nature human behaviour

Registered Report

https://doi.org/10.1038/s41562-026-02562-w

Public perceptions of polygenic testing and embryo selection for non-medical traits

Received: 30 January 2025

Accepted: 22 July 2026

Published online: 26 August 2026

Check for updates

Edmond Awad[1,2,3] ✉, Clara Colombatto[4], Joanna Demaree-Cotton[1], Brian D. Earp[1,5], Jim A. C. Everett[6], Peng Liu[5,7], G. Owen Schaefer[5], Ilina Singh[2], Dominic Wilkinson[1,5,8] ✉ & Julian Savulescu[1,5,9] ✉

Some advances in reproductive technologies raise substantial ethical and psychological challenges, for example, the use of preimplantation genetic testing to select embryos based on non-medical traits. While studies have explored public willingness to use preimplantation genetic testing for medical or non-medical attributes, stated willingness may not reflect implantation decisions when such information is available. This large cross-national study examined public views on polygenic embryo testing. In a US sample ($N = 1,467$), participants were more willing to test for medical conditions (for example, heart disease) than non-medical traits (for example, antisocial behaviour or low intelligence), although over half supported testing for non-medical traits. In a forced-choice implantation task, participants used medical and non-medical information to a similar extent when making decisions. Similar patterns were observed in Chinese participants ($N = 623$). These findings suggest variability in choices across specific traits and conditions, rather than a uniform distinction between medical conditions and non-medical traits. The Stage 1 protocol for this Registered Report was accepted in principle on 7 October 2025. The protocol, as accepted by the journal, can be found at https://osf.io/vb9c2.

It is increasingly understood that investigation of public attitudes towards new and emerging technologies is necessary to devise responsive and responsible public policy[1]. Rapid advances in reproductive technologies, including in vitro fertilization and preimplantation genetic testing (PGT), raise complex psychological and social issues that require empirical and normative analysis. These technologies not only transform the nature of reproductive decision-making but also present profound ethical dilemmas[2]: although they empower prospective parents to improve their future children's health outcomes[3], they also open the door to selecting embryos based on traits beyond health, such as intelligence or other non-medical characteristics, raising the spectre of historical eugenic practices[4,5].

PGT has emerged as a widely available option, offered commercially through private companies and clinics, purporting to give intended parents greater control over their children's health outcomes before birth[6]. Initially developed to detect severe genetic disorders, PGT is now also offered in some clinics for assessing polygenic risk scores for common diseases, such as diabetes and heart disease, and even non-disease traits, such as sex, eye colour and cognitive ability[7–10]. Moreover, PGT is commonly said to be superior to other technologies such as genome editing, which carries higher risks and raises distinctive ethical concerns[1]. It may also be preferable to non-invasive prenatal testing, which, although less physically invasive, may lead to difficult decisions such as termination of an established pregnancy.

[1]Uehiro Oxford Institute, University of Oxford, Oxford, UK. [2]Department of Psychiatry, University of Oxford, Oxford, UK. [3]Department of Economics, University of Exeter, Exeter, UK. [4]School of Psychology, University of Waterloo, Waterloo, Ontario, Canada. [5]Centre for Biomedical Ethics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [6]School of Psychology, University of Kent, Canterbury, UK. [7]Center for Psychological Sciences, Zhejiang University, Hangzhou, China. [8]John Radcliffe Hospital, Oxford University Hospitals NHS Foundation Trust, Oxford, UK. [9]Biomedical Ethics Research Group, Murdoch Children's Research Institute, Parkville, Victoria, Australia. ✉e-mail: e.awad@exeter.ac.uk; dominic.wilkinson@uehiro.ox.ac.uk; jsavules@nus.edu.sg

Research on public perceptions of PGT consistently shows widespread acceptance of its use for detecting lethal or severe medical conditions before birth, and a general willingness to use PGT for these purposes[11–14]. While there is strong support for allowing patients to make their own choices regarding the use of PGT, there is broad consensus that it should not be used for selecting non-medical traits, especially aesthetic characteristics, such as perceived physical attractiveness[11–14]. Moreover, the public appears to disapprove of testing for non-medical traits that are not aesthetic, such as intelligence or the big five personality factors[15] (including conscientiousness and neuroticism). In one study[14], although the approval rate for PGT (the combined percentages of 'approve' and 'strongly approve') for 12 medical conditions ranged between 55% and 80%, approval for 12 non-medical traits ranged between 20% and 37%. This clear distinction between approval of PGT for medical conditions versus non-medical traits could be related to ethical concerns about false expectations, overly controlling parental attitudes and potential societal implications[14] (for example, concerns about eugenics, inequality and stigmatization).

Despite the extensive research on public perceptions (including patients' decisions and experiences; see ref. 13 for a review) regarding the use of PGT to test for specific attributes, there has been less focus on how this information, once acquired, influences people's preferences on which embryos to choose with the aim of achieving implantation (hereafter referred to as implantation decisions). Such information may be provided for attributes even when not explicitly requested, or may become available via incidental findings. As such, this question has important practical, psychological and ethical importance because it bears directly on how individuals weigh perceived risks and benefits of specific attributes under realistic decision-making conditions. In turn, this helps reveal potential underlying biases and value systems in key domains, such as reproductive choices and perceptions of life quality.

While the apparent strong aversion to testing for non-medical traits might suggest that parents would disregard such information when making implantation decisions, it is also plausible that, once provided, information about an embryo's chances of developing such traits could still substantially impact their implantation choices. This impact may occur even if parents initially choose not to learn about these traits—a well-documented psychological phenomenon known as deliberate ignorance[16]. In this context, deliberate ignorance can function as a strategic tool, allowing parents to avoid social or moral accountability (for example, an explicit preference to select for traits like intelligence), or as a cognitive-sustainability mechanism, helping them manage conflicting desires—such as wanting intelligent offspring while simultaneously recognizing the ethical concerns and potential eugenic implications of selecting for such traits.

This study investigated the possible discrepancy between screening and implantation decisions by testing the following predictions: (1) when it comes to decisions about conducting screening, US-based participants would generally choose to avoid testing for non-medical traits, replicating previous findings; however, (2) when information about the likelihood of embryos possessing non-medical traits was provided, this would influence participants' implantation decisions in the selection of embryos. We further hypothesised that such avoidance in implantation decisions (3) would not be limited to US-based or broadly Western participants, but would replicate in a Chinese sample, and (4) may be driven by the perceived level of non-controllability of the consequences of having undesirable non-medical traits (for example, the difficulty in altering or mitigating the negative effects of these attributes).

Given the literature reviewed earlier and the pilot studies we performed, we derived the following specific hypotheses (see Table 1 for an overview of our design). First, previous studies on willingness to perform polygenic testing for medical conditions and non-medical traits on embryos led to the following two hypotheses:

Hypothesis 1 (H1): Less than half of US-based participants will be willing to perform polygenic testing for non-medical traits.

Hypothesis 2 (H2): US-based participants will be more willing to perform polygenic testing for medical conditions than for non-medical traits.

These two hypotheses focused on the public's limited willingness to seek information about non-medical traits in both absolute (H1) and relative (H2) terms. If confirmed, these two hypotheses might seem to suggest that because people were not willing to test for non-medical traits, they would also tend to avoid using information derived from such tests if offered to them. However, on the basis of studies of deliberate ignorance in psychology[16] and results from two pilot studies, we hypothesised that when such information is provided, it would be used, even to a degree similar to that of medical conditions. The following two hypotheses concern the avoidance of undesirable non-medical traits in implantation decisions, both in absolute terms and compared with the avoidance of adverse medical conditions:

Hypothesis 3 (H3): US-based participants will use information about the likelihood of undesirable non-medical traits to avoid those traits (that is, they will be more willing to choose to implant an embryo with average rather than above-average chances of undesirable non-medical traits).

Moreover, a tendency to avoid undesirable non-medical traits in implantation decisions would be at least comparable to the tendency to avoid adverse medical conditions:

Hypothesis 4 (H4): US-based participants will exhibit an avoidance of undesirable non-medical traits to a degree that is at least as strong as (that is, either not statistically significantly different from or statistically greater than) their avoidance of adverse medical conditions.

Our pilot studies suggested that the above hypotheses would be supported for US-based participants. However, their generalizability to other countries, especially those outside the Western world, was not clear. While there was not sufficient evidence to make an informed prediction on whether people would express similar preferences in other cultural contexts, work using Singapore-based participants (with a similar set-up to ref. 1) shows that more than 50% of participants approve of PGT for educational attainment[17] (a non-medical trait). This tendency could be similarly high in other non-Western populations with similar or even higher Chinese ethnic representation. With that in mind, one would have expected that participants based in China would be willing to use information about non-medical traits to at least a similar (if not higher) degree than US-based participants.

Hypothesis 5 (H5): Participants based in China will show similar (or stronger) preferences to US-based participants with respect to H3 and H4: (1) Chinese participants will use information about the likelihood of undesirable non-medical traits to avoid those traits, and (2) Chinese participants will exhibit an avoidance of the undesirable non-medical traits to a degree that is at least as strong as their avoidance of adverse medical conditions.

Finally, we hypothesised that the tendency to avoid undesirable attributes (be it a medical condition or a non-medical trait) would be mediated by the perceived non-controllability of the potential negative consequences (or harm) of this attribute. An undesirable attribute whose existence or harmful effects are hard to potentially control would be expected to have a more negative impact on a child's future well-being and would thus be seen as less desirable:

Hypothesis 6 (H6): US-based participants will be more likely to avoid attributes considered undesirable (both medical conditions and non-medical traits) that are perceived to have lower controllability compared with those that are perceived to have higher controllability.

To test these hypotheses, we performed two studies (Extended Data Fig. 1). Study 1a tested hypotheses H1–H4 with US-based participants. This study randomly allocated participants to complete one of two 'arms': (1) the willingness task arm (H1 and H2), or (2) the implantation task arm (H3 and H4). This study also used a

Table 1 | Design table

Question	Hypothesis	Sampling plan (for example, power analysis)	Analysis plan	Interpretation given to different outcomes
Question 1: Will participants show willingness to perform polygenic testing for non-medical traits?	Less than half of US-based participants will be willing to perform polygenic testing for non-medical traits.	Power analyses indicate that a sample size of $N=600$ will provide over 95% power to detect an effect size of $d=-0.2$, which is considered small. This effect size is more conservative than those estimated in previous studies and pilot studies.	We will estimate the mean proportion of 'yes' responses for the two non-medical traits and compare it to chance (0.5). Data preparation: We will recode 'yes' responses as 1 and 'no' or 'not sure' responses as 0, subtract 0.5 from all values and set 'traits' as the baseline. Data will be pooled across the attributes. Model: An LPM will be used to estimate the intercept, which directly represents the mean deviation of 'yes' responses for traits from 0.5. We will test whether this intercept is negative and significantly different from zero to evaluate our hypothesis. Clustering: Participant-level clustered standard errors will be used to account for correlations within participants.	A significant negative coefficient will be interpreted as (supportive or suggestive) evidence that less than half US-based participants are willing to perform polygenic testing for non-medical traits. In the case of no significance, we will conduct a TOST procedure[47], using $d=0.2$ as a threshold. If the larger P value is smaller than the chosen threshold (0.05), we will interpret this as evidence for null effect.
Question 2: Will the willingness to perform polygenic testing differ between medical conditions and non-medical traits?	US-based participants will be more willing to perform polygenic testing for medical conditions than for non-medical traits.	Power analyses indicate that a sample size of $N=600$ will provide over 95% power to detect an effect size of $d=-0.2$, which is considered small. This effect size is more conservative than those estimated in previous studies and pilot studies.	We will estimate the ATE of the treatment (medical conditions versus non-medical traits) on the binary outcome variable (1 for yes, 0 for no or not sure). Data preparation: We will pool data across the two attributes in each group (total of four), with medical conditions as the control group and non-medical traits as the treatment group. Model: An LPM will be used to compute the ATE as the difference in mean responses between the two groups (a non-parametric estimator). We will test whether this coefficient is negative and significantly different from zero to evaluate our hypothesis. Clustering: Participant-level clustered standard errors will be used to account for correlations within participants.	A significant negative coefficient will be interpreted as (supportive or suggestive) evidence that US-based participants are more willing to perform polygenic testing for medical conditions than for non-medical traits. In the case of no significance, we will conduct a TOST procedure, using $d=0.2$ as a threshold.
Question 3: Will the choice to implant an embryo differ based on chance levels of non-medical traits?	US-based participants will use information about the likelihood of undesirable non-medical traits to 'avoid those traits' (that is, they will be more willing to choose to implant an embryo with average chances of undesirable non-medical traits compared with one with above-average chances of the traits).	Power analyses indicate that a sample size of $N=600$ will provide over 95% power to detect an effect size of $d=-0.2$, which is considered small. This effect size is more conservative than those estimated in pilot studies.	We will estimate the AMCE of the above-average chance level for non-medical traits using conjoint analysis[49], pooling data across both traits. Data preparation: We will filter the dataset to include only rows where the attribute type is 'trait' and where the below-average chance level is excluded (that is, only average and above-average chance levels remain). The average chance level will be set as the baseline category. Model: Using an LPM, we will include the dummy variable Above_Average to estimate the difference between above-average and average chance levels. The AMCE will be tested by examining whether the coefficient for Above_Average is negative and significantly different from zero. Clustering: Participant-level clustered standard errors will be used to account for correlations within participants.	A significant negative coefficient for above-average treatment will be interpreted as (supportive or suggestive) evidence that US-based participants prefer average chances of traits more than above-average chances of those traits. In the case of no significance, we will conduct a TOST procedure, using $d=0.2$ as a threshold.
Question 4: Will the degree of avoidance for non-medical traits be comparable to that for medical conditions?	US-based participants will exhibit an avoidance of undesirable non-medical traits to a degree that is at least as strong as (that is, either not statistically significantly different from or statistically greater than) their avoidance of adverse medical conditions.	Power analyses indicate that a sample size of $N=600$ will provide over 95% power to detect an effect size of $d=-0.2$, which is considered small. This effect size is more conservative than those estimated in pilot studies.	We will estimate the ACIE of the interaction between above-average chance levels and attribute type (non-medical traits versus medical conditions). Data preparation: We will filter the dataset to exclude rows where the below-average chance level is present, ensuring only average and above-average chance levels remain. Model: Using an LPM, the interaction term Above_Average × Attribute_Type will be included to estimate the differential effect of above-average chance levels depending on whether the attribute is a non-medical trait or medical condition. The coefficient of the interaction term will be interpreted as the ACIE, representing the difference in effects between traits and conditions. Clustering: Participant-level clustered standard errors will be calculated to account for correlations within participants. Inferiority test: As H4 states that 'US-based participants will exhibit an avoidance of non-medical traits at least as strongly as medical conditions', we will conduct an inferiority test[47]. The inferiority threshold will be set to 0.2 (considered a conventionally small standardised effect size), indicating that non-medical traits must show at least this level of avoidance compared with medical conditions. We will test this using a one-sided t test. The null hypothesis is that the interaction effect (Above_Average × Attribute_Type) is greater than or equal to the threshold (that is, effect ≥ 0.2). The alternative hypothesis is that the interaction effect is below the threshold (that is, effect < 0.2). Above_Average × Attribute_Type is the interaction between chance level (above-average versus average) and attribute type (trait versus condition).	A significant negative interaction effect will be interpreted as (supportive or suggestive) evidence for higher aversion to above-average levels for non-medical traits compared with aversion of above-average levels of medical conditions by US-based participants. In the case of no significance, we will look at the results of the inferiority test (a one-sided t test procedure), using $d=0.2$ as a threshold.

Table 1 (continued) | Design table

Question	Hypothesis	Sampling plan (for example, power analysis)	Analysis plan	Interpretation given to different outcomes
Question 5: Will Chinese participants show similar preferences to US participants regarding chance levels of non-medical traits and medical conditions?	Participants based in China will show similar (or stronger) preferences to US-based participants with respect to H3 and H4.	Power analyses indicate that a sample size of N=600 will provide over 95% power to detect an effect size of d=−0.2, which is considered small. This effect size is assumed to match those proposed for H3 and H4.	This will follow the same plan as H3 and H4 for data collected in study 1b with Chinese participants.	Interpretation will match that for H3 and H4, but for Chinese participants.
Question 6: Will the avoidance of embryos with undesirable attributes differ based on perceived controllability?	US-based participants will be more likely to avoid attributes considered undesirable (both medical conditions and non-medical traits) that are perceived to have lower controllability compared with those that are perceived to have higher controllability.	Power analyses indicate that a sample size of N=1,800 will provide over 95% power to detect an effect size of d=−0.2, which is considered small. This effect size is assumed to match those proposed for H3 and H4.	This will be tested in the following steps across two studies. Study 1a (exploratory measurement-of-mediation design): A measurement-of-mediation analysis will be performed using established mediation analysis methods[35,36]. If study 1a provides statistical evidence for mediation, study 2 will be conducted. Study 2 (confirmatory measurement-of-mediation design): Data from the control group (mediator varies freely) will be used to confirm the mediation effect observed in study 1a. Study 2 (manipulation-of-mediator design): The blockage and enhancement manipulations will be compared with the control group to test whether controllability is a mediator. An interaction term between chance levels (above average versus average) and controllability (low controllability versus freely varying versus high controllability) will be tested to assess whether controllability is a partial mediator. Moreover, an AMCE of chance levels will be calculated for the blockage condition to determine whether controllability is a full mediator.	Study 1a: If study 1a provides statistical evidence that controllability mediates the relationship between attribute type and avoidance, study 2 will be conducted. Study 2: A significant interaction between chance levels and high controllability (versus freely varying) or between chance levels and low controllability (versus freely varying) will be interpreted as (supporting or suggestive) evidence that controllability is a 'partial mediator'. A significant effect of AMCE will be interpreted as (supportive or suggestive) evidence that the controllability is not a full mediator. In the case of no significance, we will conduct a TOST procedure (using d=0.2 as a threshold) to test for full mediation (that is, blocking controllability eliminates the effect of attribute type on avoidance). In all other cases of no significance, we will conduct a TOST procedure, using d=0.2 as a threshold.

measurement-of-mediation design to test H6 (as an exploration of the mediator), the result of which was used to determine whether the third study was needed. Study 1b only included the implantation task (second arm) and was otherwise identical in design and run with Chinese participants to test H5. Finally, study 2 was planned to use a concurrent double randomization design combining measurement-of-mediation and manipulation-of-mediator designs to test H6. However, on the basis of the measurement-of-mediation results, this study was not run.

Both studies considered four attributes: two medical conditions (heart disease and vision impairment) and two non-medical traits (low IQ and antisocial behaviour). Participants assigned to the willingness task arm in study 1a were asked about their willingness to perform PGT for all four attributes. Participants who completed the implantation task arm (those assigned to the implantation task arm in study 1a and all participants in study 1b; Fig. 1) were randomly assigned to make implantation decisions about two attributes (out of the four) in ten implantation dilemmas involving a choice between two embryos (embryo A versus embryo B) that feature those two attributes with different chance levels (significantly above-average chances, significantly below-average chances and average chances).

This study provides several contributions to the literature on public attitudes towards PGT. First, it goes beyond simply assessing the public's expressed willingness to perform testing by examining whether and how individuals use information about both medical conditions and non-medical traits to guide implantation decisions. Second, it expands the scope of existing research by including an international comparative perspective, directly measuring preferences of US-based and Chinese participants—two culturally distinct populations. Third, it analyses the mechanisms underlying why certain attributes, particularly non-medical traits, influence public preferences, focusing on the role of perceived controllability in shaping implantation decisions.

Results

Final sample size and demographics

We collected data for study 1a and study 1b. For study 1a, we recruited participants on Prolific with a target sample of 1,500 US-based adults, quota-sampled to be nationally representative by age, gender and ethnicity. Of the 1,501 participants who completed the survey and were compensated, 1,467 passed both attention checks and were included in the main analyses. Study 1b recruited participants using Credamo with a target sample of 800 China-based adults, quota-sampled to be nationally representative by age and gender. Of the 800 participants who completed the survey and were compensated, 623 passed both attention checks and were included in the main analyses. Table 2 summarises

https://doi.org/10.1038/s41562-026-02562-w

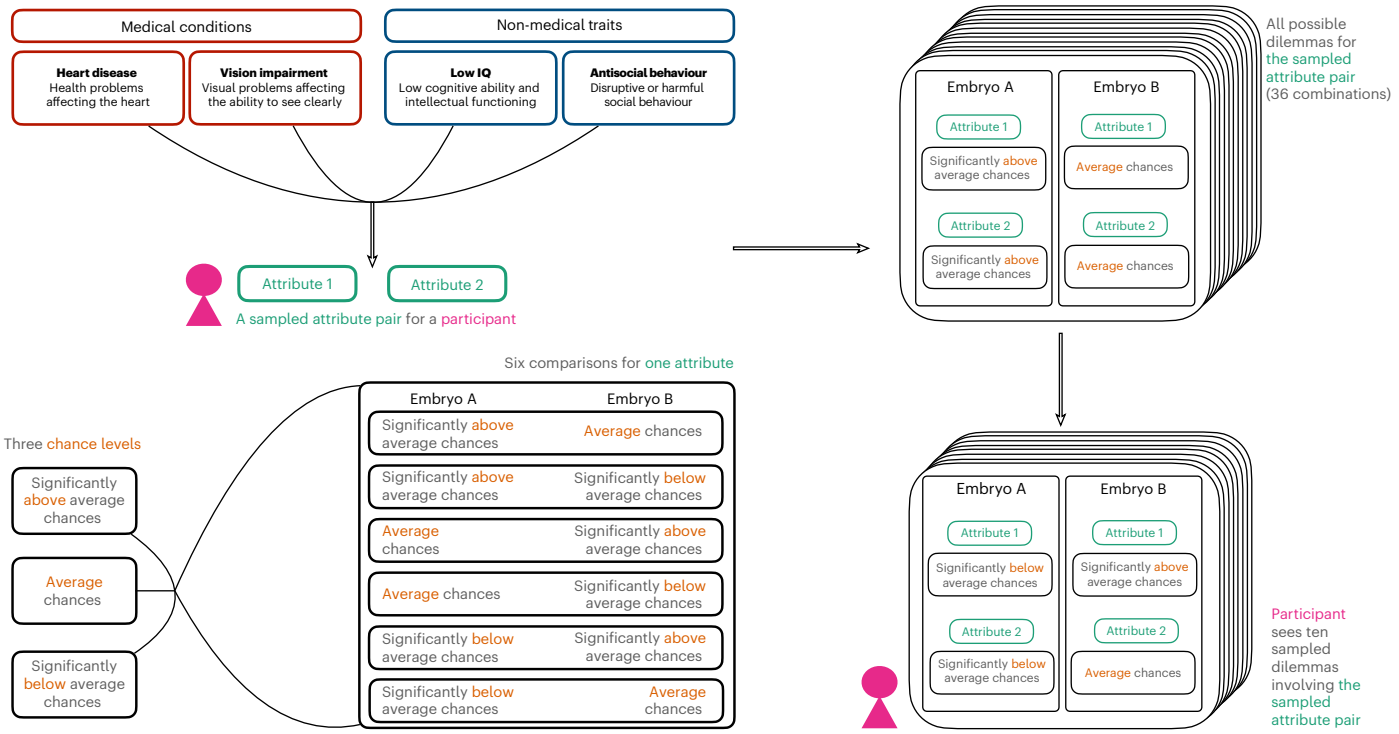


Fig. 1 | Experimental design and dilemma sampling in the implantation task arm. Participants were presented with embryo implantation dilemmas involving two attributes sampled from a set of four (two medical conditions and two non-medical traits). Each attribute could take one of three chance levels (significantly below average, average or significantly above average), yielding six ordered comparisons per attribute. Combining 2 attributes yields 36 possible embryo–pair dilemmas for a given attribute pair. Each participant was shown a random subset of 10 dilemmas drawn from these 36 and, for each comparison, they were asked to choose between embryo A and embryo B. Attribute pairings and chance configurations were counterbalanced across participants.

the focal results and conclusions for all preregistered hypotheses. For the number of participants in each experimental condition, see Supplementary Tables 8 and 9 for study 1a and 1b, respectively; for demographic details, see Supplementary Tables 10 and 11 for study 1a and 1b, respectively.

Robustness analyses (study 1a and study 1b). As registered, we repeated all main analyses using alternative exclusion criteria to assess their robustness. Specifically, we replicated the results (1) using all paid participants who passed at least one attention check, and (2) using the more restrictive sample of participants included in the main analysis who also passed all comprehension checks (Supplementary Tables 12 and 13), separately for the US and China samples. Across all specifications, the direction, magnitude and statistical significance of the key results were mostly unchanged (with the exception of one case moving from supported to suggested). Detailed robustness analyses are reported in Extended Data Table 1.

Preregistered analyses

Willingness to test (study 1a). Half of the participants in study 1a were randomly assigned to a willingness task arm in which they indicated whether they would be willing to perform polygenic testing for each of the four attributes (responses: yes, no or not sure).

H1 (non-medical traits, USA). In contrast to our preregistered hypothesis H1, more than half of participants (61.4%) expressed willingness to perform polygenic testing for non-medical traits (Fig. 2a). A linear probability model (LPM) with participant-level clustered standard errors showed that the estimated proportion of yes responses for non-medical traits was significantly above 50% ($\beta = 0.114$, where β is the coefficient from the linear probability model; 95% confidence interval (CI), [0.083, 0.145]; $t(734) = 7.34$; $P < 0.001$).

H2 (medical conditions versus non-medical traits, USA). Consistent with our preregistered hypothesis H2, participants were significantly more willing to perform polygenic testing for medical conditions (66.7% answered yes) than for non-medical traits (61.4% answered yes; Fig. 2a). An LPM with participant-level clustered standard errors showed that the estimated difference in willingness to test between non-medical traits and medical conditions was statistically significant and in the predicted direction ($\beta = -0.053$; 95% CI, [−0.086, −0.020]; $t(734) = -3.12$; $P = 0.0018$). This result meets our preregistered threshold for supportive evidence ($P < 0.0025$).

Although participants were, on average, more willing to test for medical conditions than for non-medical traits, Fig. 2b reveals heterogeneity across attributes. The higher willingness to test for medical conditions is driven primarily by a very high willingness to test for heart disease. By contrast, willingness to test for vision impairment was similar or lower than the average willingness to test for non-medical traits. This pattern suggests that willingness to test is not straightforwardly determined by a binary distinction between 'medical' and 'non-medical' but is dependent on the specific nature of individual attributes.

Implantation decisions (study 1a). Each participant assigned to the implantation task arm was allocated two attributes sampled from four possible ones (Fig. 1). Our preregistered analyses tested whether participants were less likely to implant embryos with above-average chances of non-medical attributes relative to average chances (which we call avoidance of such attributes; H3), and how that compared with avoidance of medical conditions (H4).

H3 (avoidance of undesirable non-medical traits, USA). Consistent with preregistered hypothesis H3, US-based participants avoided the undesirable non-medical traits. An LPM with participant-level clustered standard errors and pooling across both traits showed that embryos

Table 2 | Summary of results for preregistered hypotheses

Hypothesis	Sample	Focal test	Key result	Conclusion	Finding
H1	USA	Willingness to test non-medical traits versus 50%	61.4% answered yes; β=0.114; 95% CI, [0.083, 0.145]; t=7.34; P<0.001	Not supported	More than half of US-based participants were willing to test for non-medical traits
H2	USA	Willingness to test medical conditions versus non-medical traits	β=−0.053; 95% CI, [−0.086, −0.020]; t=−3.12; P=0.0018	Supported	US-based participants were more willing to perform polygenic testing for medical conditions than for non-medical traits
H3	USA	Avoidance of undesirable non-medical traits (above-average versus average chances)	β=−0.341; 95% CI, [−0.366, −0.316]; t=−25.56; P<0.001	Supported	US-based participants were more willing to choose to implant an embryo with average rather than above-average chances of undesirable non-medical traits
H4	USA	Avoidance of non-medical traits versus medical conditions	β=−0.050; 95% CI, [−0.089, −0.011]; t=−2.51; P=0.012; non-inferiority criterion met (d=0.20, one-sided P<0.001)	Supported	US-based participants chose to avoid undesirable non-medical traits to a degree at least as strong as their avoidance of adverse medical conditions
H5a	China	Avoidance of undesirable non-medical traits (above-average versus average chances)	β=−0.411; 95% CI, [−0.438, −0.384]; t=−30.07; P<0.001	Supported	China-based participants were more willing to choose to implant an embryo with average rather than above-average chances of undesirable non-medical traits
H5b	China	Avoidance of non-medical traits versus medical conditions	β=−0.090; 95% CI, [−0.131, −0.049]; t=−4.35; P<0.001; non-inferiority criterion met (d=0.20, one-sided P<0.001)	Supported	China-based participants chose to avoid undesirable non-medical traits to a degree at least as strong as their avoidance of adverse medical conditions
H6	USA	Mediation by perceived controllability	Across six pairwise contrasts, three ACMEs were non-significant and three were significant in the opposite direction to that predicted (Supplementary Table 14)	Not supported	US-based participants' ratings of (non-)controllability of attributes did not mediate their decisions to avoid those attributes

with above-average chances of undesirable non-medical traits were significantly less likely to be chosen for implantation than embryos with average chances (β = −0.341; 95% CI, [−0.366, −0.316]; t(612) = −25.56; P < 0.001; Fig. 3a). This result meets our registered threshold for supportive evidence (P < 0.0025).

H4 (medical conditions versus non-medical traits, USA). We next examined whether avoidance of undesirable non-medical traits was at least as strong as avoidance of adverse medical conditions. An LPM with an interaction between chance level (above versus average) and attribute type (traits versus conditions) showed a negative interaction (β = −0.050; 95% CI, [−0.089, −0.011]; t(731) = −2.51; P = 0.012), indicating greater avoidance of undesirable non-medical traits relative to medical conditions. Consistent with the preregistered non-inferiority test, this effect exceeded the specified non-inferiority margin (d = 0.20, where d is Cohen's d, a standardised effect size; one-sided P < 0.001), providing confirmatory support for H4.

Despite the supporting evidence for H4, Fig. 3a shows meaningful variation in implantation preferences between the four attributes. Avoidance of attributes was most pronounced for heart disease. Low IQ and antisocial behaviour were avoided to a similar but slightly weaker degree than heart disease, and to a significantly greater degree than visual impairment. A follow-up exploratory analysis provided suggestive evidence that heart disease was avoided more than low IQ (β = 0.080; 95% CI, [0.025, 0.135]; t(731) = 2.91; P = 0.004), whereas the avoidance for heart disease and that for antisocial behaviour were not significantly different (β = 0.04; 95% CI, [−0.013, 0.093]; t(731) = 1.51; P = 0.131). A two-one-sided test (TOST) with a smallest effect size of interest of d = 0.20 indicated statistical equivalence (P_{TOST} = 0.018).

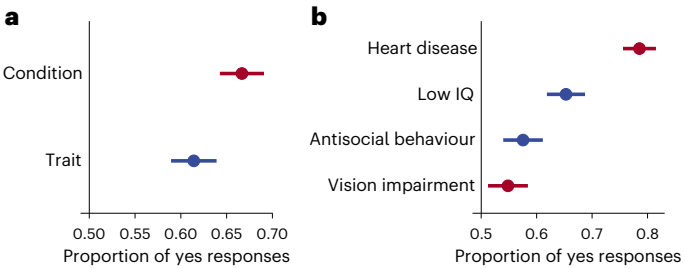


Fig. 2 | Willingness to perform genetic testing by attribute type and attribute in study 1a (US sample). a, Proportion of yes responses to a ternary (yes, no or not sure) question asking whether participants would be willing to perform polygenic testing, averaged by attribute type (medical conditions versus non-medical traits; n = 1,470 responses per type). **b**, Proportion of yes responses for each individual attribute (n = 735 per attribute). Medical conditions (heart disease and vision impairment) are shown in red; non-medical traits (low IQ and antisocial behaviour) are in blue. Points indicate means; error bars represent 95% CIs (±1.96 s.e., clustered by participant). N = 735 participants (willingness arm, US sample). Each participant responded to all four attributes, contributing two responses per attribute type and one per attribute. No control group was used; attributes were compared within participants.

Implantation decisions (study 1b). We next examined implantation decisions among participants in China, following the same registered analytic approach as for US participants. In this study, all participants were assigned to the implantation task arm. As in study 1a, our focus is on avoidance of the undesirable non-medical traits and how that compares with avoidance of the medical conditions.

H5a (avoidance of undesirable attributes, China). Consistent with our preregistered hypothesis, when confronted with implantation decisions, Chinese participants used information about the likelihood of undesirable non-medical traits to avoid those traits. An LPM with participant-level clustered standard errors showed that when pooling

https://doi.org/10.1038/s41562-026-02562-w

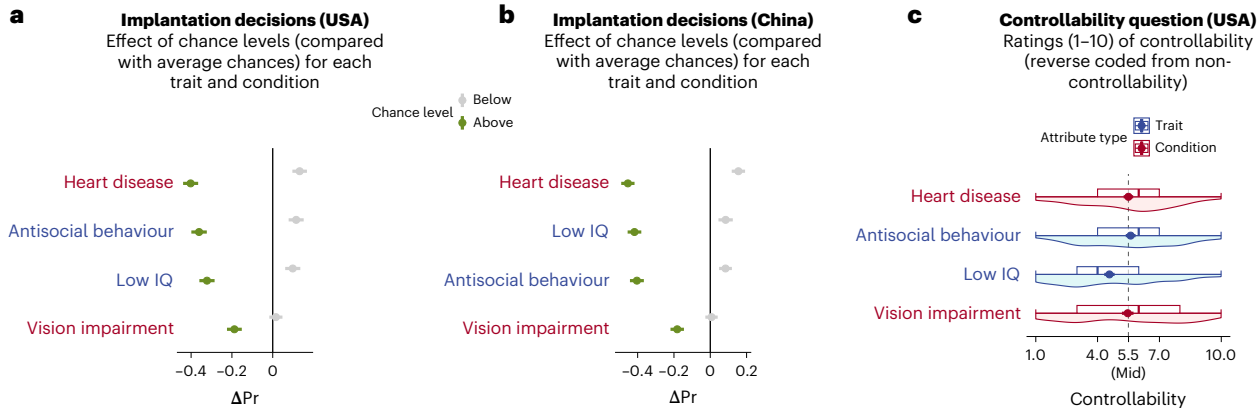


Fig. 3 | Implantation decisions and perceived controllability in studies 1a and 1b. a, US sample. Change in the probability of selecting an embryo (ΔPr) when its chance level for a given attribute is significantly above average (green) or below average (grey) relative to average chances, estimated separately for each attribute. Points show means from LPMs with cluster-robust standard errors; error bars indicate 95% CIs (±1.96 s.e., clustered by participant). Number of participants per attribute: heart disease = 355, antisocial behaviour = 357, low IQ = 378, vision impairment = 374 participants; number of ratings per attribute = 10 × *N* (each participant completed 10 dilemmas per assigned attribute; implantation arm, US sample). Each participant was randomly assigned two of the four attributes. The reference category in all models is average chances. **b**, China sample. Same analysis as in **a**. Number of participants per attribute: heart disease = 320, antisocial behaviour = 313, low IQ = 300, vision impairment = 313 participants; number of ratings per attribute = 10 × *N* (each participant completed 10 dilemmas per assigned attribute; implantation arm, China sample). **c**, Perceived controllability ratings for each attribute (US sample), measured on a 1–10 scale (reverse-coded from non-controllability; higher values indicate greater perceived controllability). Violin plots show the full response distribution; box plots show median (centre line), interquartile range (box bounds, 25th–75th percentile) and whiskers extending to the scale limits (1 and 10); points show means with error bars indicating 95% CIs (±1.96 s.e., clustered by participant). Each participant rated each assigned attribute once. Medical conditions are shown in red; non-medical traits are in blue. No control group was used; attributes were compared within the same participants.

across both traits (low IQ and antisocial behaviour), embryos with above-average chances of undesirable non-medical traits were significantly less likely to be chosen for implantation than embryos with average chances ($\beta = -0.411$; 95% CI, [−0.438, −0.384]; $t(509) = -30.07$; $P < 0.001$; Fig. 3b). This result meets our preregistered threshold for supportive evidence ($P < 0.0025$), providing confirmatory support for H5a.

H5b (medical conditions versus non-medical traits, China). We next tested whether avoidance of undesirable non-medical traits was at least as strong as avoidance of adverse medical conditions. An LPM with participant-level clustered standard errors showed that the interaction between above-average chance levels (versus average chance) and non-medical traits (versus medical conditions) was negative and statistically significant ($\beta = -0.090$; 95% CI, [−0.131, −0.049]; $t(622) = -4.35$; $P < 0.001$), indicating greater avoidance of undesirable non-medical traits relative to medical conditions. Consistent with the preregistered non-inferiority test, this effect exceeded the specified non-inferiority margin ($d = 0.20$; one-sided $P < 0.001$), providing confirmatory support for H5b.

Despite this, Fig. 3b suggests a slightly different attribute-level pattern from the US sample: in the Chinese sample, avoidance of low IQ and antisocial behaviour was closer to that of heart disease. Exploratory follow-up analyses indicated that avoidance of heart disease did not differ significantly from avoidance of either low IQ ($\beta = 0.036$; 95% CI, [−0.019, 0.091]; $t(622) = 1.31$; $P = 0.191$) or antisocial behaviour ($\beta = 0.050$; 95% CI, [−0.005, 0.105]; $t(622) = 1.77$; $P = 0.077$). TOSTs, using a smallest effect size of interest of $d = 0.2$, further indicated that differences in avoidance between heart disease and both non-medical traits were statistically equivalent (low IQ, $P_{TOST} = 0.013$; antisocial behaviour, $P_{TOST} = 0.042$). Nevertheless, as in the US sample, these results suggest that although category-level contrasts support the preregistered hypotheses, the Chinese attribute-level pattern is also less consistent with a simple medical versus non-medical distinction.

Controllability as a potential mechanism (study 1a). *H6 (controllability)*. After completing the implantation task arm, participants rated the perceived controllability of the two attributes they were assigned. To assess whether perceptions of non-controllability of an attribute mediated the extent to which they were avoided, we used measurement of mediation to estimate mediation of non-controllability for each pairwise contrast between attributes, with avoidance aggregated to the participant-by-attribute level (Supplementary Table 14). The results did not support the predicted mediation pattern: three contrasts were non-significant (vision impairment versus low IQ, vision impairment versus heart disease and antisocial behaviour versus heart disease) and the remaining three (vision impairment versus antisocial behaviour, low IQ versus antisocial behaviour and low IQ versus heart disease) were significant (suggestive; at 0.05) in the opposite direction to that expected (for example, antisocial behaviour was avoided to a greater extent relative to low IQ, but was also perceived as more controllable). Perceived controllability thus does not seem to mediate avoidance of different attributes. Taken together, these findings indicate that perceived controllability does not provide a plausible explanation for the observed differences in avoidance across attributes.

Exploratory analyses

Morality, legality and the decision to implant the only available embryo. Participants in the willingness task arm of study 1a were also asked additional questions about (1) the morality of performing polygenic testing, (2) whether such testing should be legally banned, and (3) whether they would implant an embryo if it was the only one available and it had a significantly above-average chance of a given attribute (Fig. 4). These questions were not registered as primary outcomes and are reported here to contextualise the main findings.

Fewer than half of responses judged that polygenic testing for non-medical traits is morally wrong (18%) or should be legally banned (18.1%). LPMs with participant-level clustered standard errors showed that the estimated proportions of these responses for non-medical traits were significantly below 50% (for 'morally wrong', $\beta = -0.320$; 95% CI, [−0.345, −0.295]; $t(734) = -24.54$; $P < 0.001$; Fig. 4a; for 'should be legally banned', $\beta = -0.319$; 95% CI, [−0.344, −0.294]; $t(734) = -24.37$; $P < 0.001$; Fig. 4b). Furthermore, testing for medical conditions was judged morally wrong even less often than testing for non-medical traits ($\beta = 0.059$; 95% CI, [0.037, 0.081]; $t(734) = 5.50$; $P < 0.001$; Fig. 4a).

https://doi.org/10.1038/s41562-026-02562-w

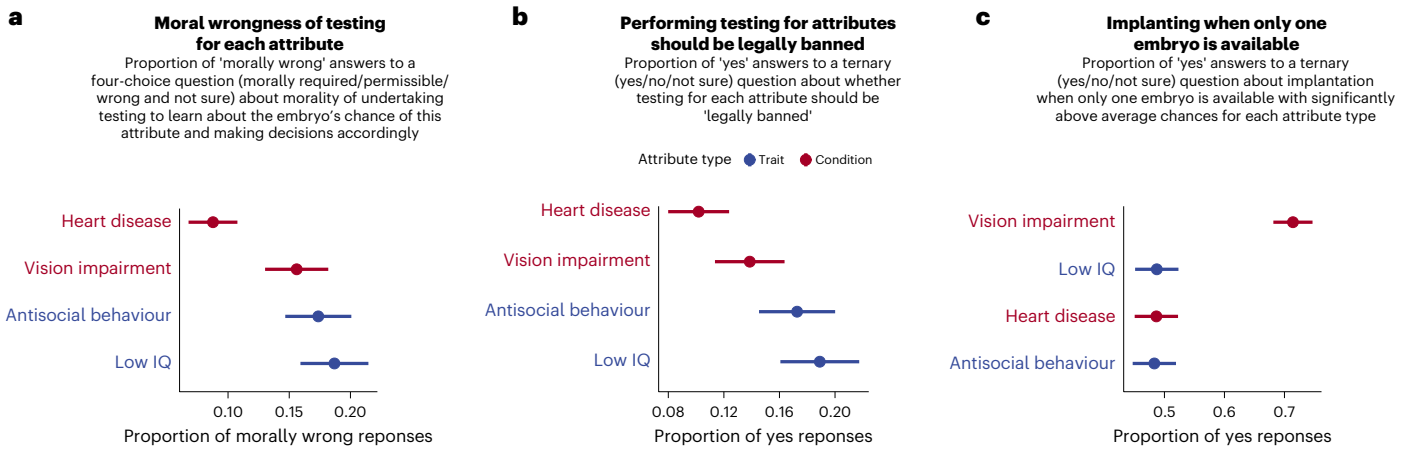


Fig. 4 | Normative and policy attitudes towards testing and implantation in study 1a (US sample). a, Proportion of responses judging testing for each attribute as morally wrong, based on a four-option question (morally required, morally permissible, morally wrong or not sure) about the permissibility of undertaking testing to learn about an embryo's chance of the attribute and acting on that information. **b**, Proportion of respondents answering yes to whether testing for each attribute should be legally banned, measured using a ternary response format (yes, no or not sure). **c**, Proportion of respondents answering yes to whether they would implant an embryo when only one embryo is available and it has a significantly above-average chance of the attribute. In **a**–**c**, medical conditions are shown in red and non-medical traits are in blue; points indicate means and error bars show 95% CIs (±1.96 s.e., clustered by participant). $N = 735$ participants (willingness arm, US sample); each participant responded to all four attributes. No control group was used; attributes were compared within the same participants.

Similarly, support for legal bans was significantly lower for medical conditions than for non-medical traits ($\beta = 0.061$; 95% CI, [0.039, 0.083]; $t(734) = 5.56$; $P < 0.001$; Fig. 4b). Taken together, these findings provide converging support for H1 and H2. Participants not only expressed greater willingness to test for medical conditions than for non-medical traits (H2) but also fewer responses judged testing for medical conditions to be morally wrong and deserving of legal restriction. Also, participants not only expressed a majority willingness (61.4%) to test for non-medical traits (H1) but also a minority of responses judged such testing to be morally wrong or deserving of legal restriction, indicating that resistance to testing for the two non-medical traits is less widespread than previously assumed.

Finally, when asked whether they would implant an embryo if it were the only viable option and had a significantly above-average chance of a given attribute, participants were substantially more willing to implant embryos associated with medical conditions than with non-medical traits ($\beta = -0.115$; 95% CI, [−0.142, −0.088]; $t(734) = -7.99$; $P < 0.001$; Fig. 4c). At the attribute level, willingness to implant was highest for embryos with above-average chance of vision impairment (71%), whereas willingness to implant embryos associated with above-average chance of antisocial behaviour and low IQ was 48.3% and 48.7%, respectively, with heart disease falling in between (48.6%). Notably, even under this constrained scenario, participants were nearly evenly split for three of the four attributes, indicating clear limits to permissiveness in implantation decisions.

Head-to-head attribute dilemmas. To further characterise how participants prioritised different attributes, we conducted exploratory head-to-head analyses of implantation decisions in which embryos differed on two attributes at the same time. This approach provides a direct comparison of how each attribute fares against the others in pairwise trade-offs. For each attribute pair, we aggregated choices across dilemmas comparing embryos with above-average chances on one attribute and below-average chances on the other, and the reverse comparison (above–below versus below–above). We call these crossed dilemmas. We also included dilemmas contrasting embryos where both attributes were above average with those where both were below average (above–above versus below–below; Fig. 5), which we call aligned dilemmas. These analyses focus on the subset of dilemmas in which both attributes deviated from average chances, allowing the two attributes to be directly contrasted (Fig. 1 for the full sampling scheme). Of primary interest are the crossed dilemmas (above–below versus below–above), where the two attributes are in direct conflict. In these dilemmas, we treat avoidance of attribute A as 'prioritised' over avoidance of attribute B if participants more often chose the embryo that was better on A (below-average chance of A) and worse on B (above-average chance of B) than the reverse option (above-average chance of A and below-average chance of B).

For each pair, we tested whether the choice share differed from 0.5 using an intercept-only LPM with participant-clustered standard errors. Across these head-to-head comparisons in the US sample, clear and structured preference patterns emerged. Avoidance of vision impairment was consistently less prioritised, losing to all other attributes. In other words, a majority of participants chose an embryo with a higher chance of visual impairment if it would have a lower chance of heart disease, low IQ or antisocial behaviour (versus heart disease, $\beta = 0.290$; 95% CI, [0.180, 0.400]; $t(247) = 5.23$; $P < 0.001$; versus antisocial behaviour, $\beta = 0.271$; 95% CI, [0.167, 0.375]; $t(279) = 5.16$; $P < 0.001$; versus low IQ, $\beta = 0.329$; 95% CI, [0.233, 0.425]; $t(279) = 6.74$; $P < 0.001$). The remaining three attributes formed a near non-transitive pattern: heart disease was prioritised over antisocial behaviour ($\beta = -0.144$; 95% CI, [−0.271, −0.017]; $t(235) = -2.22$; $P = 0.027$), antisocial behaviour was prioritised over low IQ ($\beta = 0.162$; 95% CI, [0.048, 0.276]; $t(283) = 2.78$; $P = 0.006$), whereas heart disease and low IQ were not statistically different from equal preference (50%; $\beta = 0.022$; 95% CI, [−0.107, 0.151]; $t(267) = 0.34$; $P = 0.734$). Nevertheless, a TOST with a smallest effect size of interest $d = 0.20$ did not support statistical equivalence ($P_{TOST} = 0.120$). Similar patterns were observed in the Chinese sample (Extended Data Fig. 2). Full results are in Supplementary Tables 22 and 23.

Cross-country comparisons (USA versus China). We conducted exploratory cross-country comparisons focusing on implantation decisions when embryos had above-average chances for each attribute. Across attributes, the overall structure of preferences was broadly similar in the US and Chinese samples, with above-average chances consistently reducing choice to implant (Fig. 3a,b and Extended Data Fig. 3). A notable exception concerned low IQ, for which avoidance was significantly stronger in the Chinese sample than in the US sample. Formally,

https://doi.org/10.1038/s41562-026-02562-w

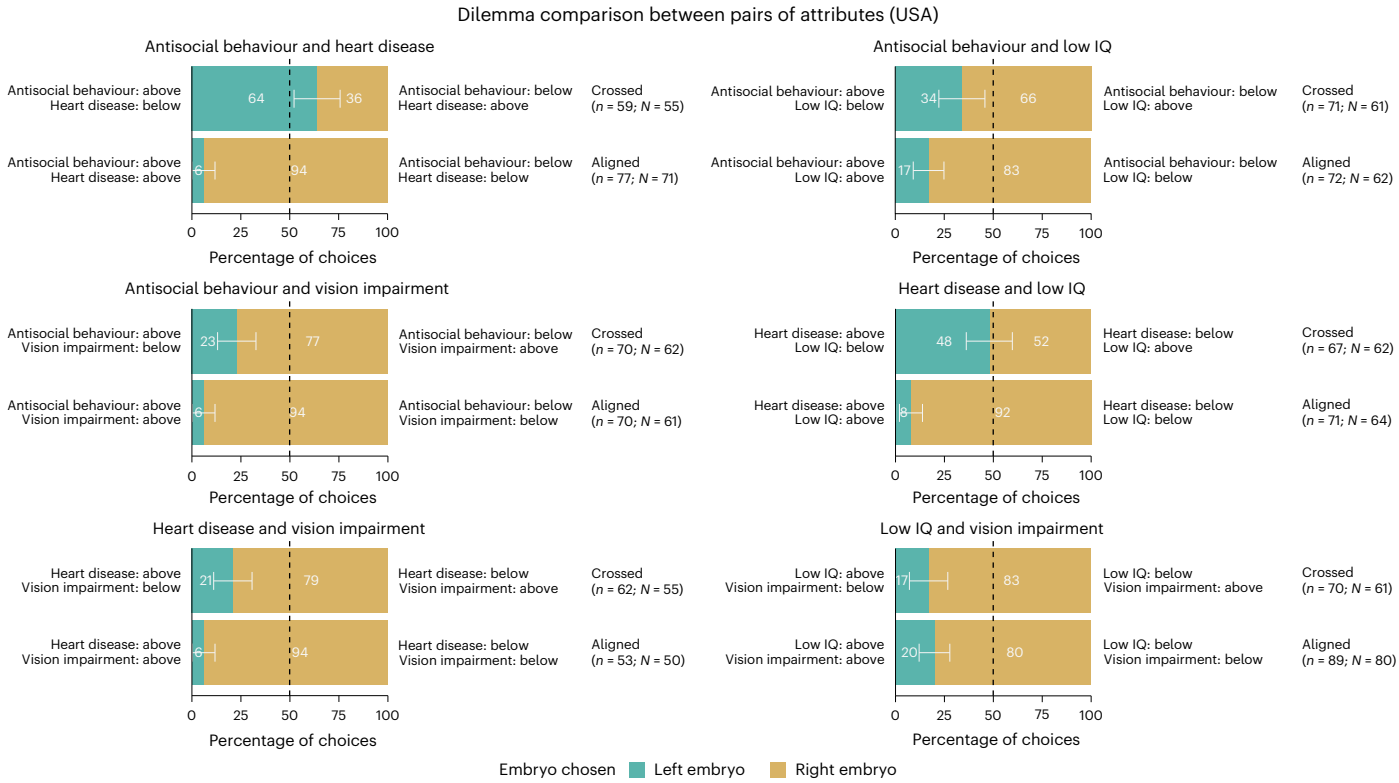


Fig. 5 | Head-to-head implantation dilemmas between pairs of attributes (US sample). Each plot shows implantation choices in dilemmas where embryos differ on two attributes, shown separately for crossed dilemmas (one embryo above average on one attribute and below average on the other) and aligned dilemmas (one embryo above average on both attributes and the other below average on both). Bars show the proportion of choices favouring the left embryo (teal) and right embryo (gold). The dashed vertical line marks 50% (equal preference). Error bars indicate 95% CIs (±1.96 s.e., clustered by participant). *n* (responses) and *N* (participants) per dilemma type and attribute pair are indicated on the figure. Participants were randomly assigned two of four attributes, so *N* varies by pair. No control group was used; the reference point is equal choice (50%).

this difference was captured by a negative interaction between above-average chance (versus average chances) and country (China versus USA) in an LPM with participant-level clustered standard errors ($\beta = -0.095$; 95% CI, [−0.148, −0.042]; $t = -3.53$; $P < 0.001$), indicating a larger reduction in implantation probability in China relative to the USA. This effect remains statistically significant as supportive evidence after Bonferroni correction (0.005/4 = 0.0012) for a family of four attribute-level tests (see Extended Data Table 2 for full results).

Controllability as a potential mechanism in the Chinese sample. Although perceived controllability was registered as a potential mechanism only for the US sample, we conducted an exploratory mediation analysis in the Chinese sample using the same specification. To assess measurement of mediation, we estimated mediation separately for each pairwise contrast between attributes, as in study 1a (Supplementary Table 15). The results did not support the predicted mediation pattern: four contrasts were non-significant, and the remaining two were significant in the opposite direction to that expected. Consistent with this, Extended Data Fig. 4 shows that low IQ was rated as less controllable than the other three attributes despite other attributes being avoided to a greater extent, and the overall ordering of attributes by perceived controllability did not match their ordering by avoidance. Taken together, these findings indicate that perceived controllability does not provide a plausible explanation for the observed differences in avoidance across attributes in the Chinese sample.

Other potential mechanisms (USA and China). In addition to controllability, participants in the implantation arm rated each attribute on three further dimensions: generality (whether the negative consequences of the attribute affect many areas of life), medical relevance (whether the attribute represents a medical condition) and seriousness (whether the negative consequences of the attribute are serious). These were included to explore alternative explanations for attribute-level avoidance (Fig. 6). In the US sample, both perceived seriousness and perceived generality closely mirrored the ordering of avoidance across attributes.

To assess whether these dimensions help account for differences in avoidance, we conducted exploratory mediation analyses in the US sample using the same pairwise specification as in the controllability analyses, estimating mediation separately for each contrast between attributes (Supplementary Tables 16–18). These analyses suggest that seriousness provides the strongest candidate explanation: four of the six pairwise contrasts yielded positive and statistically significant average causal mediation effects (ACMEs), all in the predicted direction. Generality showed a weaker but still broadly supportive pattern, with three contrasts significant in the predicted direction. By contrast, medical relevance did not show a consistent mediation pattern: only one contrast was statistically significant in the predicted direction. Taken together, these results suggest that perceived seriousness, and to a lesser extent perceived generality, can partially account for some avoidance differences across attributes, whereas medical relevance was not supported as a general mechanism for avoidance.

Exploratory inspection of the Chinese data (Extended Data Fig. 5) suggests both similarities and differences. As in the US sample, seriousness tracks the ordering of avoidance across attributes, whereas generality shows a different ordering—most notably with low IQ perceived as having the most general life impact. Pairwise mediation analyses in the Chinese sample (Supplementary Tables 19–21) likewise suggest that seriousness is the strongest of the three alternative mechanisms, with three of the six contrasts yielding significant ACMEs in the predicted

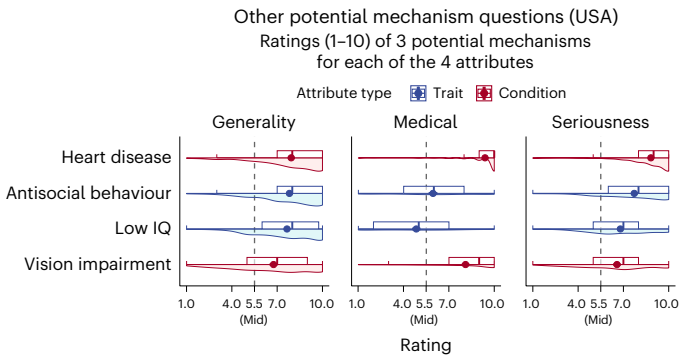


Fig. 6 | Exploratory ratings of potential psychological mechanisms (US sample). Distributions of participant ratings (1–10 scale) for three candidate mechanisms—generality (extent to which the attribute affects many areas of life), medical relevance (extent to which the attribute is considered a medical condition) and seriousness (severity of potential consequences)—assessed for each of the four attributes. Higher values indicate greater perceived generality, medical relevance or seriousness. Violin plots show the full response distribution; box plots show median (centre line), interquartile range (box bounds, 25th–75th percentile) and whiskers extending to the scale limits (1 and 10); points show means with error bars indicating 95% CIs (±1.96 s.e., clustered by participant). Dashed vertical lines indicate the scale midpoint (5.5). Medical conditions are shown in red; non-medical traits are in blue. Each participant rated each assigned attribute once, so $n = N$ per attribute: heart disease = 355, antisocial behaviour = 357, low IQ = 378 and vision impairment = 374 participants (implantation arm, US sample). Each participant was randomly assigned two of the four attributes and rated both. These measures were collected for exploratory purposes and were not preregistered as mediators.

direction. Evidence for generality was more limited and mixed, with only two significant contrasts, one of which was in the opposite direction to that predicted. Medical relevance again showed little evidence of statistical mediation.

Discussion

PGT is increasingly accessible and is now used to assess polygenic risks for a widening range of medical conditions beyond severe monogenic disorders and, in some cases, including non-medical traits[6,18,19]. It is important to distinguish between PGT for monogenic disorders, which has an established clinical evidence base, and polygenic testing, the predictive validity of which remains considerably more limited. This study focuses specifically on public responses to this latter form of testing because it is primarily through polygenic PGT that there will be substantial trade-offs between different traits and conditions, which in turn should be informed by the relevant public preferences examined in our study. Several major professional bodies—including the European Society of Human Genetics and the European Society of Human Reproduction and Embryology—have issued statements cautioning against the routine clinical use of polygenic PGT[20,21], and medical professionals' views on its current use are generally more cautious than the public support documented in the survey literature reviewed later[22,23]. Across countries, regulation of this expansion has largely relied on a distinction between medical and non-medical uses of testing[24,25], and public-attitudes research has broadly mirrored this pattern, documenting greater acceptance of testing for serious medical conditions than for non-medical traits[11–15]. However, most of this work focuses on whether people expressly approve of or are willing to obtain particular kinds of genetic information, rather than whether they would actually use such information if it were available[1,14,17].

This distinction matters because willingness to seek information and downstream decision-making may reflect different psychological processes and normative considerations[16,26]. Moreover, clinics and commercial providers sometimes present prospective parents with genetic information without asking them to select which attributes to test for[27]. Motivated by this, we designed this study to disentangle attitudes towards testing from implantation decisions based on genetic information. Participants were asked to report either their willingness to obtain information about the likelihood that embryos would develop different conditions and traits, or their decision to implant embryos with different genetic profiles, where this information was already available.

Our findings qualify the apparent consensus in previous work. Although participants expressed greater willingness to test for medical conditions than for non-medical traits (H2), more than half were still willing to test for the two non-medical traits (H1). More importantly, once genetic information was available, the medical–non-medical distinction did not carry over to implantation decisions (which were made in a forced-choice paradigm). In both the US and Chinese samples, participants avoided embryos with above-average chances of undesirable non-medical traits (H3 and H5a), and this avoidance was at least as strong as avoidance of adverse medical conditions (H4 and H5b). Exploratory analyses further showed that low IQ and antisocial behaviour were generally treated as close to heart disease in terms of overall avoidance and in direct trade-offs in both countries (Figs. 3a,b and 5 and Extended Data Fig. 2). However, the preregistered hypothesis regarding the drivers of such avoidance was not supported: the perceived controllability of the negative consequences of the condition did not mediate these avoidance patterns (H6).

The lack of support for H1 departs from both our pilot A study (see 'Pilot data' in Methods for details) and previous work[14], which found that only a minority of participants expressed willingness to test for non-medical traits. One plausible explanation is attribute framing. Pilot A and earlier studies focused on positively framed traits (for example, high IQ and prosocial behaviour), whereas study 1a examined negatively framed traits (for example, low IQ and antisocial behaviour). Another possibility is that willingness to undergo polygenic testing has increased over time, as familiarity with genetic testing and public discussion of its uses have expanded; pilot A and the study reported in ref. 14 were conducted more than 18 months earlier. Consistent with this interpretation, willingness to test for vision impairment and heart disease—both framed similarly in pilot A—also appeared higher in this study (54.8% and 78.6%, respectively) than in the pilot (46.7% and 70.6%, respectively).

Our findings also suggest that current regulatory frameworks may not map well onto how genetic information is used once it is available. Across countries, governance frameworks commonly rely on broad distinctions between medical and non-medical uses of genetic testing, often operationalised through terms such as serious disease, medical indication or medical necessity[24,25]. Previous survey-based work suggests that these distinctions also shape attitudes towards testing, with higher acceptance of testing for medical rather than non-medical attributes[14]. Our results are partly consistent with this pattern at the level of willingness to test. However, they also show that the relevance of these category-based distinctions does not extend to implantation decisions. Once information is available, participants' choices are guided more by evaluations of specific attributes than by whether those attributes are classified as medical or non-medical. This is broadly consistent with a welfarist approach to regulation, under which what matters is the effect of a condition or trait on well-being rather than its category label[28].

It is important to note that the non-medical traits and medical conditions examined here were not intended to represent the full range of PGT applications. Rather, they were selected as plausible test cases for the medical versus non-medical distinction. Low IQ and antisocial behaviour may be viewed as adjacent to clinically recognised conditions, such as intellectual disability or psychiatric and behavioural disorders. However, participants did not view these traits as medical to the same extent as the medical conditions in our scenarios, as

reflected in exploratory classification (Fig. 6, middle). As such, these two non-medical traits still provide a useful test of whether medical and non-medical categories map onto how genetic information is actually used in concrete implantation choices. Our results suggest that they do not do so clearly.

In addition, our results show that the established psychological distinction between decisions to avoid information and decisions made based on that information, if available, is important in the context of PGT[16,26]. Much of the existing literature assumes that the tendency to seek genetic information and the decision to act on it are aligned. Our design allowed these stages to be separated. Willingness to test captures abstract judgements about what kinds of selection are appropriate to endorse, whereas implantation decisions reflect responses to explicit trade-offs affecting a future child. This helps explain why avoidance of medical over non-medical traits was strong in attitudes towards testing (H2) but not implantation decisions (H4). One interpretation is consistent with accounts of deliberate ignorance, in which individuals avoid acquiring information that carries moral or psychological burden while still responding strongly once such information is available[16].

With respect to underlying mechanisms, we preregistered perceived controllability, or lack thereof, as a candidate mediator of avoidance behaviour (H6), given prior evidence that perceptions of whether the consequences of a condition can be managed are linked to stigmatization, anger and reduced willingness to provide support in some domains, although not in others[29,30]. Contrary to this hypothesis, perceived controllability did not mediate avoidance behaviour in our study. Exploratory analyses instead suggested closer alignment with dimensions such as perceived seriousness and the generality of anticipated life consequences. In other words, when deciding whether to avoid certain conditions, participants seemed to consider the seriousness of their consequences, regardless of whether these could be controlled or managed in practice. This pattern is consistent with the broader finding that category-based distinctions between medical and non-medical attributes may be less informative than how consequential particular attributes are perceived to be. Nevertheless, these observations are descriptive and should be interpreted cautiously, as the study was not designed to adjudicate between competing psychological mechanisms.

These findings have different implications in different regulatory contexts. In more permissive, market-driven systems such as the USA, where commercial providers have broad discretion over which traits to offer, the key issue may be the scope and framing of the information presented to prospective parents. Our results suggest that once such information is made available, it can shape implantation decisions even when testing is formally optional. This has direct implications for informed consent: if providers routinely present unsolicited trait information in a uniform format regardless of predictive validity, the mere act of signing a consent form may not constitute meaningful informed choice, and the architecture of information presentation may do more to shape reproductive decisions than any explicit policy or guideline. In more restrictive, indication-based systems, such as those found in the UK, China and many European countries[24,25,28,31], the challenge may lie in the bluntness of category-based boundaries that do not match how particular attributes are evaluated in practice. As the commercial landscape of PGT expands, with a growing number of providers offering polygenic scores across a wider range of attributes[7–10], decisions about which attributes are included in testing panels and how they are described may play an increasingly important role in structuring reproductive choices.

Several limitations should be noted. First, our study focused on four attributes selected to span a wide range of perceived impact, rather than to represent the full space of medical conditions and non-medical traits. As a result, the findings should not be interpreted as establishing general rules about all traits or conditions, but as demonstrating the existence of plausible counterexamples to category-based distinctions. Second, the scenarios necessarily simplified complex clinical realities, including uncertainty around prediction accuracy and non-uniform percentage point deviations from average chances. In practice, within-family polygenic scores for behavioural traits such as antisocial behaviour and low cognitive ability are subject to greater attenuation of predictive power, meaning that expected 'gains' from embryo selection are smaller for behavioural outcomes than our uniform presentation implies[7]. Third, the study measured stated preferences and choices in hypothetical scenarios rather than real-world behaviour, which may be shaped by additional emotional, relational and institutional factors. Fourth, although the inclusion of participants from both the USA and China provides some cross-national perspective and follows a parsimonious strategy of sampling culturally distant contexts[32], the findings do not capture the full diversity of regulatory environments or stakeholder views. Fifth, although the study found no evidence for perceived controllability as a mediating mechanism in preregistered mediation analyses and identified alternative mechanisms—such as perceived seriousness and the breadth of anticipated life consequences—as potentially relevant, these were examined only in exploratory analyses and were not directly manipulated. It is worth noting that directly manipulating controllability salience (as was planned for study 2) may yield different results from measuring abstract perceptions as mediators. Future research could test this. Sixth, in the implantation task, participants made forced choices based on attribute information provided to them, regardless of whether they would have sought that information. Participants may therefore treat it as decision-relevant even if they would not seek it out or endorse its use outside the experimental context. Several features of our design mitigated this concern, including trade-off dilemmas requiring prioritization, the explicit instruction that participants were not required to use all information, and the fact that some attributes had no significant effect on decisions in pilot studies (see more details in Methods, 'Task 2'). Finally, our study reveals only public preferences. Consideration of ethical principles, theories and concepts would be necessary to form justifiable public policy[33].

Future research could extend this work by addressing some of these limitations, such as examining a broader set of attributes, including those that fall closer to or further from existing regulatory boundaries, and by incorporating additional countries with distinct governance frameworks. Studies involving other stakeholders, such as clinicians, genetic counsellors and regulators, would further clarify how institutional contexts shape decision-making. Experimental designs that more closely approximate clinical settings, including sequential testing and selection decisions or varying presentation formats, may also help to identify mechanisms underlying observed behaviours. Such work would contribute to a more detailed understanding of how emerging reproductive technologies interact with psychological judgement and regulatory practice. Ultimately such preferences would need to be integrated with ethical considerations in revising public policy.

Methods

All deviations from the methods and results can be found in the 'Deviations from registered protocol' section. The data, code and other materials can all be found at https://osf.io/vb9c2.

Ethics information

These studies were reviewed and approved by the Social Sciences and Humanities Interdivisional Research Ethics Committee (SSH IDREC) at the University of Oxford (protocol ID: R80692), and by the Research Ethics Board of the Center for Psychological Sciences at Zhejiang University, China (protocol ID: 2025-002). The research complied with all relevant ethical regulations and informed consent was obtained from all human participants. Participants were compensated financially for their participation in accordance with national minimum wage and/or the standard rates of the online platforms.

Pilot data

To assess feasibility and refine task design, we conducted three pilot studies with US-based participants (N = 280, 289 and 465 after exclusions). The pilot studies were approved by the SSH IDREC at the University of Oxford (protocol ID: R80692). Pilot A examined attitudes towards polygenic testing for 12 attributes (6 conditions and 6 traits), showing that participants were more willing to test for medical conditions than non-medical traits and viewed such testing as more morally and legally acceptable. Pilot B used a discrete choice format where participants selected between embryos differing in chance levels of one condition and one trait. Results showed that participants were more sensitive to undesirable chance levels (for example, increased risk of a condition) than to desirable ones for medical conditions, and that moral trade-offs varied across attribute pairings. Pilot C was done using more granular chance levels and probability descriptions (for example, 2%, 5% and 10% versus 5%, 10% and 25%). We found that participants were more likely to avoid undesirable chance levels for antisocial behaviour and high IQ (both non-medical traits) than for vision impairment, but their responses did not differ significantly from those for heart disease. We also found that presenting large versus small risk differences had minimal impact on decision patterns. These studies informed key design decisions for studies 1a and 1b, such as using a uniform ±5 percentage point risk difference to enhance interpretability and internal validity. Full pilot methods, results (including inferential statistics) and comprehension data are provided in the Supplementary Information.

Design

Overview. We planned to run three studies in sequence. The first two studies (study 1a and study 1b) were similar in terms of design and content, with an important difference that one was using US-based participants (study 1a) and an additional task arm, while the other was using China-based participants (study 1b). Data from study 1a were used to test hypotheses H1–H4. In addition, study 1a was also used to test hypothesis H6 using a measurement-of-mediation design. Study 1b was used to test hypothesis H5 (equivalent to H3 and H4 with Chinese participants). The third study (study 2) was planned to be used to test for H6 in a concurrent double randomization design, combining measurement-of-mediation and manipulation-of-mediator designs[34,35]. Given that study 1a was used to explore the possibility of a mediator, whether to run study 2 was conditional on establishing statistical evidence of mediation in study 1a. Due to the substantially larger sample size required for study 2, our plan was to only conduct this follow-up in the US sample to ensure feasibility. As study 1a provided no evidence for a mediating role of non-controllability, study 2 was not run. Supplementary Fig. 16 illustrates the design of the three planned studies and Extended Data Fig. 1 illustrates the design of the two studies.

Study 1a. After passing a Qualtrics-provided reCAPTCHA, providing their consent, reading instructions and answering comprehension questions, participants were randomly allocated to one of two arms: (1) the willingness task arm or (2) the implantation task arm. For the willingness task arm, participants were assigned a random order for the four attributes that applied across tasks. Participants were presented with four tasks in the following order: (1.1) willingness questions, (1.2) morality questions, (1.3) legality questions, and (1.4) only-available-embryo questions. The second, third and fourth tasks were added for exploratory purposes and to ensure comparable completion times across the two arms. For each task, participants responded to each of the four attributes on separate pages in the randomised order assigned to them. Participants therefore completed tasks across 16 pages (4 attributes × 4 questions). In the implantation task arm, participants were randomly assigned two attributes in a random order (with this random order being fixed across tasks for the same participant). They were then presented with three tasks in the following order: (2.1) implantation questions, (2.2) focal mechanism question, and (2.3) exploratory mechanism questions. All participants answered demographics questions as the final task. Participants faced two attention checks, one right before the first task, and one right before the demographics task. As per Prolific rules for surveys over 5 minutes, participants can be excluded only after failing two attention checks. For participants who failed both checks, the experiment was terminated, while those who passed at least one of the attention checks moved to the third and fourth tasks. Participants also answered two comprehension questions after reading the instructions.

Study 1b. This study had two main differences from study 1a. First, it did not include the willingness task arm. Second, it was run with China-based participants. After passing a Qualtrics-provided reCAPTCHA, providing their consent, reading instructions and answering comprehension questions, participants were randomly assigned two attributes in a fixed order in which these attributes appeared across the tasks. They were then presented with four tasks in the following order: (1) implantation questions, (2) focal mechanism question, (3) exploratory mechanism questions, and (4) demographics questions. All other aspects of the procedure were identical to study 1a.

Study 2. This study was planned to allocate participants randomly into one of three groups. All participants were to follow similar steps as in the implantation task arm of study 1a (or just like study 1b). The goal of task 3 (focal mechanism question) in study 2 was to perform measurement of mediation in one group and a manipulation check in the two other groups. Nevertheless, on the basis of the results from study 1a, this study was not run.

Experimental design. *Study 1*. Four attributes (two medical conditions and two non-medical traits) were ordered randomly for each participant. For the willingness task arm, participants were asked about all four attributes using this random order. For the implantation task arm, participants were randomly assigned two attributes in a fixed order in which these attributes appeared across the tasks. This could have been two undesirable non-medical traits, two medical conditions or one of each. The following are the four attributes:

Medical conditions:

- Heart disease: Health problems affecting the heart.
- Vision impairment: Visual problems affecting the ability to see clearly.

Non-medical traits:

- Low IQ: Low cognitive ability and intellectual functioning.
- Antisocial behaviour: Disruptive or harmful social behaviour.

We do not assume that the selected traits and conditions are representative of their broader categories. Rather, we examined whether people evaluate some non-disease traits differently from some disease traits when making hypothetical embryo selection decisions, reflecting how public reasoning may reinforce or challenge current policy distinctions between medical conditions and non-medical traits.

Study 2. Our concurrent double randomization design was planned to target the effect of controllability on implantation decisions by either increasing (enhancement manipulation) or decreasing (blockage manipulation) the mediator's effect compared with the control case in which the mediator is allowed to vary freely[35,36]. To do this, participants were planned to be randomly assigned to one of the following three groups:

(A) Mediator varies freely (control): This would be identical to the implantation task arm of study 1a.
(B) Blockage manipulation: This would be identical to the implantation task arm of study 1a, except that the description of attributes would emphasise that they all have high controllability.

(C) Enhancement manipulation: This would be identical to the implantation task arm of study 1a, except that the description of attributes would emphasise that they all have low controllability.

Here is the description of attributes in the three conditions:
(A) Vary freely (control): similar to Study 1.
(B) Blockage (high controllability; changes from study 1 are in italic font):

Medical conditions:

- Heart disease: Health problems affecting the heart, *often manageable with lifestyle changes and medication*.
- Vision impairment: Visual problems affecting the ability to see clearly, *often manageable with glasses or contacts*.

Non-medical traits:

- Low IQ: Low cognitive ability and intellectual functioning, *often manageable with specialised education and support*.
- Antisocial behaviour: Disruptive or harmful social behaviour, *often manageable with medication or therapy*.

(C) Enhancement (low controllability; changes from study 1 are in italic font):

Medical conditions:

- Heart disease: Health problems affecting the heart, *hard to manage even with lifestyle changes and medication*.
- Vision impairment: Visual problems affecting the ability to see clearly, *hard to manage even with glasses or contacts*.

Non-medical traits:

- Low IQ: Low cognitive ability and intellectual functioning, *hard to manage even with specialised education and support*.
- Antisocial behaviour: Disruptive or harmful social behaviour, *hard to manage even with medication or therapy*.

Tasks. On the basis of the study and condition, participants completed a subset of the following tasks:

Task 1 (willingness question). Participants read a question asking them about their willingness to perform polygenic testing for the purpose of making decisions on which embryos to implant (Supplementary Fig. 1). This question was repeated four times; once for each of the four attributes.

Task 2 (implanting questions). Participants read 10 dilemmas (randomly sampled from 36 possible dilemmas) on which of 2 embryos to implant (A or B) based on attributes and chance levels. We used dilemmas that traded off the two sampled attributes. For chance levels, we used three levels:

- Average chances.
- Significantly above-average chances: an absolute increase of 5 percentage points in chances compared with average chances.
- Significantly below-average chances: an absolute decrease of 5 percentage points in chances compared with average chances.

It is worth noting the following on our design for task 2:

- We selected a uniform difference of ±5 percentage points for all attributes to enhance comparability across dilemmas and minimise potential confounds arising from varying base rates. While this does not reflect real-world settings, where absolute risk differences vary substantially across attributes—ranging from around 1% for many traits to closer to 10% for some conditions—we believe this simplification supports internal validity by ensuring that observed differences in responses reflect participants' value judgements rather than statistical artefacts or scaling effects.
- We present differences in absolute percentage points relative to average chances (for example, '5 percentage points higher than average chances') following formats used by some polygenic screening providers (for example, Nucleus Genomics). This framing also improves clarity and comprehension in our online task, as supported by results from pilot A and pilot C (Supplementary Information).
- While traits and conditions differ in predictive accuracy at present, our study was designed to elicit how participants value different attributes under uniform presentation, not how they evaluated the technical reliability of each score. Introducing predictive accuracy estimates would risk confounding these value judgements with perceptions of test quality. Moreover, predictive accuracy for polygenic scores is evolving rapidly and may change substantially over time. This approach also aligns with current practice in reports from some providers, such as Orchid, which present risk estimates without disclosing predictive accuracy figures.

Each dilemma involved one of the three levels above for each of the two embryos and the two attributes, with the restrictions that the two embryos could not have the same chance level for the same attribute. This created 3 (chance levels for embryo A) × 2 (chance levels for embryo B) pairings for each attribute, thus 36 possible dilemmas. An example of these dilemmas is presented in Supplementary Fig. 2, and the dilemmas were preceded by the following preamble:

> "Imagine that genetic testing of the only two viable embryos has been completed. And imagine that in this scenario, the clinic did not ask you which attributes to test for, but has provided the following information regardless. You must decide which of the two embryos you want to implant and which you will not implant, based on the likelihood of certain conditions and traits (revealed by the genetic testing). Apart from the presented characteristics, you can assume that the two embryos are otherwise identical and equally likely to be successfully implanted. This is the type of decision you will be asked to make in every scenario. You may use any of the information provided in each scenario to inform your decision—but you are not required to use all of it."

This was followed by the implantation question (Supplementary Fig. 2), which was repeated ten times with different chance levels.

This design was appropriate for several reasons. First, these dilemmas closely resemble real-life implantation decisions: as argued in ref. 37, users of polygenic embryo screening have to trade off different attributes when deciding which embryos to implant. Second, the use of different chance levels (with average chances as control) allows for an estimate of desirability of each attribute in terms of 'expected gain' (difference from embryo selected at random without the use of testing), a measure that is argued in ref. 7 to be a more realistic way to capture a user's preference as opposed to questions that compare two embryos with extreme scores. It also provides a better measure than directly asking about preference to implant an embryo based on possessing high or low chances of specific attributes, which may be harder to interpret.

One potential concern was that participants may feel encouraged to use all presented information. However, our design mitigated this in several ways. The dilemmas often involved trade-offs—such as one embryo having a higher chance of a condition but a lower chance of a trait—requiring prioritization rather than uniform endorsement. Our pilot studies also showed that participants did not always act on all information: for instance, some attributes had no significant effect on decisions, even when highlighted. Moreover, our design allowed for internal comparison of the relative influence of different

attributes (for example, traits versus conditions), which helped interpret responses even if participants engage with the information to some degree. We also explicitly informed participants that they are not required to use all the information provided when making their decision.

Task 3 (focal mechanism question). We planned to study mechanisms using measurement-of-mediation and manipulation-of-mediator designs. In study 1a and study 2, we planned to ask the same questions. In the case of study 1a, the measurement-of-mediation design was employed and considered exploratory and was used to determine the need for study 2. It showed that study 2 was not needed. In study 2, the measurement-of-mediation design was planned to be employed with the control group, while a manipulation-of-mediator design would have been employed with each of the two other groups (compared with the control group), in which case the mechanism questions would have served as a manipulation check.

(1) Non-controllability: For each of the two sampled attributes, participants were asked whether they believed that potential negative consequences could be controlled or managed; see Supplementary Fig. 3 for full wording.

Task 4 (exploratory mechanism questions). For exploratory purposes, we also asked three other questions about other potential mechanisms, such as generality, seriousness of harm and medical status of attributes.

(2) Generality: For each of the two sampled attributes, participants were asked whether they believed that potential negative consequences would affect many areas of a person's life (see Supplementary Fig. 4 for full wording).
(3) Seriousness of harm: For each of the two sampled attributes, participants were asked how serious the potential negative consequences would be (see Supplementary Fig. 5 for full wording).
(4) Perceived medical nature: For each of the two sampled attributes, participants were asked whether they represent a medical condition (see Supplementary Fig. 6 for full wording).

Task 5 (morality question). For exploratory analysis, participants were asked to report their views on the morality of performing polygenic testing for the purpose of making decisions on which embryos to implant (see Supplementary Fig. 7 for full wording). A similar question was used in pilot A. This question repeated four times; once for each of the four attributes.

Task 6 (legality question). For exploratory analysis, participants were asked to report their views on whether there should be laws banning polygenic testing for the purpose of making decisions on which embryos to implant (see Supplementary Fig. 8 for full wording). A similar question was used in pilot A. This question repeated four times; once for each of the four attributes.

Task 7 (only-embryo question). For exploratory analysis, participants were asked to report their views on whether implanting should be performed when only one embryo is available (see Supplementary Fig. 9 for full wording). This question repeated four times; once for each of the four attributes.

Task 8 (demographics questions). We asked questions about participants' gender, age, number of children under 18, household income, education, employment status, whether they or their partners underwent in vitro fertilization and PGT, and whether they have done surveys similar to this one.

Finally, participants were asked a debriefing question in the form of a free text box.

Translations. As one of the locations of participants was China, a non-English speaking country, we translated materials following a standard forward- and back-translation procedure[38,39].

Attention checks. Two attention checks were used. By the rules of Prolific, surveys that last more than 5 minutes should include at least two attention questions and participants can only be excluded if they fail both questions. Prolific rules also mandate that the answer to the attention question will have to be presented in the same page. We asked the following two questions (the first appeared right before the first task, and the second appeared right before the demographics task), with the order of response options randomised:

Attention check 1: At the start of the next section, you will see the first set of questions. To get to the next page, you should answer Blue from the options below.

Based on the text above, which eye colour is the rarest?

- Brown
- Blue
- Green
- Grey
- Hazel
- Amber

Attention check 2: In studies like ours, there are sometimes a few people who do not carefully read the questions they are asked and just quickly click through the survey. These random answers are problematic because they compromise the results of the studies. It is very important that you pay attention and read each question. Please answer YouTube in the question below to show that you read our questions carefully (and regardless of your own opinion).

When an important event is happening or is about to happen, many people try to get informed about the development of the situation. Based on the text above, where do people get their information from?

- TV
- X (formerly Twitter)
- Facebook
- YouTube
- Reddit
- Radio
- Newspaper
- TikTok
- Other

Comprehension checks. Two comprehension checks were used. The correct answer was provided after. These were used for robustness. The questions appeared directly after the instructions.

Comprehension check 1: Imagine the following situation. Typically, 20 individuals out of 100 experience depression (a chance of 20% on average). A polygenic risk score indicates that an embryo has an absolute increase of 10 percentage points in the chance of depression compared with the average chance. Out of 100 people with this polygenic risk score, how many should be expected to experience depression? Among the choices below which number is most likely to be correct?

- 100
- 12
- 20
- 30

Comprehension check 2: What does preimplantation genetic testing (PGT) screen for?

- The mother's genetic makeup
- The genetic traits of the embryo before it is implanted
- The father's potential genetic diseases
- The blood type of the fetus

The following page showed the correct answers:
'The correct answers are:
Q1. The answer is 30. Since it says that the risk is increased, this is equal to the average risk (20%) plus the added risk from the genetic test (+10%).
Q2. The genetic traits of the embryo before it is implanted.'

Sampling plan

Participants. The study was completed online by participants based in the USA and China. US-based participants were recruited via Prolific while China-based participants were recruited via Credamo. We aimed to recruit samples that were nationally representative with respect to age, gender and ethnicity for the USA, and with respect to age and gender for China. We predicted that the oldest age group in China would prove to be difficult to recruit online, in which case we would prioritise statistical power over representativeness by combining the oldest two groups or three groups and noting this in the report. Participants were compensated financially for their participation in accordance with national minimum wage and/or the standard rates of the online platforms. For US-based participants, sampling took place over 1 week. As expected, most of the sample was recruited in the first few days. As predicted, the remaining participants from underrepresented categories required a longer time, and the sampling strata were relaxed when the corresponding quotas were not achieved after the planned 1-week sampling period. According to Prolific, 'most representative samples reach at least 90% completion within 48 hours and fill completely after loosening requirements. This typically results in a sample accuracy of 95% or higher[40].' For Chinese participants, sampling took place over an initial 2-week period. After this stage, sampling strata were relaxed (as planned). Recruitment was handled by the survey platform (Credamo), which resulted in additional requirements. For example, Credamo requires a rejection rate below 30%. We balanced such platform requirements with our own sampling criteria, ensuring that the main analysis includes only participants who answered both attention checks correctly.

Expected effect sizes. This section and the two following ones made estimates based on the three pilots mentioned earlier and further described in Supplementary Information. These pilot studies were not identical to the planned study but they provided a good approximation. See the list of caveats later.

H1. An estimate of −0.24 (standardised $d = -0.65$) was calculated from pilot A as a coefficient for the intercept of trait on PGT decisions. This number corresponds to 26% of participants (0.26 − 0.5 = −0.24) expressing willingness to perform testing for two non-medical traits (high IQ and prosocial behaviour). This estimate is consistent with other studies like ref. 14 (which included 12 non-medical traits) showing a range of 19–37% approval for testing for all non-medical traits (standardised $d = [-0.78, -0.28]$). It is also consistent with the effects when taking into account all six traits tested in pilot A, which showed a range of 13–29% approval for testing (standardised $d = [-1.09, -0.45]$).

H2. An estimate of −0.33 ($d = -0.76$) was calculated from pilot A as a coefficient for the effect of attribute type (medical, non-medical) on PGT decisions. This number corresponds to a difference in the percentage of participants willing to perform testing for two medical conditions (59%) and non-medical traits (26%). This estimate is consistent with other studies like ref. 14 (which included 12 medical conditions and 12 non-medical traits) showing a range of 18–60% difference in approval for testing between medical conditions and non-medical traits (standardised $d = [-1.5, -0.39]$). This range is consistent with the full results of pilot A that included six conditions and six traits, with a range of 1–60% difference in approval for testing between any medical condition and non-medical trait (standardised $d = [-1.24, -0.02]$; approval for testing for 'well-being' was similar to 'high IQ', while $d = -0.9$ for comparing heart disease with high IQ).

H3. Estimates of −0.43 ($d = -0.71$) and −0.16 ($d = -0.48$) were calculated from pilot C and pilot B, respectively, as a coefficient for the effect of chance level. This number corresponds to the difference in probability of choosing an embryo when it has significantly above-average chances of having a non-medical trait and the probability of choosing an embryo when it has an average chance of having that trait. The negative sign means that the probability is higher for average chances. The discrepancy between pilot C and pilot B is most likely due to using antisocial behaviour in pilot C and reverse-coded prosocial behaviour in pilot B. As such, pilot C provides a more precise estimate here.

H4. Estimates of −0.13 ($d = -0.14$) and 0.26 ($d = 0.56$) were calculated from pilot C and pilot B, respectively, as a coefficient for the interaction of chance level and trait type. This number corresponds to the difference between medical conditions and non-medical traits in the difference in probability of choosing an embryo when it has above-average chances versus average chance. The negative sign means that the difference in probability is lower (higher in absolute terms) for non-medical traits. The discrepancy between pilot C and pilot B is most likely due to using antisocial behaviour in pilot C and reverse-coded prosocial behaviour in pilot B. As such, pilot C should be trusted more.

H5. There were no estimates for H3 and H4 from Chinese participants in our pilots. Given that China is considered a collectivist society with a focus on interpersonal harmony[41,42], we expected larger estimates given that preference to avoid offspring with above-average chances of antisocial behaviour could be higher.

H6. There were no estimates from pilots or other studies. Some participants in pilots B and C noted that their lower avoidance of visual impairment was driven by the possibility of it being mitigated or fixed. The description of this attribute was 'difficulty in seeing, ranging from blurriness to blindness'.

Sample size. The sample size had been determined through a cost–benefit analysis, taking into account the resources available, the expected effect sizes (see 'Expected effect sizes'), and power analyses to confirm that the planned sample size would allow for the detection of theoretically significant effect sizes (see 'Power analysis'). The minimum final sample sizes needed for the studies were:

- Study 1a: 1,200 US-based participants (600 for each arm; allocated randomly)
- Study 1b: 600 China-based participants
- Study 2: 1,800 US-based participants (600 for each arm; allocated randomly)

Accounting conservatively for exclusion rates of up to 20% (see 'Exclusions'; the highest exclusion rate in the 3 pilots was 7%), we anticipated collecting data from a total of 4,500 participants, divided as follows: 1,500, 750 and 2,250. To account for any exclusion rate higher than 20%, our plan was as follows:

- We first recruit the planned number—1,500 for study 1a, 750 for study 1b and 2,250 for study 2.
- After data collection, we will assess the actual exclusion rate.
- If the number of retained participants is lower than 600 for any arm (two arms in study 1a, one in study 1b and three in study 2), we would recruit additional participants in batches of 50 until we reach our planned sample size of 600 valid participants per arm.

Power analysis. Power analyses had been conducted using simulations with bootstrap resampling[43] developed in R using data from the three

pilot studies (Supplementary Information) and the planned models (calculated using R package lfe). Power estimates were also replicated using the R package simr[44], and employing mixed-effect (lmer) models, replacing clustered standard errors with random effects (simr works with lm/lmer but not with lfe). We used 1,000 simulations, and estimates of means and variances from pilots A (H1 and H2), B (H3 and H4) and C (H3 and H4). The caveats of these estimates are given later. Following ref. 45, we considered a strict level of $\alpha = 0.005$ and following ref. 46 we used family-wise correction to account for dependencies within the following families of hypotheses: family 1 (H1 and H2) and family 2 (H3 and H4). Given this, we used α levels of 0.005/2 for hypotheses H1–H4. For H4, we used $d = 0.2$ as a threshold for the inferiority test.

In the studies, following ref. 45, we use 'supportive evidence' and 'suggestive evidence' to interpret results passing an initial α of 0.005 (0.0025 or 0.005) and 0.05 (0.025 or 0.05), respectively. Correction was only applied to family 1 (H1 and H2) and family 2 (H3 and H4). No correction was applied to family 3 (H5) and family 4 (H6).

Our analysis had shown that the aimed samples of 600 participants (study 1a) would be able to detect with over 95% power and a corrected α level of 0.0025 a minimal standardised effect of $d = -0.2$ for H1–H4. Following ref. 47, we chose this value since it is (1) lower than the effect size found in previous studies for H1 and H2, (2) lower than the effect size found in our pilots for H1–H3, and (3) it is conventionally considered a small effect size[48]. The negative d here is chosen due to the nature of the hypotheses, which are all expected to yield a negative effect size (with the exception of H4, which allows for positive values up to +0.2). It represents 40%, 27% and 42% of the smallest estimated effects for H1–H3. It represents 143% of the estimated effect for H4, which is acceptable given that this hypothesis is framed for inferiority testing.

Our analysis did not provide estimates for H5 and H6. Data for Chinese participants assumed higher estimates for H3 and H4 compared with US-based participants (thus higher estimate for H5). For H6, the choice of 1,800 participants was also based on the choice of 600 participants for study 1a, given that 1 of the 3 groups (control) was equivalent to study 1a.

Caveats of making estimates based on the pilots. *Pilot A (H1 and H2).*

- Attributes: This pilot used six conditions and six traits. For the purpose of power analysis, only the most relevant four were included: heart disease, vision impairment, high IQ and prosocial behaviour. The last two had opposite framing (positive) from the planned study. The description used for the first two was also slightly different from what is planned.

Pilot B (H3 and H4).

- Attributes: This pilot used six conditions and six traits. For the purpose of power analysis, only the most relevant four were included: heart disease, vision impairment, high IQ and prosocial behaviour. The last two had opposite framing (positive) from the planned study; their values were reversed. The description of the first two was also slightly different from what is planned.
- Pairing of attributes: This pilot sampled one medical condition and one non-medical trait for each participant. Given this, unlike the planned design, there were no condition–condition or trait–trait pairs in the data. This could have created unfair comparisons between traits and conditions. To simulate a scenario where each of the four attributes is paired with all three of the other attributes, the power calculation included six attributes (the four closest and two additional ones: 'Alzheimer's disease' and 'aptitude for exercising'). These two additional attributes were recoded into one of the attributes of interest to align with the planned study. For example, cases where high IQ was paired with Alzheimer's disease were recoded as high IQ being paired with prosocial behaviour. A similar approach was applied to the other three attributes of interest. Instances of Alzheimer's disease and aptitude for exercising pairings were reassigned randomly to one of the planned trait–condition pairings. This adjustment aimed to approximate the structure of the planned study to allow for minimal effect estimation.
- Sex of embryos was always varied. Each dilemma posed a male embryo versus female embryo. For the purpose of power analysis, the data were pooled over embryo sex.
- Each participant answered 15 dilemmas as opposed to 10 dilemmas in the planned study.
- Percentages for chance levels were not mentioned.

Pilot C (H3 and H4).

- Attributes: This pilot used two conditions and two traits (similar to the planned study): heart disease, vision impairment, high IQ and antisocial behaviour. The third one had opposite framing (positive); its values were reversed. The description of the other three was also slightly different from what is planned.
- Pairing of attributes: This pilot sampled one medical condition and one non-medical trait for each participant. Given this, unlike the planned design, there were no condition–condition or trait–trait pairs in the data. This could have created unfair comparisons between traits and conditions.
- Sex of embryos was always varied. Each dilemma posed a male embryo versus a female embryo. For the purpose of power analysis, the data were pooled over embryo sex.
- Chance levels: We used seven chance levels instead of four chance levels: significantly above average, moderately above average, slightly above average, average, slightly below average, moderately below average and significantly below average. For the purpose of power analysis, we retained only data for the significantly above-average, average and significantly below-average chance levels.
- Percentages for chance levels. Participants were allocated to one of two groups, each with different numbers representing absolute change in chance levels: large (25% for significantly, 10% for moderately and 5% for slightly) versus small (10% for significantly, 5% for moderately and 2% for slightly). For the purpose of power analysis, the data were pooled over this factor.
- Each participant answered 15 dilemmas as opposed to 10 dilemmas in the planned study. In power analysis, this has been accounted for by removing rows corresponding to moderately and slightly above- and below-average chances, resulting in an average of two dilemmas per participant in the dataset.

Inclusions and exclusions. Participants were included if they met all of the following criteria:

(1) Location: USA for study 1a and China for study 1b
(2) Language: fluent in English for study 1a and fluent in Mandarin for study 1b
(3) Consent: provision of consent to the study
(4) Age: confirmation that they were 18 years or older

Before data collection: participants who participated in the pilot studies of this project or some of the other pilot studies by our team on Prolific were excluded from taking our survey on Prolific (relevant for study 1a).

During data collection: To be compensated, participants had to (1) provide their platform ID, (2) complete the study (answering all questions), and (3) pass at least one attention check. Participants who did not meet all three criteria were excluded from completing the study and their data were removed.

During analysis: Participants who failed any attention check were excluded from the main analysis to ensure attentiveness. For robustness, we conducted a secondary analysis identical to the main ones, but including all participants to maintain sample representativeness. As an additional secondary robustness check, we excluded participants who failed any comprehension questions (in addition to those that failed any attention check), even though they were provided with correct answers after responding.

Analysis plan

Preprocessing. As described earlier, only participants who passed at least one attention check were paid and their data were retained. Among those, participants who failed any attention check were further excluded from the main analysis to ensure attentiveness.

Data preparation for hypotheses H3–H6. To fit the data for conjoint analysis[49], the data were reshaped as follows:

(1) Participant level to dilemma level: The data were transformed from the participant level (N rows) to the dilemma level ($10 \times N$ rows), where each dilemma represents a choice between two options.
(2) Dilemma level to choice level: The data were then reshaped to the choice level ($20 \times N$ rows), where each row corresponds to one of the two options (for example, embryo A or embryo B) in a given dilemma.
(3) Choice level to attribute-choice level: Finally, the data were further reshaped to the attribute-choice level ($40 \times N$ rows), where each row corresponds to one of the two chance level values of each of the two attributes for a single dilemma.
(4) Chance level to dummy columns: The 'chance level' column was expanded into two dummy variables:
   (a) Above_Average: 1 if the chance level is significantly above average; 0 otherwise
   (b) Below_Average: 1 if the chance level is significantly below average; 0 otherwise

The processed dataset contained one row for each combination of attribute and chance level for both options (for example, embryo A and embryo B) across the ten dilemmas per participant.

The key variables were as follows:

- ResponseId: The participant identifier, used for clustering standard errors.
- Attribute_Type: A binary variable indicating whether the attribute is a non-medical trait or a medical condition.
- Attribute_Name: A categorical variable with four values, corresponding to the four attributes (for example, vision impairment or antisocial behaviour).
- Above_Average, Average and Below_Average: Dummy variables indicating the chance level for each row.
- Choice: A binary variable indicating whether the embryo associated with the focal attribute-choice row was chosen (1 for chosen; 0 otherwise).

See Supplementary Fig. 10 for an example of the recoding procedure for a sample dilemma.

Analysis plan for hypothesis testing. All analyses were conducted in R using the lfe package. For the purpose of testing our hypotheses, following ref. 45, we used a significance threshold of α with the corresponding interpretation as in Supplementary Table 1 (see justification in 'Power analysis'). This excluded inferiority tests and equivalence tests, which used 0.05 as a threshold.

Analysis plan for H1. We estimated the mean proportion of yes responses for the two non-medical traits and compared it with chance (0.5). Data preparation: We recoded 'yes' responses as 1 and 'no' or 'not sure' responses 0, subtracted 0.5 from all values, and set 'traits' as the baseline. Data were pooled across the attributes. Model: An LPM was used to estimate the intercept, which directly represents the mean deviation of yes responses for traits from 0.5. We tested whether this intercept is negative and significantly different from zero to evaluate our hypothesis. Clustering: Participant-level clustered standard errors were used to account for correlations within participants.

Analysis plan for H2. We estimated the average treatment effect (ATE) of the treatment (medical conditions versus non-medical traits) on the binary outcome variable (1 for 'yes'; 0 for 'no' or 'not sure'). Data preparation: We pooled data across the two attributes in each group (total of four), with medical conditions as the control group and non-medical traits as the treatment group. Model: An LPM was used to compute the ATE as the difference in mean responses between the two groups (a non-parametric estimator). We tested whether this coefficient is negative and significantly different from zero to evaluate our hypothesis. Clustering: Participant-level clustered standard errors were used to account for correlations within participants.

Analysis Plan for H3. We estimated the AMCE of the above-average chance level for non-medical traits using conjoint analysis[49], pooling data across both traits. Data preparation: We filtered the dataset to include only rows where the attribute type is 'trait' and where the below-average chance level is excluded (that is, only average and above-average chance levels remain). The average chance level was set as the baseline category. Model: Using an LPM, we included the dummy variable Above_Average to estimate the difference between above-average and average chance levels. The AMCE was tested by examining whether the coefficient for Above_Average is negative and significantly different from zero. Clustering: Participant-level clustered standard errors were used to account for correlations within participants.

Analysis plan for H4. We estimated the average component interaction effect (ACIE) of the interaction between above-average chance levels and attribute type (non-medical traits versus medical conditions). Data preparation: We filtered the dataset to exclude rows where the below-average chance level is present, ensuring only average and above-average chance levels remain. Model: Using an LPM, the interaction term 'above average' × 'attribute type' was included to estimate the differential effect of above-average chance levels depending on whether the attribute is a non-medical trait or medical condition. The coefficient of the interaction term was interpreted as the ACIE, representing the difference in effects between traits and conditions. Clustering: Participant-level clustered standard errors were calculated to account for correlations within participants. Inferiority test: As H4 stated that 'US-based participants will exhibit an avoidance of non-medical traits *at least as strongly as* medical conditions', we planned to conduct an inferiority test[47]. The inferiority threshold was set to 0.2 (considered a conventionally small standardised effect size), indicating that non-medical traits must show at least this level of avoidance compared with medical conditions. We tested this using a one-sided t test. The null hypothesis is that the interaction effect (Above_Average × Attribute_Type) is greater than or equal to the threshold (that is, effect ≥ 0.2). The alternative hypothesis is that the interaction effect is below the threshold (that is, effect < 0.2). Above_Average × Attribute_Type is the interaction between chance level (above-average versus average) and attribute type (trait versus condition).

Analysis plan for H5. This followed the same plan as H3 and H4 for data collected in study 1b with Chinese participants.

Analysis plan for H6. This was planned to be tested in two steps across two studies. However, the first step showed that there is no need for the

second step: Study 1a (exploratory measurement-of-mediation design): A measurement-of-mediation analysis was performed using established mediation analysis methods[35,36]. Had study 1a provided statistical evidence for mediation, study 2 would have been conducted. Study 2 (confirmatory measurement-of-mediation design): Data from the control group (mediator varies freely) would be used to confirm the mediation effect observed in study 1a. Study 2 (manipulation-of-mediator design): The blockage and enhancement manipulations would be compared against the control group to test whether controllability is a mediator. An interaction term between chance levels (above average versus average) and controllability (low controllability versus freely varying versus high controllability) would be tested to assess whether controllability is a partial mediator. Moreover, an AMCE of chance levels would be calculated for the blockage condition to determine whether controllability is a full mediator. If the larger P value was smaller than the chosen threshold (0.05), we would have interpreted this as evidence for null effect (statistical equivalence).

Robustness checks. We repeated the above analysis using the following datasets:

(1) All paid participants (those who completed the survey and passed at least one of the two attention checks). This includes more data than in the main analysis (those who passed one but not both of attention checks). This dataset was nationally representative in terms of age, gender and ethnicity (USA), or age and gender (China).
(2) Participants who passed all attention checks and all comprehension checks. This would be a subset of the dataset used in the main analysis.

Null hypothesis testing. We planned that in the case of no significance for all hypotheses except H4, we would conduct a TOST procedure[47], using $d = 0.2$ as a threshold. However, this was not needed. In the case of H4, we planned to conduct an inferiority test: a one-sided test.

Protocol registration

The Stage 1 protocol for this Registered Report was accepted in principle on 7 October 2025. The protocol, as accepted by the journal, can be found at https://figshare.com/s/dbd00321961f088c1276?file=58830508 and https://osf.io/vb9c2.

Deviations from registered protocol. The deviations from the stage 1 plan fall into two categories: (1) typographical or terminological errors in the preregistration, and (2) adaptations required for data collection in the Chinese context.

(1) Typographical errors
H6 mediation (attribute versus attribute type): H6 was registered as a mediation by perceived controllability at the attribute level. In Table 1, the interpretation column for Question 6 instead referred to attribute type (that is, traits versus conditions) as the independent variable. This was a typographical error that appeared nowhere else in the preregistration and was inconsistent with the stated hypothesis and measurement plan. For transparency, we also tested mediation using attribute type as the independent variable. This analysis suggested partial mediation (Supplementary Table 25); however, inspection of attribute-level controllability ratings shows that the effect is driven mainly by a single attribute (low IQ; Fig. 3c) rather than a general difference between traits and conditions. As controllability was measured and intended to vary at the attribute level, we treat the attribute-level analysis as the appropriate test of H6.
Study 1a—race/ethnicity quotas: Study 1a was registered to use a sample representative for age, gender and race. Prolific implements this using the category 'ethnicity (simplified)', which corresponds to the intended race-based quota (Asian, Black, white, mixed, other).
Non-inferiority terminology: The stage 1 preregistration referred to analyses in H4 and H5b as 'inferiority tests', but the described procedure was a standard non-inferiority test against a prespecified inferiority margin ($d = 0.20$), consistent with the hypothesis that avoidance of non-medical traits would be at least as strong as avoidance of medical conditions. In stage 2, we use non-inferiority to refer to this same procedure.
(2) Adaptations for the Chinese sample
Study 1b—demographic questions (China): In study 1b, demographic answer options (income, education and age) were adapted to the Chinese context and to Credamo's standard groupings. The preregistration did not specify exact categories or currency.
Study 1b—attention check platforms (China): The preregistration listed an attention check using mostly US-based platforms (for example, YouTube, X/Twitter, Facebook, Reddit and TikTok). For the Chinese sample, these were replaced with locally relevant platforms (Youku, Weibo, WeChat, Zhihu and Douyin). The structure and purpose of the attention check remained unchanged.
Study 1b—unmatched participants (China): Credamo matches participants for payment using the last six digits of phone numbers. Participants were asked to provide these digits at the start of the study to receive payments on Credamo. For study 1b, 31 participants could not be matched. Given that using data from non-compensated participants would be unethical, their data were excluded from the main analyses and we requested an additional 31 participants. Moreover, during the collection of the final 50 participants, we ended up collecting 58 participants rather than 50. The last 8 participants were also excluded to stick to the registered number. Additional robustness analysis shows that these additional changes do not make any difference to the main analyses.

Reporting summary

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

Data availability

All data and materials associated with this research, including for the pilot studies, is available via Open Science Framework (OSF) at: https://osf.io/vb9c2.

Code availability

All analysis code (completed in R) associated with this research, including for the pilot studies, is available via OSF at: https://osf.io/vb9c2.

References

1. Meyer, M. N., Tan, T., Benjamin, D. J., Laibson, D. & Turley, P. Public views on polygenic screening of embryos. *Science* **379**, 541–543 (2023).
2. Haining, C. M., Savulescu, J., Keogh, L. & Schaefer, G. O. Polygenic risk scores and embryonic screening: considerations for regulation. *J. Med. Ethics* **51**, 719–728 (2025).
3. Taylor-Sands, M. & Bowman-Smart, H. Non-invasive prenatal testing for adult onset conditions: reproductive choice and the welfare of the future child. *Melb. Univ. Law Rev.* **45**, 730–778 (2022).
4. Polyakov, A. et al. Polygenic risk score for embryo selection—not ready for prime time. *Hum. Reprod.* **37**, 2229–2236 (2022).
5. Savulescu, J., Earp, B. D. & Schuklenk, U. Ethics of genetic research on same-sex sexual behaviour. *Nat. Hum. Behav.* **5**, 1123–1124 (2021).

6. Roura-Monllor, J. A. et al. Promises and pitfalls of preimplantation genetic testing for polygenic disorders: a narrative review. *F S Rev.* **6**, 100085 (2025).
7. Turley, P. et al. Problems with using polygenic scores to select embryos. *N. Engl. J. Med.* **385**, 78–86 (2021).
8. Brown, K. V. Aspiring parents have a new DNA test to obsess over. *The Atlantic* https://www.theatlantic.com/health/archive/2025/01/polygenic-risk-score-ivf/681323/ (2025).
9. Nietfeld, E. It's illegal in most of the world. In America, new parents are embracing it—for better or worse. *Slate* https://slate.com/technology/2024/05/ivf-daughters-toxic-masculinity-sex-selection.html (2024).
10. The Fertility Institutes. Choose your baby's eye color. *The Fertility Institutes* https://www.fertility-docs.com/programs-and-services/pgd-screening/choose-your-babys-eye-color (2021).
11. Kalfoglou, A. L., Scott, J. A. & Hudson, K. PGD patients' and providers' attitudes about the use and regulation of PGD. *Reprod. Biomed. Online* **11**, 486–496 (2005).
12. Zuckerman, S., Zeevi, D. A., Gooldin, S. & Altarescu, G. Acceptable applications of preimplantation genetic diagnosis (PGD) among Israeli PGD users. *Eur. J. Hum. Genet.* **25**, 1113–1117 (2017).
13. Hughes, T. et al. A review on the motivations, decision-making factors, attitudes and experiences of couples using pre-implantation genetic testing for inherited conditions. *Hum. Reprod. Update* **27**, 944–966 (2021).
14. Furrer, R. A. et al. Public attitudes, interests, and concerns regarding polygenic embryo screening. *JAMA Netw. Open* **7**, e2410832 (2024).
15. de Raad, B. *The Big Five Personality Factors: The Psycholexical Approach to Personality* (Hogrefe Publishing, 2000).
16. Hertwig, R. & Engel, C. *Deliberate Ignorance: Choosing Not to Know* (MIT Press, 2021).
17. Haining, C. M., Toh, H. J., Savulescu, J. & Schaefer, G. O. Singaporean attitudes to cognitive enhancement: a cross-sectional survey. *J. Med. Ethics* **51**, 747–753 (2025).
18. Smolarczyk, R. et al. Foretelling the future: preimplantation genetic testing and the coming of polygenic embryo screening. *J. Clin. Med.* **14**, 3885 (2025).
19. Black, J. Embryo scoring: 10 breakthrough technologies 2026. *Technology Review* https://www.technologyreview.com/2026/01/12/1130011/embryo-scoring-genetic-testing-2026-breakthrough-technology/ (2026).
20. Forzano, F. et al. The use of polygenic risk scores in pre-implantation genetic testing: an unproven, unethical practice. *Eur. J. Hum. Genet.* **30**, 493–495 (2022).
21. European Society of Human Reproduction and Embryology (ESHRE). ESHRE supports the position of ESHG on embryo selection based on polygenic risk scores. *ESHRE* https://www.eshre.eu/Europe/Position-statements/PRS (2022).
22. Furrer, R. A. et al. Survey of U.S. reproductive medicine clinicians' attitudes on polygenic embryo screening. *npj Genom. Med.* **10**, 79 (2025).
23. Shah, S., Cantor, R. M., Quinn, M., Shamshoni, J. & Palmer, C. G. S. Preimplantation genetic testing for polygenic disorders: viewpoints of reproductive genetic counselors and reproductive endocrinology and infertility specialists in the United States. *J. Genet. Couns.* **34**, e70072 (2025).
24. Bayefsky, M. J. Comparative preimplantation genetic diagnosis policy in Europe and the USA and its implications for reproductive tourism. *Reprod. Biomed. Soc. Online* **3**, 41–47 (2016).
25. Ginoza, M. E. C. & Isasi, R. Regulating preimplantation genetic testing across the world: a comparison of international policy and ethical perspectives. *Cold Spring Harb. Perspect. Med.* **10**, a036681 (2020).
26. Golman, R., Hagmann, D. & Loewenstein, G. Information avoidance. *J. Econ. Lit.* **55**, 96–135 (2017).
27. Genomic Prediction. Embryo health score test. *LifeView* https://www.lifeview.com/ehs (2026).
28. Munday, S. & Savulescu, J. Three models for the regulation of polygenic scores in reproduction. *J. Med. Ethics* **47**, e91 (2021).
29. Muschetto, T. & Siegel, J. T. Attribution theory and support for individuals with depression: the impact of controllability, stability, and interpersonal relationship. *Stigma Health* **4**, 126–135 (2019).
30. Muschetto, T. & Siegel, J. T. Perceived stability of depressive symptomology and willingness to help relational partners: an attributional perspective. *Curr. Psychol.* **42**, 14076–14091 (2023).
31. Human Fertilisation and Embryology Authority. Approved PGT-M and PTT conditions. *HFEA* https://www.hfea.gov.uk/treatments/embryo-testing-and-treatments-for-disease/approved-pgt-m-and-ptt-conditions (2016).
32. Awad, E. Digital pathways beyond Western-centric participants. *Behav. Res Methods* **57**, 229 (2025).
33. Savulescu, J., Kahane, G. & Gyngell, C. From public preferences to ethical policy. *Nat. Hum. Behav.* **3**, 1241–1243 (2019).
34. Mark, M. M. From program theory to tests of program theory. *N. Dir. Prog. Eval.* **1990**, 37–51 (1990).
35. Pirlott, A. G. & MacKinnon, D. P. Design approaches to experimental mediation. *J. Exp. Soc. Psychol.* **66**, 29–38 (2016).
36. Cooper, J., Zanna, M. P. & Taves, P. A. Arousal as a necessary condition for attitude change following induced compliance. *J. Pers. Soc. Psychol.* **36**, 1101–1106 (1978).
37. Lázaro-Muñoz, G., Pereira, S., Carmi, S. & Lencz, T. Screening embryos for polygenic conditions and traits: ethical considerations for an emerging technology. *Genet. Med.* **23**, 432–434 (2021).
38. Brislin, R. W. Back-translation for cross-cultural research. *J. Cross Cult. Psychol.* **1**, 185–216 (1970).
39. Kowal, M. Translation practices in cross-cultural social research and guidelines for the most popular approach: back-translation. *Anthropol. Rev.* **87**, 19–32 (2024).
40. Using representative samples on Prolific—FAQs. *Prolific* https://researcher-help.prolific.com/en/articles/445162-using-representative-samples-on-prolific-faqs (2026).
41. Hofstede, G. *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations* (SAGE, 2001).
42. Li, C., Düring, D. & Kwok, S. H. *Harmony in Chinese Thought: A Philosophical Introduction* (Rowman & Littlefield, 2021).
43. Strong, R. W. & Alvarez, G. Using simulation and resampling to improve the statistical power and reproducibility of psychological research. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/2bt6q (2019).
44. Green, P. & MacLeod, C. J. SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods Ecol. Evol.* **7**, 493–498 (2016).
45. Benjamin, D. J. et al. Redefine statistical significance. *Nat. Hum. Behav.* **2**, 6–10 (2018).
46. Lakens, D. Why you don't need to adjust your alpha level for all tests you'll do in your lifetime. *The 20% Statistician* http://daniellakens.blogspot.com/2016/02/why-you-dont-need-to-adjust-you-alpha.html (2016).
47. Lakens, D., Scheel, A. M. & Isager, P. M. Equivalence testing for psychological research: a tutorial. *Adv. Methods Pract. Psychol. Sci.* **1**, 259–269 (2018).
48. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* (Routledge, 2013).
49. Hainmueller, J., Hopkins, D. J. & Yamamoto, T. Causal inference in conjoint analysis: understanding multidimensional choices via stated preference experiments. *Polit. Anal.* **22**, 1–30 (2014).

Acknowledgements
The authors thank Y. Ye and S. Li of the Women's Hospital, Zhejiang University School of Medicine and Y. Chu of the Center for Psychological Sciences, Zhejiang University, for the forward- and back-translation of the study 1b materials into Chinese, and X. Xiao of the Center for Psychological Sciences, Zhejiang University, for coordinating the translation of the study 1b survey. We note here that several minor changes were made to the Introduction at stage 2, following reviewer feedback and to accurately reflect the study as conducted. The sentence 'Should Study 1a suggest a potential mediating effect of controllability on implantation decisions, this effect will be directly tested in a new Study 2, where participants will be randomly allocated into one of three controllability groups: one that is identical to the implantation task in study 1a, one that aims to enhance the effect of the mediator, and one that aims to block (or mitigate) the effect of the mediator' was removed. The sentence 'However, based on the "measurement-of-mediation" results, this study was not run' was added in a separate location to accurately reflect that study 2 was not conducted. The parenthetical example, '(for example, genetic markers for a non-medical attribute might be detected during screening for other unrelated attributes)', was also removed because this example is not well supported in the case of highly polygenic behavioural traits. Finally, three qualifiers were added to statements for greater precision: 'in some clinics' in 'PGT is now also offered [in some clinics] for assessing polygenic risk scores for common diseases', 'possible' in 'This study investigated this [possible] discrepancy', and 'expressed' in 'it goes beyond simply assessing public's [expressed] willingness to perform testing'.

Author contributions
E.A. and J.S. conceived the research. J.S., D.W., I.S., and P.L. secured funding for the research. E.A., J.S., D.W., J.D.-C., B.D.E., G.O.S. and I.S. designed the pilot studies. E.A. collected and analysed the pilot data in consultation with J.S. E.A. conducted the power analysis in consultation with C.C. All authors contributed to the experimental design of the planned studies. E.A. and P.L. collected data. P.L. and E.A. oversaw the prompt translation for the Chinese sample (study 1b). E.A. developed the analysis plan, conducted the analyses, produced the visualizations and drafted the paper with feedback from all authors. All authors reviewed, revised and approved both stages of the paper.

Funding
This research was funded in part by the Wellcome Trust (grant number: 226801) for Discovery Research Platform for Transformative Inclusivity in Ethics and Humanities Research (ANTITHESES). For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. This research was also supported by the National University of Singapore under J.S.'s NUS Start-Up grant (NUHSRO/2022/078/Startup/13). It was also supported by the Social Science Research Council (Singapore) and administered by the Ministry of Education, Singapore, under its Social Sciences Research Thematic Grant (SSRC2023-SSRTG-006). This research was also supported by the National Institute for Health and Care Research (NIHR) Oxford Health Biomedical Research Centre (NIHR203316). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. Study 1b was supported by P.L.'s Start-up Fund for 'Hundred Talents Program' of Zhejiang University. B.D.E. was supported by NUSMed and ODPRT (NUHSRO/2024/035/Startup/04) for the project 'Experimental Philosophical Bioethics and Relational Moral Psychology'. J.A.C.E. was supported through a Philip Leverhulme Prize (PLP-2021-095). J.S. was further supported through the Murdoch Children's Research Institute by the Victorian State Government's Operational Infrastructure Support (OIS) Program. None of the funders had any role in study design, data collection and analysis, decision to publish or preparation of the article.

Competing interests
J.S. is a Bioethics Committee consultant for Bayer. The other authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41562-026-02562-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-026-02562-w.

Correspondence and requests for materials should be addressed to Edmond Awad, Dominic Wilkinson or Julian Savulescu.

Peer review information *Nature Human Behaviour* thanks David John Amor and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41562-026-02562-w

Extended Data Table 1 | Summary of focal results and conclusions for preregistered hypotheses across the main and robustness samples

Hypo.	Focal test	Sample	Key result	Conclusion
H1	Willingness to test non-medical traits vs. 50%	US - Main	61.4% "yes"; β = 0.114, 95% CI [0.083, 0.145], t = 7.34, p < 0.001	Not supported
		US - R1	61.5% "yes"; β = 0.115, 95% CI [0.086, 0.144], t = 7.49, p < 0.001	Not supported
		US - R2	62.2% "yes"; β = 0.122, 95% CI [0.089, 0.155], t = 7.18, p < 0.001	Not supported
H2	Willingness to test medical conditions vs. non-medical traits	US - Main	β = −0.053, 95% CI [-0.086, -0.020], t = −3.12, p = 0.0018	Supported
		US - R1	β = −0.053, 95% CI [-0.086, -0.020], t = −3.19, p = 0.0014	Supported
		US - R2	β = −0.036, 95% CI [-0.071, -0.001], t = −1.96, p = 0.0496	Suggested*
H3	Avoidance of undesirable non-medical traits (above-average vs. average chances)	US - Main	β = −0.341, 95% CI [-0.366, -0.316], t = −25.56, p < 0.001	Supported
		US - R1	β = −0.340, 95% CI [-0.365, -0.315], t = −25.86, p < 0.001	Supported
		US - R2	β = −0.358, 95% CI [-0.385, -0.331], t = −25.68, p < 0.001	Supported
H4	Avoidance of non-medical traits vs. medical conditions	US - Main	β = −0.050, 95% CI [-0.089, -0.011], t = −2.51, p = 0.012; non-inferiority criterion met (d = 0.20, one-sided p < 0.001)	Supported
		US - R1	β = −0.048, 95% CI [-0.087, -0.009], t = −2.45, p = 0.014; non-inferiority criterion met (d = 0.20, one-sided p < 0.001)	Supported
		US - R2	β = −0.052, 95% CI [-0.093, -0.011], t = −2.46, p = 0.014; non-inferiority criterion met (d = 0.20, one-sided p < 0.001)	Supported
H5a	Avoidance of undesirable non-medical traits (above-average vs. average chances)	China - Main	β = −0.411, 95% CI [-0.438, -0.384], t = −30.07, p < 0.001	Supported
		China - R1	β = −0.396, 95% CI [-0.420, -0.372], t = −32.02, p < 0.001	Supported
		China - R2	β = −0.415, 95% CI [-0.442, -0.388], t = −29.28, p < 0.001	Supported
H5b	Avoidance of non-medical traits vs. medical conditions	China - Main	β = −0.090, 95% CI [-0.131, -0.049], t = −4.35, p < 0.001; non-inferiority criterion met (d = 0.20, one-sided p < 0.001)	Supported
		China - R1	β = −0.102, 95% CI [-0.139, -0.065], t = −5.45, p < 0.001; non-inferiority criterion met (d = 0.20, one-sided p < 0.001)	Supported
		China - R2	β = −0.089, 95% CI [-0.130, -0.048], t = −4.15, p < 0.001; non-inferiority criterion met (d = 0.20, one-sided p < 0.001)	Supported
H6	Mediation by perceived controllability	US - Main	Across six pairwise contrasts, three ACMEs were non-significant and three were significant in the opposite direction to that predicted (see Supplementary Table 14)	Not supported
		US - R1	Across six pairwise contrasts, three ACMEs were non-significant and three were significant in the opposite direction to that predicted.	Not supported
		US - R2	Across six pairwise contrasts, three ACMEs were non-significant and three were significant in the opposite direction to that predicted.	Not supported

The table follows the format of Table 2 in the main text. "Main" denotes the primary analytic sample, "R1" includes all paid participants who passed at least one attention check, and "R2" includes participants in the main analysis who also passed all comprehension checks. Results were substantively unchanged across specifications, with one result shifting from supported to suggested (see starred conclusion).

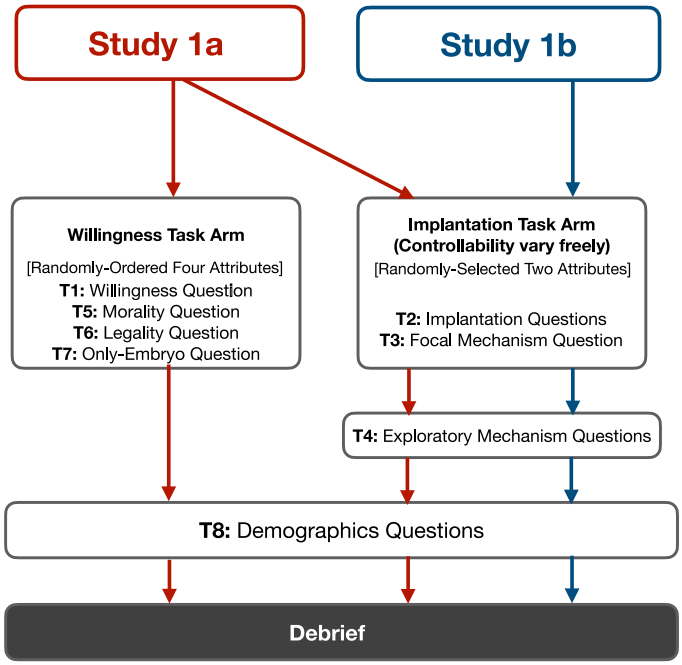
https://doi.org/10.1038/s41562-026-02562-w

Extended Data Table 2 | Cross-country differences in avoidance for above-average chances (China vs. US)

Attribute	n (obs.)	N (pps.)	Estimate	SE (cluster)	t-value	p-value
Low IQ	9000	678	-0.095	0.027	-3.529	<0.001
Heart Disease	8960	675	-0.051	0.027	-1.943	0.052
Antisocial Behaviour	8894	670	-0.043	0.027	-1.571	0.116
Vision Impairment	9117	687	0.006	0.026	0.221	0.825

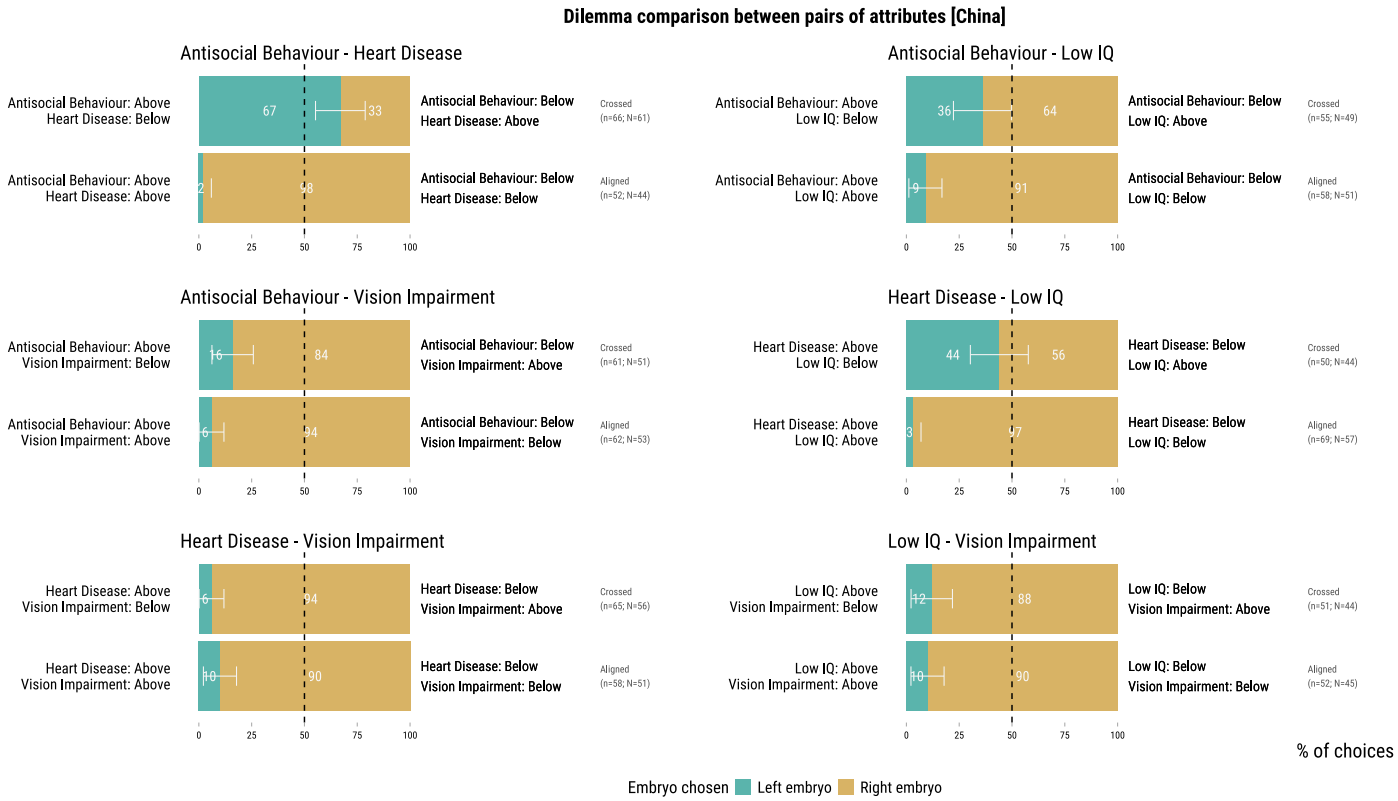
The table reports, for each attribute, the difference in the avoidance effect between China and the US, estimated using linear probability models with participant-level clustered standard errors. Negative coefficients indicate that above-average chances (vs. average chances) reduce implantation more strongly in China than in the US.

https://doi.org/10.1038/s41562-026-02562-w



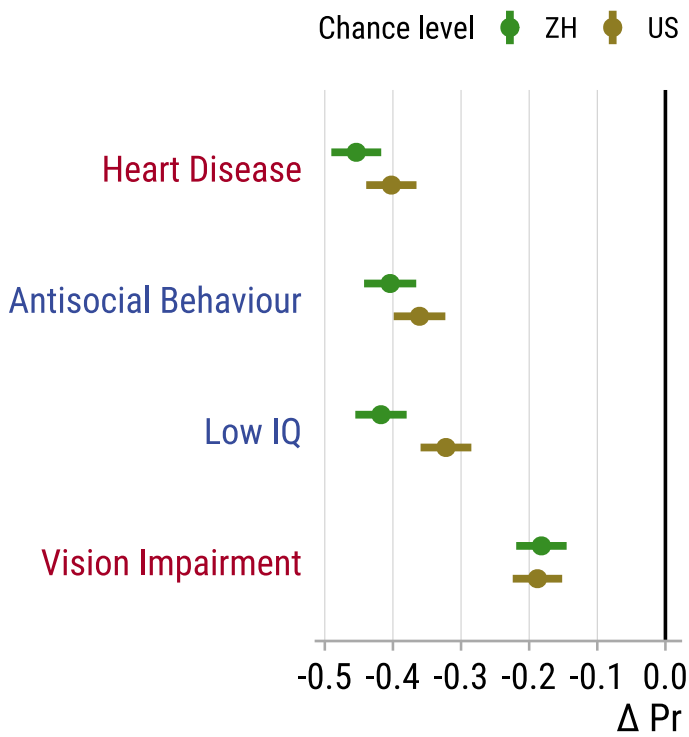
Extended Data Fig. 1 | Study design. Both Study 1a and Study 1b included an Implantation Task Arm, in which two attributes were randomly selected and Controllability varied freely. This arm covered the implantation and focal mechanism questions (T2–T3), followed by exploratory mechanism questions (T4). Study 1a also included a Willingness Task Arm, in which four attributes were shown in random order across the willingness, morality, legality, and only-embryo questions (T1, T5–T7). All arms then ended with demographics questions (T8) and a debrief.

https://doi.org/10.1038/s41562-026-02562-w



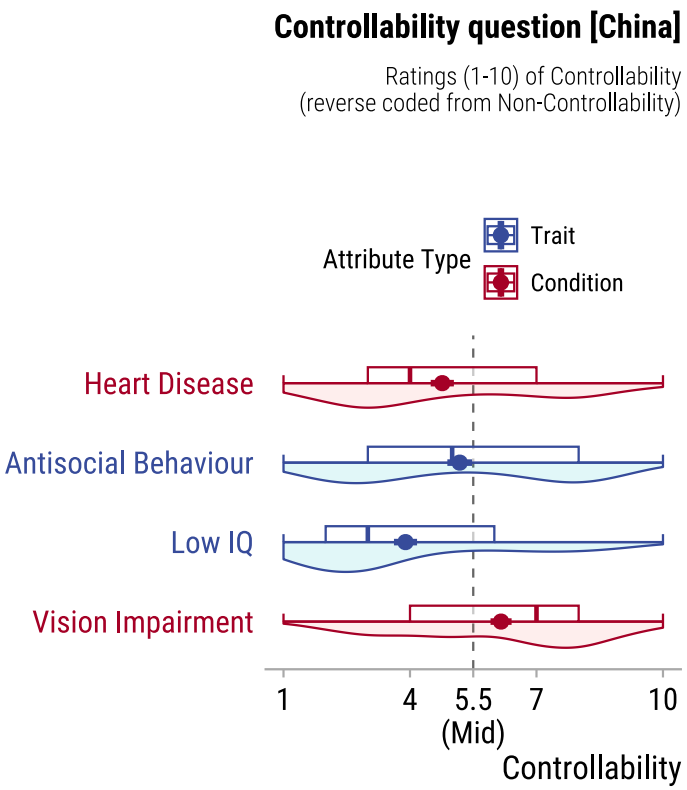
Extended Data Fig. 2 | Head-to-head implantation dilemmas between pairs of attributes (China sample). Each panel shows implantation choices in dilemmas where embryos differ on two attributes, shown separately for crossed dilemmas (one embryo above average on one attribute and below average on the other) and aligned dilemmas (one embryo above average on both attributes and the other below average on both). Bars show the proportion of choices favouring the left embryo (teal) and right embryo (gold). The dashed vertical line marks 50% (equal preference). Error bars indicate 95% confidence intervals (±1.96 SE, clustered by participant). n (responses) and N (participants) per dilemma type and attribute pair are indicated on the figure. Participants were randomly assigned two of four attributes, so N varies by pair. No control group was used; the reference point is equal choice (50%). This figure mirrors Fig. 5 for the Chinese sample.

https://doi.org/10.1038/s41562-026-02562-w



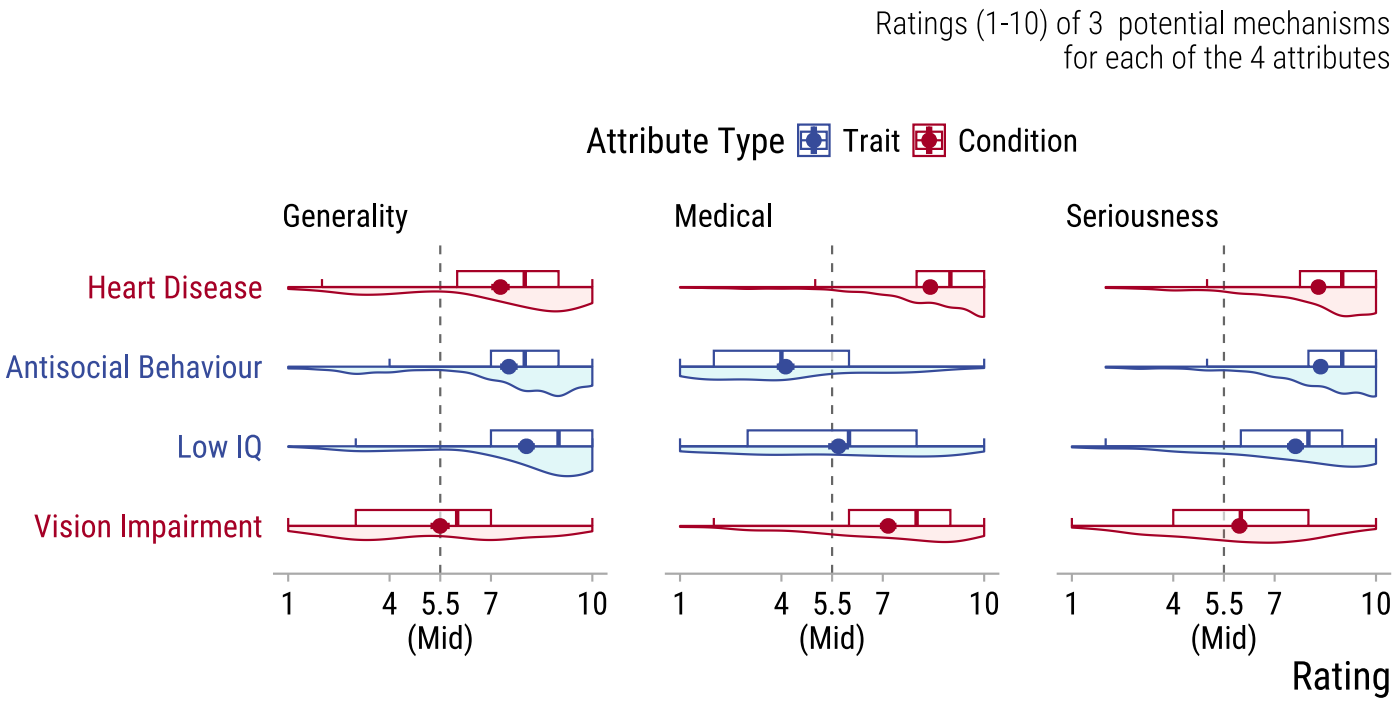
Extended Data Fig. 3 | Implantation decisions in the US and China. Estimated effect of significantly above-average chances (relative to average chances) on the probability of choosing to implant an embryo, shown separately for each attribute in the US (olive) and China (green) samples. Points are coefficient estimates from linear probability models with cluster-robust standard errors; error bars indicate 95% confidence intervals (±1.96 SE, clustered by participant). Non-medical traits are shown in blue; medical conditions in red. N per attribute (US): Heart Disease = 355, Antisocial Behaviour = 357, Low IQ = 378, Vision Impairment = 374 participants; N per attribute (China): Heart Disease = 320, Antisocial Behaviour = 313, Low IQ = 300, Vision Impairment = 313 participants; n per attribute = 10 × N in each sample (each participant completed 10 dilemmas per assigned attribute; as in Fig. 3a, b). Each participant was randomly assigned two of the four attributes. The reference category in all models is average chances. No control group was used; attributes were compared within participants.

https://doi.org/10.1038/s41562-026-02562-w



Extended Data Fig. 4 | Ratings of perceived controllability (China sample). Perceived controllability ratings for each attribute (China sample), measured on a 1–10 scale (reverse-coded from non-controllability; higher values indicate greater perceived controllability). Violin plots show the full response distribution; box plots show median (centre line), interquartile range (box bounds, 25th–75th percentile), and whiskers extending to the scale limits (1 and 10); points show means with error bars indicating 95% confidence intervals (±1.96 SE, clustered by participant). Each participant rated each assigned attribute once, so n = N per attribute: Heart Disease = 320, Antisocial Behaviour = 313, Low IQ = 300, Vision Impairment = 313 participants (Implantation arm, China sample; N per attribute as in Fig. 3b). Medical conditions are shown in red; non-medical traits in blue. No control group was used; attributes were compared within the same participants. This figure corresponds to Fig. 3c for the Chinese sample.

https://doi.org/10.1038/s41562-026-02562-w



Extended Data Fig. 5 | Exploratory ratings of potential psychological mechanisms (China sample). Distributions of participant ratings (1–10 scale) for three candidate mechanisms — generality (extent to which the attribute affects many areas of life), medical relevance (extent to which the attribute is considered a medical condition), and seriousness (severity of potential consequences) — assessed for each of the four attributes. Higher values indicate greater perceived generality, medical relevance, or seriousness. Violin plots show the full response distribution; box plots show median (centre line), interquartile range (box bounds, 25th–75th percentile), and whiskers extending to the scale limits (1 and 10); points show means with error bars indicating 95% confidence intervals (±1.96 SE, clustered by participant). Dashed vertical lines indicate the scale midpoint (5.5). Medical conditions are shown in red; non-medical traits in blue. Each participant rated each assigned attribute once, so n = N per attribute: Heart Disease = 320, Antisocial Behaviour = 313, Low IQ = 300, Vision Impairment = 313 participants (Implantation arm, China sample). Each participant was randomly assigned two of the four attributes and rated both. These measures were collected for exploratory purposes and were not preregistered as mediators. This figure corresponds to Fig. 6 for the Chinese sample.

nature portfolio

Corresponding author(s): Edmond Awad

Last updated by author(s): May 18, 2026

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	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	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)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	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 were collected using Qualtrics survey software.
Data analysis	Data cleaning, analysis and visualisation were conducted in R version 4.2.2 using RStudio version 2023.9.1.494, with the following packages: data.table (1.18.4), lfe (3.0.0), mediation (4.5.1), car (3.1.2), tidyverse (2.0.0), hrbrthemes (0.9.3), gghalves (0.1.4), knitr (1.51), kableExtra (1.4.0), gtsummary (1.7.2), gt (1.3.0), sciplot (1.2.0), sysfonts (0.8.9), and showtext (0.9.7).

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

All data and materials associated with this research, including for the pilot studies, are available on Open Science Framework (OSF) website at this link: https://osf.io/vb9c2/?view_only=6507069145df420d99f1363f6b53afb0.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Participants reported gender as part of the demographic questionnaire, using the response options: Female; Male; Non-binary / third gender; Prefer not to say; and Prefer to self-describe. Gender breakdowns reported in the "Behavioural & social sciences study design" section below are based on this survey question. Recruitment quotas for US-based participants (Study 1a) were set through Prolific to approximate the US adult population by age, gender and ethnicity (simplified), using the binary gender categories available for quota sampling on the platform. Recruitment quotas for Chinese participants (Study 1b) were set through Credamo to approximate the China adult population by age and gender. We did not conduct analyses stratified by gender, and the preregistered confirmatory analyses were not designed to test gender-specific effects. Breakdown by gender is provided for the two registered studies in the "Behavioural & social sciences study design" section below.
Reporting on race, ethnicity, or other socially relevant groupings	We did not include a race or ethnicity question in the study questionnaire. For US-based participants (Study 1a), recruitment quotas were set through Prolific using Prolific's ethnicity (simplified) variable, with five categories: Asian, Black, White, Other and Mixed. This variable was used to approximate the ethnic composition of the US adult population during recruitment and to describe the sample. It was not used as a proxy for socioeconomic status and was not included as a covariate or main explanatory variable in the preregistered confirmatory analyses.
Population characteristics	See "Behavioural & social sciences study design" section below.
Recruitment	Participants were recruited online. US-based participants (Study 1a) were recruited through Prolific and were required to be adults based in the United States. Recruitment for Study 1a used Prolific quotas for age, sex and ethnicity (simplified) to approximate the US adult population. Chinese participants (Study 1b) were recruited through Credamo and were required to be adults based in China. Recruitment for Study 1b used quotas for age and gender to approximate the China adult population. As with other online-panel studies, participants were members of research panels who chose to take part in a paid online study, which may limit generalisability to people who do not participate in online research. The use of quota sampling was intended to reduce, but cannot remove, this source of selection bias.
Ethics oversight	The studies were reviewed and approved by the Social Sciences and Humanities Interdivisional Research Ethics Committee (SSH IDREC) at the University of Oxford (Protocol ID: R80692), and by the Research Ethics Board of the Center for Psychological Sciences at Zhejiang University, China (Protocol ID: 2025-002). The research complied with all relevant ethical regulations and informed consent was obtained from all human participants. Participants were compensated financially for their participation in accordance with national minimum wage and/or the standard rates of the online platforms.

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

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

Behavioural & social sciences study design

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

Study description	Experimental design, quantitative data. Participants completed online survey tasks.
Research sample	The research sample consisted of adult online-panel participants in the United States and China. Study 1a recruited US-based adults through Prolific, using quotas for age, gender and ethnicity (simplified) to approximate the US adult population. After preregistered exclusions, the final analysed Study 1a sample was 1,467 participants. Participants were 11% aged 18–24, 18% aged 25–34, 18% aged 35–44, 16% aged 45–54, 23% aged 55–64 and 13% aged 65 or older. The sample was 51% female, 48% male, 1.3% non-binary / third gender, 0.3% preferred not to say and 0.1% preferred to self-describe. Most participants had no children under 18 (63%), were working full-time or part-time (70%), and had at least some university education (79%). Household income was distributed across the response categories, with the largest groups reporting $25,000–$49,999 (22%), $100,000–$149,999 (19%) and $50,000–$74,999 (18%). Study 1b recruited China-based adults through Credamo, using quotas for age and gender to approximate the China adult population. After preregistered exclusions, the final analysed Study 1b sample was 623 participants. Participants were 12% aged 18–25, 11% aged 26–30, 15% aged 31–35, 11% aged 36–40, 7.9% aged 41–45, 9.0% aged 46–50, 8.5% aged 51–55 and 26% aged 56 or older. The sample was 50% female, 48% male, 0.8% preferred not to say and 0.6% preferred to self-describe. Most participants had one child under 18 (45%) or no children under 18 (42%), were working full-time (80%), and had a bachelor's degree or graduate degree (75%). Household income was most commonly CNY100,001–CNY250,000 (40%) or CNY250,001–CNY500,000 (21%).

Sampling strategy	We used voluntary response sampling from online research panels, with quota sampling in the two main registered studies. Study 1a recruited US-based adults through Prolific, using quotas for age, gender and ethnicity (simplified) to approximate the US adult population. Study 1b recruited China-based adults through Credamo, using quotas for age and gender to approximate the China adult population. Sample size was determined at Stage 1 through a cost-benefit analysis that took into account available resources, expected effect sizes and simulation-based power analyses using pilot data. The minimum planned final sample size was 1,200 participants for Study 1a, with 600 participants per randomly assigned arm, and 600 participants for Study 1b. To allow for exclusions, we initially recruited 1,500 participants for Study 1a and 750 participants for Study 1b, with a preregistered plan to recruit additional participants in batches of 50 if fewer than 600 valid participants were retained in any arm. The initial Study 1a recruitment was sufficient; for Study 1b, one additional batch of 50 participants was collected, following the preregistered plan. The final analysed samples were 1,467 participants in Study 1a and 623 participants in Study 1b.
Data collection	Data collection was conducted online via Qualtrics. Participants completed the survey remotely and unsupervised. Blinding is not applicable, as data were collected automatically by the platform without any researcher involvement during the session.
Timing	Data were collected for Study 1a between 27 October 2025 and 3 November 2025, and for Study 1b between 26 November 2025 and 12 December 2025.
Data exclusions	As specified in the Stage 1 protocol, participants were eligible if they were adults, consented to take part, met the location and language criteria for the relevant study, and completed the survey. For Study 1a, participants had to be based in the United States and fluent in English; for Study 1b, participants had to be based in China and fluent in Mandarin. Before data collection for Study 1a, Prolific participants who had taken part in the pilot studies for this project or in selected related pilot studies by our team were prevented from taking the survey. During data collection, participants had to provide their platform ID, complete the study, and pass at least one of two attention checks to be compensated. For the main preregistered analyses, we then excluded compensated participants who failed either attention check (as specified in the Stage 1 protocol). In Study 1a, 1,501 participants completed the survey and were compensated; 1,467 passed both attention checks and were included in the main analyses. In Study 1b, 800 participants completed the survey and were compensated; 623 passed both attention checks and were included in the main analyses. In Study 1b, 31 additional participants could not be matched by the platform and were not compensated, so their data were not included. A further 8 participants collected beyond the registered target were compensated but excluded from analysis to keep to the registered sample size.
Non-participation	In Study 1a, 1,580 individuals attempted the survey. Of these, 1,501 completed the survey and were compensated. The remaining 79 individuals did not enter the final participant pool because they left at the welcome page before seeing the consent page (N = 4), did not provide a valid or matching Prolific ID (N = 1), did not pass the consent page because they either did not consent to take part or reported being under 18 (N = 17), or dropped out after consent but before completion (N = 57). No participants failed both attention checks before completing the study. In Study 1b, 1,043 individuals attempted the survey. Of these, 808 completed the survey and were compensated. The remaining 235 individuals did not enter the final compensated participant pool because they did not pass the consent page because they either did not consent to take part or reported being under 18 (N = 24), dropped out after consent but before completion (N = 152), failed both attention checks before completing the study (N = 28), or could not be matched by the recruitment platform (N = 31). Of the 808 participants who were compensated, 8 were collected beyond the registered target and excluded from analysis to keep to the registered sample size.
Randomization	In Study 1a, participants were randomly assigned at the beginning of the experiment to one of two arms: the Willingness Task Arm or the Implantation Task Arm. In the Willingness Task Arm, participants saw all four attributes in a random order, and the same attribute order was used across the willingness, morality, legality and only-embryo questions. In the Implantation Task Arm, participants were randomly assigned two of the four attributes, in a random order that was then fixed across tasks for that participant. For these two attributes, participants completed 10 implantation dilemmas randomly sampled from 36 possible embryo-pair dilemmas. In Study 1b, all participants completed the Implantation Task only, following the same randomisation procedure for attribute assignment, attribute order and dilemma sampling. Attribute pairings and chance configurations were counterbalanced across participants.

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	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A