in cooperative AI. Preprint at <https://arxiv.org/abs/2012.08630> (2020).
- Christiano, P. et al. Deep reinforcement learning from human preferences. Preprint at <https://arxiv.org/abs/1706.03741> (2017).
- Noothigattu, R. et al. A voting-based system for ethical decision making. Preprint at <https://doi.org/10.48550/arXiv.1709.06692> (2018).
- Olson, M. *The Logic of Collective Action: Public Goods and the Theory of Groups* (Harvard Univ. Press, 2003).
- Ostrom, E. *Governing the Commons: the Evolution of Institutions for Collective Action* (Cambridge Univ. Press, 1991).
- Rawls, J. *A Theory of Justice* (Oxford Univ. Press, 2000).
- von Platz, J. *Theories of Distributive Justice: Who Gets What and Why* (Routledge, 2020).
- Dworczak, P., Kominers, S. D. & Akbarpour, M. *Redistribution Through Markets* Becker Friedman Institute for Research in Economics Working Paper No. 2018-16 (SSRN, 2018); <https://doi.org/10.2139/ssrn.3143887>
- Nozick, R. *Anarchy, State, and Utopia* (Basic Books, 1974).
- Isaac, R. M., McCue, K. F. & Plott, C. R. Public goods provision in an experimental environment. *J. Public Econ.* **26**, 51–74 (1985).
- Fehr, E. & Schurtenberger, I. Normative foundations of human cooperation. *Nat. Hum. Behav.* **2**, 458–468 (2018).
- Nielsen, K. Radical egalitarian justice: justice as equality. *Soc. Theory Pract.* **5**, 209–226 (1979).
- Falkinger, J., Fehr, E., Gächter, S. & Winter-Ebmer, R. A simple mechanism for the efficient provision of public goods: experimental evidence. *Am. Econ. Rev.* **90**, 247–264 (2000).
- Dworkin, R. *Sovereign Virtue: The Theory and Practice of Equality* (Harvard Univ. Press, 2002).
- Buckley, E. & Croson, R. Income and wealth heterogeneity in the voluntary provision of linear public goods. *J. Public Econ.* **90**, 935–955 (2006).
- Cherry, T. L., Kroll, S. & Shogren, J. F. The impact of endowment heterogeneity and origin on public good contributions: evidence from the lab. *J. Econ. Behav. Organ.* **57**, 357–365 (2005).
- Brañas-Garza, P., Molis, E. & Neyse, L. Exposure to inequality may cause under-provision of public goods: experimental evidence. *J. Behav. Exp. Econ.* **92**, 101679 (2021).
- Schulman, J., Heess, N., Weber, T. & Abbeel, P. Gradient estimation using stochastic computation graphs. *NeurIPS* **28**, 3528–3536 (2015).
- Carroll, M. et al. On the utility of learning about humans for human-AI coordination. Preprint at <https://doi.org/10.48550/arXiv.1910.05789> (2020).
- Dezfooli, A., Nock, R. & Dayan, P. Adversarial vulnerabilities of human decision-making. *Proc. Natl Acad. Sci. USA* **117**, 29221–29228 (2020).
- Abramson, J. et al. Imitating interactive intelligence. Preprint at <https://doi.org/10.48550/arXiv.2012.05672> (2021).
- Strouse, D. J., McKee, K., Botvinick, M. M., Hughes, E. & Everett, R. Collaborating with Humans without Human Data. *NeurIPS* **34**, 14502–14515 (2021).
- Zheng, S. et al. The AI economist: improving equality and productivity with AI-driven tax policies. Preprint at <https://doi.org/10.48550/arXiv.2004.13332> (2020).
- Gurerk, O. The competitive advantage of sanctioning institutions. *Science* **312**, 108–111 (2006).
- Dannenberg, A., Haita-Falah, C. & Zitzelsberger, S. Voting on the threat of exclusion in a public goods experiment. *Exp. Econ.* **23**, 84–109 (2020).
- Aumann, R. J., Kurz, M. & Neyman, A. Voting for public goods. *Rev. Econ. Stud.* **50**, 693 (1983).
- Tyran, J.-R., Rauchdobler, J. & Sausgruber, R. Voting on thresholds for public goods: experimental evidence. *FinanzArchiv* **64**, 34 (2010).
- Beahm, D. L. *Conceptions of and Corrections to Majoritarian Tyranny* (Lexington Books, 2002).
- Mohamed, S., Png, M.-T. & Isaac, W. Decolonial AI: decolonial theory as sociotechnical foresight in artificial intelligence. *Phil. Technol.* **33**, 659–684 (2020).
- Bogert, E., Schechter, A. & Watson, R. T. Humans rely more on algorithms than social influence as a task becomes more difficult. *Sci. Rep.* **11**, 8028 (2021).
- Hertwig, R. & Erev, I. The description-experience gap in risky choice. *Trends Cogn. Sci.* **13**, 517–523 (2009).

44. Katzenbach, C. & Ulbricht, L. Algorithmic governance. *Internet Policy Rev.* <https://doi.org/10.14763/2019.4.1424> (2019).
45. Sætra, H. S. A shallow defence of a technocracy of artificial intelligence: examining the political harms of algorithmic governance in the domain of government. *Technol. Soc.* **62**, 101283 (2020).
46. Peters, B. G. *The Politics of Bureaucracy: an Introduction to Comparative Public Administration* (Routledge, 2010).

### Acknowledgements

We thank M. Bakker, M. Chadwick, H. Sheahan, K. Kuvakcuoglu, D. Hassabis and L. Weidinger for comments on an earlier version of this paper; and C. Ahlheim, J. Merel, D. Fritz and M. Patel for technical assistance and/or helpful suggestions. We thank C. Hilbe, J. Jamison and S. Terry for feedback on potential mechanisms. The authors received no specific funding for this work.

### Author contributions

R.K., J.B., A.W., A.T., C.S., M.B., O.H. and T.Z. contributed to conceptualization, formal analysis, investigation and writing of the paper. L.C.-G. and P.T. administered the project. C.S. and M.B. supervised the project. D.W. provided software and resources.

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41562-022-01383-x>.

**Correspondence and requests for materials** should be addressed to Christopher Summerfield.

**Peer review information** *Nature Human Behaviour* thanks Pablo Branas-Garza, The Anh Han and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022

## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | n/a                                 | Confirmed  |
|-------------------------------------|--|
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided<br><i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of all covariates tested   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons  |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. $F$ , $t$ , $r$ ) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted<br><i>Give <math>P</math> values as exact values whenever suitable.</i>                            |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's $d$ , Pearson's $r$ ), indicating how they were calculated  |

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection

Data analysis

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Human data will be made freely available at [https://github.com/deepmind/hcmd\\_dai](https://github.com/deepmind/hcmd_dai)

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences       Behavioural & social sciences       Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The study involved a quantitative experimental design.
Research sample	The sample consisted of participants on well-known crowdsourcing platforms and no demographic data was recorded. Estimates of sample sizes were created by simulations using neural networks.
Sampling strategy	The sampling strategy was convenience sampling from the crowdsourcing platforms. Sample sizes were determined by simulations based on neural networks.
Data collection	Participants performed the task computerized over the internet. The researcher was not blind to the experimental conditions but did not interact with the participants prior to them doing the task.
Timing	The data was collected between the 7th of November 2019 to the 13th of August 2021.
Data exclusions	No participants were excluded from the data.
Non-participation	Of the 4804 participants in the main experiment, 48 did not submit a vote (for unknown reasons). These were excluded from the analysis. An unknown number of participants voluntarily aborted participation in the task and were excluded before all analyses.
Randomization	Allocation was fully random.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

### Methods

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	We did not collect demographic information of participants.
Recruitment	Participants were self-selected by being users of popular crowdsourcing platforms in the UK and USA. The nature of the sample, and its effect on our behavioral results are unknown. However, our experiment makes no claims about a specific effect in relation to any population, but rather demonstrated a modelling pipeline that can effectively capture the behavior of an arbitrary subpopulation.
Ethics oversight	The study was approved by HuBREC (Human Behavioral Research Ethics Committee), which is a research ethics committee run within Deepmind but staffed/chaired by academics from outside the company.

Note that full information on the approval of the study protocol must also be provided in the manuscript.