Behavioural Markers of Consciousness Shape Moral Status Judgments

Matan Mazor[1], Arianna Risoli[2], Anna Eberhardt[2], & Stephen M. Fleming[2,3,4]

[1] All Souls College and Department of Experimental Psychology, University of Oxford

[2] Department of Experimental Psychology, University College London

[3] Functional Imaging Laboratory, University College London

[4] Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London

Author note

Correspondence concerning this article should be addressed to Matan Mazor, All Souls College, High Street, Oxford OX1 4AL. E-mail: matan.mazor@psy.ox.ac.uk

Abstract

Whether an individual is conscious or not has important ethical and legal implications. For example, animal welfare regulations seek to minimize suffering in conscious animals, and AI ethics discussions focus on the possibility of artificial consciousness as a key concern. Critically, consciousness cannot be directly perceived, but must be inferred from external cues such as behaviour and biological makeup. It remains unknown how people form such inferences. In three experiments, 652 English-speaking adults formed beliefs about consciousness based on behaviours that are considered markers of consciousness in the scientific literature. A between-subject staircasing procedure revealed that beliefs about consciousness contribute to moral status, and that such beliefs partly explain the moral significance of biological similarity to humans. Finally, markers of visual awareness contributed more to moral status than markers of self-awareness, despite similar effects on the perception of consciousness. We discuss implications for societal and legal discussions of non-human ethics.

Keywords: moral status, consciousness, ethics, meta-science

Behavioural Markers of Consciousness Shape Moral Status Judgments

Introduction

An entity can be said to have *moral status* when "it or its interests morally matter to some degree for the entity's own sake" (Jaworska & Tannenbaum, 2013). The attribution of moral status to a being affects how their interests are taken into account in everyday decision making. For example, the choice not to eat meat is often motivated by an attribution of non-trivial moral status to farm animals. But how do we decide which entities are worthy of a higher moral status, or, conversely, which can be exploited and harmed without concern for their interests?

Philosophers have long debated which factors should be taken into account when determining the moral status of individuals. Utilitarian philosophers Jeremy Bentham (1789) and John Stuart Mill (1861) attached moral status to the capacity to suffer, or to sentience more generally. Contemporary philosophical writings about the origins of moral status debate the extent to which they should be grounded in private, qualitative experience, or rather in functional or behavioural features which can be observed by others and scientifically quantified (Carruthers, 2019; Danaher, 2020; Dawkins, 2003; Levy, 2014). Crucially, regardless of what the normative answer to this question is, in actual moral decision-making we only have access to the observable properties of other individuals, such as their behaviour and physiology, and need to infer their internal states from these third-person observations. In other words, regardless of whether moral status *should* be based on public observations or private experiences, in practice it *must* be based on observable properties, because we can never directly perceive the subjective experience of others.

In psychology, behavioural observations allows inferences on internal mental processes, which are often assumed to have an experiential nature for the subject. In comparative psychology (the study of animal cognition), clever experimental manipulations allow scientists to deduce latent mental variables from observable behaviour. The mirror mark test is one such manipulation, in which an animal's response to an unfamiliar mark on its own reflection in the mirror is taken as a measure of self-awareness – a mental property that cannot be directly observed (Gallup, 1970). Other well-known examples are the study of caching behaviour as a measure of episodic memory (Clayton & Dickinson, 1998) and the use of trace-conditioning — the learning of associations between temporally separated events — as a measure of conscious perception (Clark, Manns, & Squire, 2002). In all of the above examples, scientists explain an observable behaviour as emerging from an internal mental state.

In turn, beliefs about internal mental states have been shown to correlate with moral concerns toward others. Specifically, previous research has demonstrated that participants attach moral significance to intelligence (Piazza & Loughnan, 2016; Wilks, Caviola, Kahane, & Bloom, 2021) and to the presence of subjective experience (H. M. Gray, Gray, & Wegner, 2007). Evidence from these and similar studies have led K. Gray, Young, and Waytz (2012) to argue that mind perception – the ascription of mental states to other entities – underpins judgments of moral concern.

Here we combine these two separate lines of work — comparative research on animal cognition and research in social psychology on the origins of human morality — to answer two questions about the links between perceived consciousness and morality. First, we ask whether members of the general public infer the presence of "consciousness" – as they interpret this word – from observed behaviours that are often taken by comparative psychologists as markers of

private mental states. And second, we ask whether these inferences in turn shape their moral decision making in relation to hypothetical questions of life and death. We separately assess the relevance of behavioural markers of a) perceptual (visual) consciousness, b) the capacity to experience rich valenced states such as suffering, c) self-awareness, and d) the unity of experience in time and across senses. In doing so, we not only ask whether perceived consciousness exerts influence over moral decision making, but also determine which dimensions of conscious experience (Birch, Schnell, & Clayton, 2020) contribute to moral status more than others.

Experiment 1

Methods

In order to isolate the unique effects of single behavioural and physiological features, independent of preconceptions about specific alien species or AI architectures, we followed Piazza and Loughnan (2016) and described imaginary aliens on a distant planet. According to the story, scientists have discovered several alien species on the planet, which all have two eyes and hand-like limbs, and all feed on space berries that grow on the planet. The scientists sorted the alien species into 8 pairs such that within each pair the two species were identical except for two differences. The experiment then consisted of descriptions of the 8 alien pairs, followed by questions. For each pair, participants were asked to describe in their own words the main difference between the two alien species. This allowed us to monitor the clarity of the descriptions and participants' attention. Then, participants were told that a fire started on the planet, and that two groups of aliens were caught in the fire, one of each species. We asked the participants which group they would rather save, assuming that the other group will die in the fire (Wilks et al., 2021). Lastly, participants were asked to use two sliders to indicate the extent to

which they thought each species was conscious. Crucially, consciousness was intentionally not defined to participants at any point, making their consciousness ratings reflective of their own interpretation of this term as they use it in their language, rather than based on a prescribed definition.

For the fire dilemma, the number of aliens in the feature-positive group (see below) was always 10. We determined the number of aliens in the feature-negative group based on the moral decisions of previous participants, following a Markov Chain Monte Carlo with People procedure (Sanborn, Griffiths, & Shiffrin, 2010). Specifically, 5 chains of 10 participants completed the experiment. Within each chain, the first participant decided between two groups of 10 aliens for all dimensions. In case the participant decided to save the feature-positive group, the number of feature-negative aliens was increased by 1 for the next participant in the chain. In case they decided to save the feature-negative group, the number of feature-negative aliens was decreased by 1 for the next participants in the chain. The same rule was then followed for all 10 participants in the chain.

Within each pair, the aliens varied along one of the following dimensions of interest: phenomenal richness, evaluative richness, unity, temporality, selfhood, size, physical resemblance to humans in appearance, and biological resemblance to humans (see Fig. 1, left panel). The first five dimensions are based on a taxonomy of animal consciousness, described in Birch et al. (2020), and the last three describe physical, rather than mental, dimensions. The *feature-positive* species had more of the mental capacity of interest, or resembled humans more, compared with the *feature-negative* species (the two species were given one-syllable gibberish names, fully counterbalanced across participants). Each dimension was presented as two scientific findings (accompanied by cartoon figures of the experimental design), followed by their

interpretation by the scientists. Below we describe the 8 dimensions and their experimental operationalizations, based on animal studies. The full descriptions as presented to our subjects are available in the Supplementary Materials.

1. *Phenomenal Richness:* Phenomenal richness is roughly defined as the 'level of detail with which [animals] consciously perceive aspects of their environment' (Birch et al., 2020). Phenomenal richness can vary between different sensory modalities, but for our study we focused on phenomenal richness in visual experience. While both alien species could see things and learn simple rules, only the feature-positive aliens could "tell the difference between very similar objects (for example, very similar shades of red)" (*fine-grained discrimination learning*, Pearce, Esber, George, & Haselgrove, 2008). In addition, only feature-positive aliens could learn to associate a light and a treat even when the treat is given as long as one second after the light (as opposed to only when the treat is given immediately) (*trace-conditioning*, Clark et al., 2002).

2. *Evaluative Richness:* Evaluative richness is to valence what phenomenal richness is to sense data. It is roughly defined as the ability to evaluate small changes in valence and to engage in complex affect-based decision-making (Birch et al., 2020). While both alien species had good and bad moods, liked sweet berries and disliked the cold, only feature-positive aliens had a tendency to look away from angry faces when in a bad mood (*affective bias*, Reimert, Fong, Rodenburg, & Bolhuis, 2017), and tended to sustain uncomfortably cold tempertaures for sweeter berries (*motivational trade-off*, Balasko & Cabanac, 1998).

3. *Unity:* Unity is roughly defined as having a "single, unified perspective as opposed to multiple perspectives" (Birch et al., 2020). Only feature-positive aliens could recognize a

fruit by touch alone after having seen it, without ever touching it before (*crossmodal integration*, Narins, Grabul, Soma, Gaucher, & Hödl, 2005). Also, only feature-positive aliens could generalise a trained association from their trained left eye to the untrained right eye (*interoccular transfer*, Ortega, Stoppa, Güntürkün, & Troje, 2008).

4. *Temporality:* Temporality is roughly defined as having an integrated stream of experience, as opposed to "a staccato series of fragmented experiences" (Birch et al., 2020). Temporality can be defined over short and long time scales — both suggested to constitute a dimension of animal consciousness (Birch et al., 2020). Here we focused on future planning over longer time scales (days). We described two future-planning paradigms: only feature-positive aliens chose to keep a tool that would help them to solve a problem at a later time (Kabadayi & Osvath, 2017), and preferred to wait for unripe fruit to ripen before eating them (Hillemann, Bugnyar, Kotrschal, & Wascher, 2014).

5. *Selfhood:* Selfhood is "the conscious awareness of oneself as distinct from the world outside" (Birch et al., 2020). While both alien species liked to keep themselves clean, only the feature-positive group would act to remove a fleck of dirt on their forehead if they see it in the mirror (*mirror self-recognition*, Gallup, 1970). Also, while both alien species sometimes stole food from their neighbours, only the feature-positive aliens attempted to hide themselves when caught (*experience projection*, De Waal, 1986).

6. *Size:* Feature-positive aliens were described as having average weight and height of 45kg and 120cm, while feature-negative aliens had average weight and height of 1 gram and 1 cm.

7. *Physical resemblance to humans:* Feature-positive aliens were described as having right and left eyes, and two hand-like limbs. Feature-negative aliens had upper and lower eyes, and 5 hand-like limbs.

8. *Biological resemblance to humans:* Feature-positive aliens were described as having red blood, and DNA that is composed of the same four bases as human DNA. Feature-negative aliens had yellow blood, and DNA that is composed of entirely different bases.

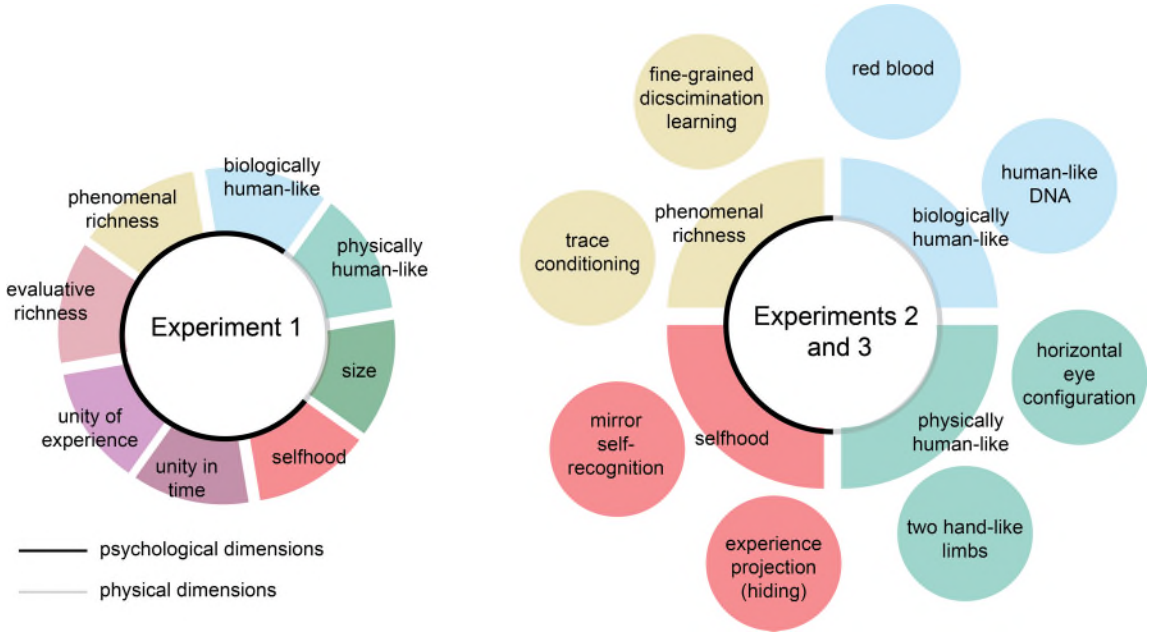

Figure 1. Manipulated dimensions. In Exp. 1, we manipulated eight dimensions: five psychological and five physical. In Exp. 2 and 3, we focused on a subset of four dimensions: phenomenal richness in vision, selfhood, physical resemblance to humans, and biological resemblance to humans, and separated each dimension into two scientific observations.

Results

The research complied with all relevant ethical regulations and was approved by the Research Ethics Committee of University College London (study ID number 1260/003).

Participants were recruited via Prolific and gave informed consent prior to their participation. They were selected based on their acceptance rate (>95%) and for being native English speakers. A total of 50 participants took part in Exp. 1, in 5 chains of 10 participants each. Participants provided meaningful descriptions of the different dimensions. All recorded responses, including verbal descriptions, are openly available on github.com/matanmazor/dimensions_of_moral_status. This first experiment was not pre-registered.

Consciousness ratings. Participants rated the consciousness level of the different alien species on a scale of 1 to 100. Overall, consciousness ratings were high, with a distinctive peak at the maximum rating of 100 (mean rating = 75.90). Feature-positive species were perceived as substantially more conscious than feature-negative species ($M = 5.81$, 95% CI [2.87,8.75], $t(49) = 3.97$, $p < .001$; see Fig. 2). This difference in consciousness ratings was significant for the dimensions phenomenal richness ($M = 10.14$, 95% CI [5.24,15.04], $t(49) = 4.16$, $p < .001$), unity ($M = 8.90$, 95% CI [4.37,13.43], $t(49) = 3.95$, $p < .001$), temporality ($M = 10.88$, 95% CI [4.75,17.01], $t(49) = 3.56$, $p < .001$), and selfhood ($M = 17.24$, 95% CI [9.56,24.92], $t(49) = 4.51$, $p < .001$). Perceived consciousness was also higher for aliens that are more biologically similar to humans ($M = 9.88$, 95% CI [4.43,15.33], $t(49) = 3.64$, $p < .001$). These effects survived a Bonferroni correction across the 8 comparisons. The effects for evaluative richness and size on consciousness ratings were significant but did not survive correction for multiple comparisons.

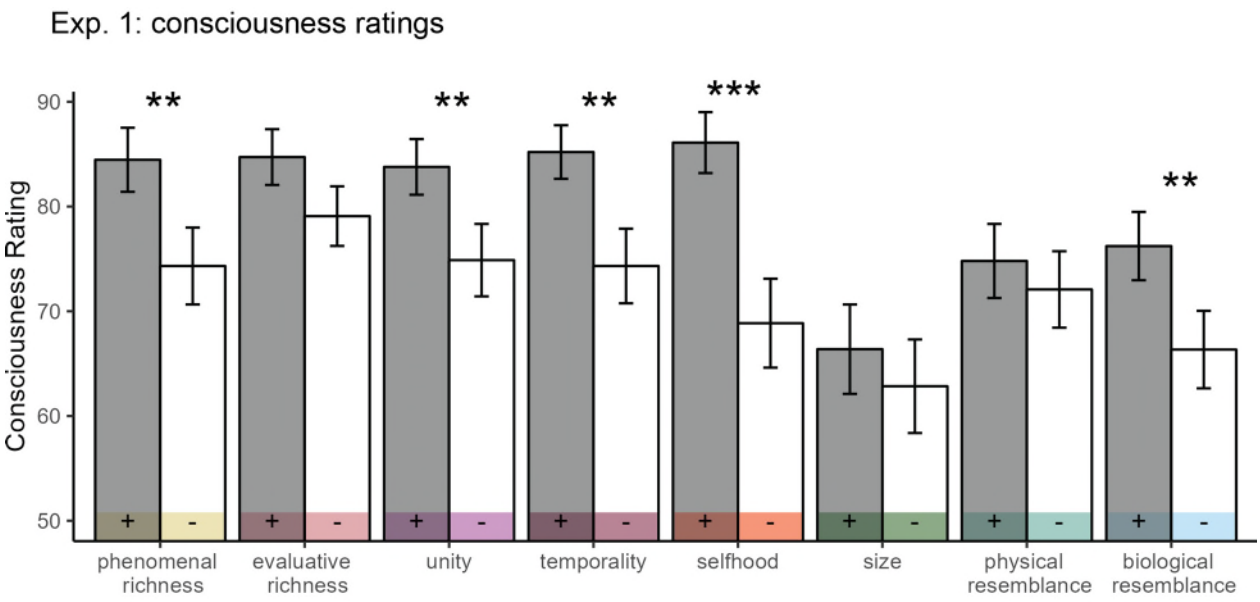

Figure 2. Mean consciousness ratings in Exp. 1. The feature-positive and feature-negative aliens are presented in grey and white, respectively. Error bars represent the standard error of the mean. Starts represent statistical significance after a Bonferroni correction for multiple comparisons. p<0.01: **, p<0.001: ***

Moral judgments. For each dimension, participants decided whether they would rather save aliens from the feature-negative or the feature-positive species. The number of feature-positive aliens (N^+) was always set to 10, and the number of feature-negative aliens (N^-) followed a staircase procedure determined by the moral judgments of previous participants — decreasing after decisions to save the feature-negative aliens, and increasing following decisions to save the feature-positive aliens (see Fig. 3A). We ran 5 short (10 participant) chains in an attempt to validate our method and establish a directional effect for our dimensions of interest. The chains are unlikely to have converged after 10 participants, so final numbers should not be interpreted as reflecting a true 'conversion rate' between feature-positive and feature-negative aliens.

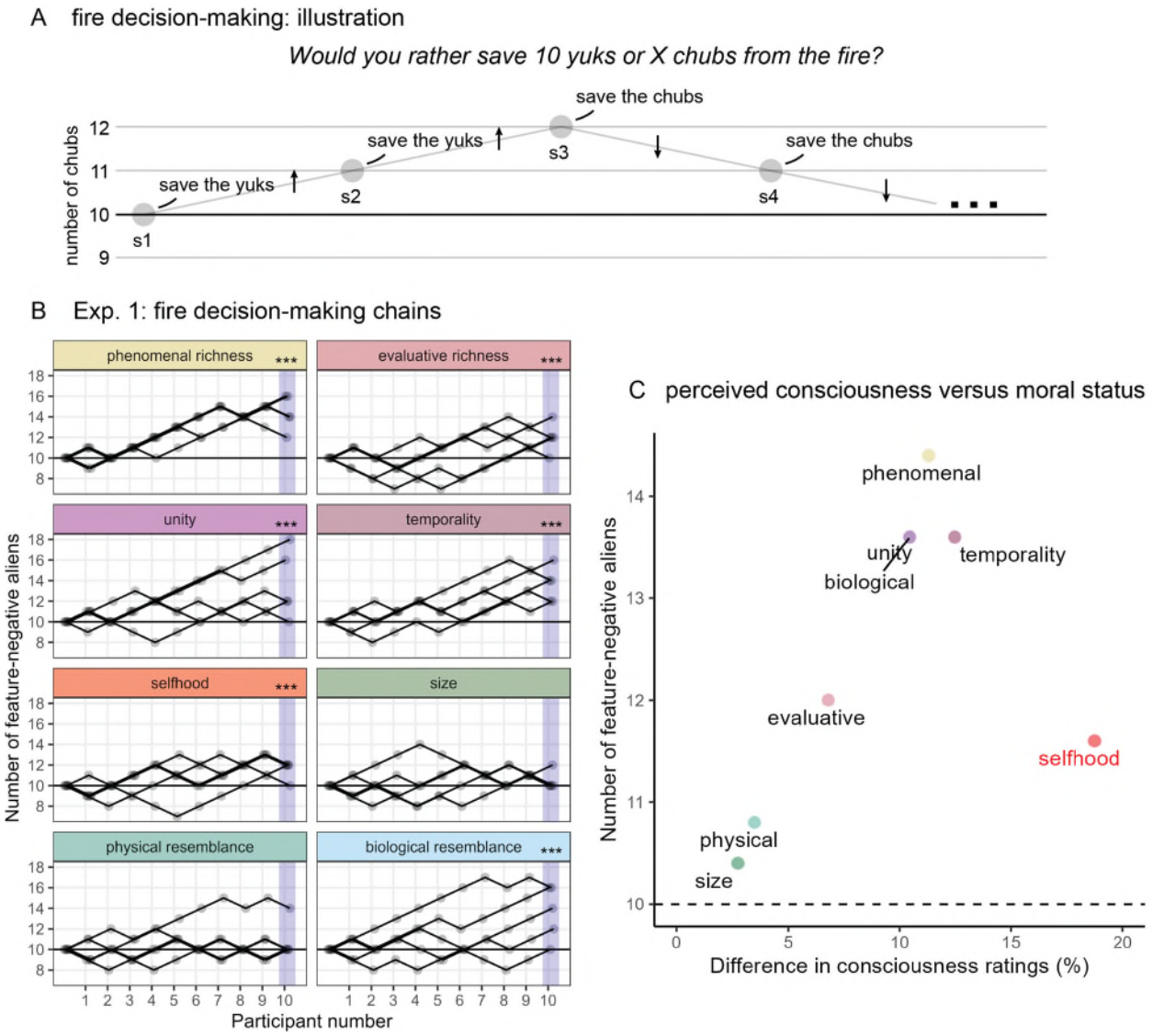

Figure 3. Exp. 1: results. A: the between-subject staircase procedure. Each circle represents one participant. The number of feature-negative aliens increased following decisions to save the feature-positive group, and decreased following decisions to save the feature-negative group. B: the number of feature-negative aliens in the five chains, for each of the eight dimensions. C: mean number of feature-negative aliens at the end of the chain, plotted against the difference in consciousness ratings between feature-positive and feature-negative aliens per individual dimensions.

We focused on the average value of N^- at the end of the chain. Once converged, N^- denotes the number of feature-negative aliens which participants consider equally deserving of

saving as 10 feature-positive aliens. N^- was equal or higher than 10 for all dimensions (see Fig. 3B). Specifically, it was 14.40 for phenomenal richness, 12.00 for evaluative richness, 13.60 for temporality and 13.60 for unity. Paralleling the effect of biological resemblance to humans on consciousness, aliens that were biologically similar to humans were also more likely to be saved from the fire (N^-=13.60).

If our participants were only considering the number of aliens in each group, without any consistent effect of the manipulated dimensions, we would expect N^- to be exactly 10 by the 10th participant. This was the case in four out of five chains for the appearance and size dimensions. As a more conservative null model, we simulated data from a cohort of participants that choose to save the larger group of aliens on 80% of the trials, and choose randomly on the remaining 20%. The average N^- for the last participant exceeded 11 (or went below 9) only on 368 of 10,000 simulations, and never exceeded 12. Hence, we find strong evidence against this conservative null-effect model for all 5 dimensions of consciousness, as well as for biological resemblance to humans, but not for size and appearance. This implies that participants' decisions of whom to save were significantly affected by descriptions of behavioural and biological attributes. We next asked whether these biases in moral decision making systematically covaried with beliefs about consciousness.

The relation between moral judgments and consciousness ratings. We show above that the perceived consciousness of imaginary aliens was informed by descriptions of behaviours which are interpreted in the scientific literature on animal consciousness as signs of phenomenal richness, evaluative richness, unity, temporality, and selfhood (Birch et al., 2020), as well as by beliefs about their biological resemblance to humans. We also showed that these same factors contributed to moral judgments about these imaginary aliens. Our experimental design

was not optimized to test for a causal link between these two findings, but it did allow us to quantify the extent to which variability in beliefs about consciousness explains variability in judgments of moral status.

Across dimensions, participants' ascriptions of moral status (high values of N^-) were associated with a more pronounced difference between consciousness ratings for the feature-positive and feature-negative alien species (see Fig. 3C). One notable exception was selfhood, which scored highest in consciousness ratings (with a mean difference of 18.75 between feature-positive and feature-negative aliens), but exerted only moderate effects on moral status judgments (N^-=11.6). Indeed, the effect on perceived consciousness correlated with N^- only when excluding the selfhood dimension ($r = .97$, 95% CI $[.81, > .99]$, $t(5) = 9.15$, $p < .001$), but not when including it ($r = .48$, 95% CI $[-.33, .89]$, $t(6) = 1.35$, $p = .224$). This is surprising in light of philosophical traditions, such as Kantian ethics, that associate moral status with selfhood and self-awareness (Kant, 1785; McMahan, 2002; Quinn, 1984; Tooley, 1972). In light of this unexpected result, in Experiments 2 and 3 we specifically examine the two operationalizations of selfhood in our study (mirror self-recognition and a tendency to hide after a shameful act), and ask whether this dissociation between consciousness ratings and moral worth judgments is common to both.

Experiment 2

In Experiment 1, we found that short descriptions of hypothetical scientific findings informed participants' attributions of consciousness to imaginary aliens as well as the value they attached to their lives. We also found evidence for a relationship between these two effects, reflected in the alignment of consciousness ratings and moral status judgments across subjects

and dimensions. One exception to this alignment was the selfhood dimension, where a strong effect on the attribution of consciousness did not translate to a strong effect on moral status judgments. Finally, while size and physical appearance had no effect on the attribution of consciousness and moral status judgments, biological resemblance to humans had strong effects on both. In Experiment 2, we zero in on these effects and focus on the dimensions of selfhood, biological resemblance to humans, phenomenal richness, and physical resemblance to humans in appearance (see Fig. 1, right panel).

Methods

Experiment 2 followed a similar procedure to Experiment 1, with the following changes. First, each alien pair corresponded to a single scientific observation: mirror self-recognition (Gallup, 1970), experience projection (hiding) (De Waal, 1986), blood color, DNA building blocks, discrimination learning (Pearce et al., 2008), trace conditioning (Clark et al., 2002), eye position (vertical or horizontal configuration) and number of limbs (2 or 5; see Supplementary Materials for full description). This way, single dimensions of consciousness now mapped to two alien pairs, each corresponding to one scientific observation. Second, participants were not given information about the way scientists interpreted the findings. Third, to simplify and shorten the experiment, we simplified some of the descriptions, and omitted all figures. Finally, in order to allow the MCMC chains to converge, we ran two longer chains of around 100 participants each. Experiment 2 was not pre-registered.

Results

A total of 212 English-speaking adults took part in the experiment, in 2 chains of 105 and 107 participants. Participants provided meaningful descriptions of the different dimensions. All

recorded responses, including verbal descriptions, are openly available on github.com/matanmazor/dimensions_of_moral_status.

Consciousness ratings. Participants reported the degree to which they believed the different alien species were conscious, on a scale of 1 to 100. Similar to Exp. 1, consciousness ratings were high, with a distinctive peak at the maximum rating of 100 (mean rating = 71.91; see Fig. 4A). Feature-positive species were perceived as substantially more conscious than feature-negative species ($M = 6.52$, 95% CI [3.92,9.12], $t(216) = 4.95$, $p < .001$). This difference in consciousness ratings was significant for all features (corrected for multiple comparisons) except for the number of limbs. Notably, eye configuration, blood colour, and the composition of DNA molecules, all affected the perceived consciousness of aliens.

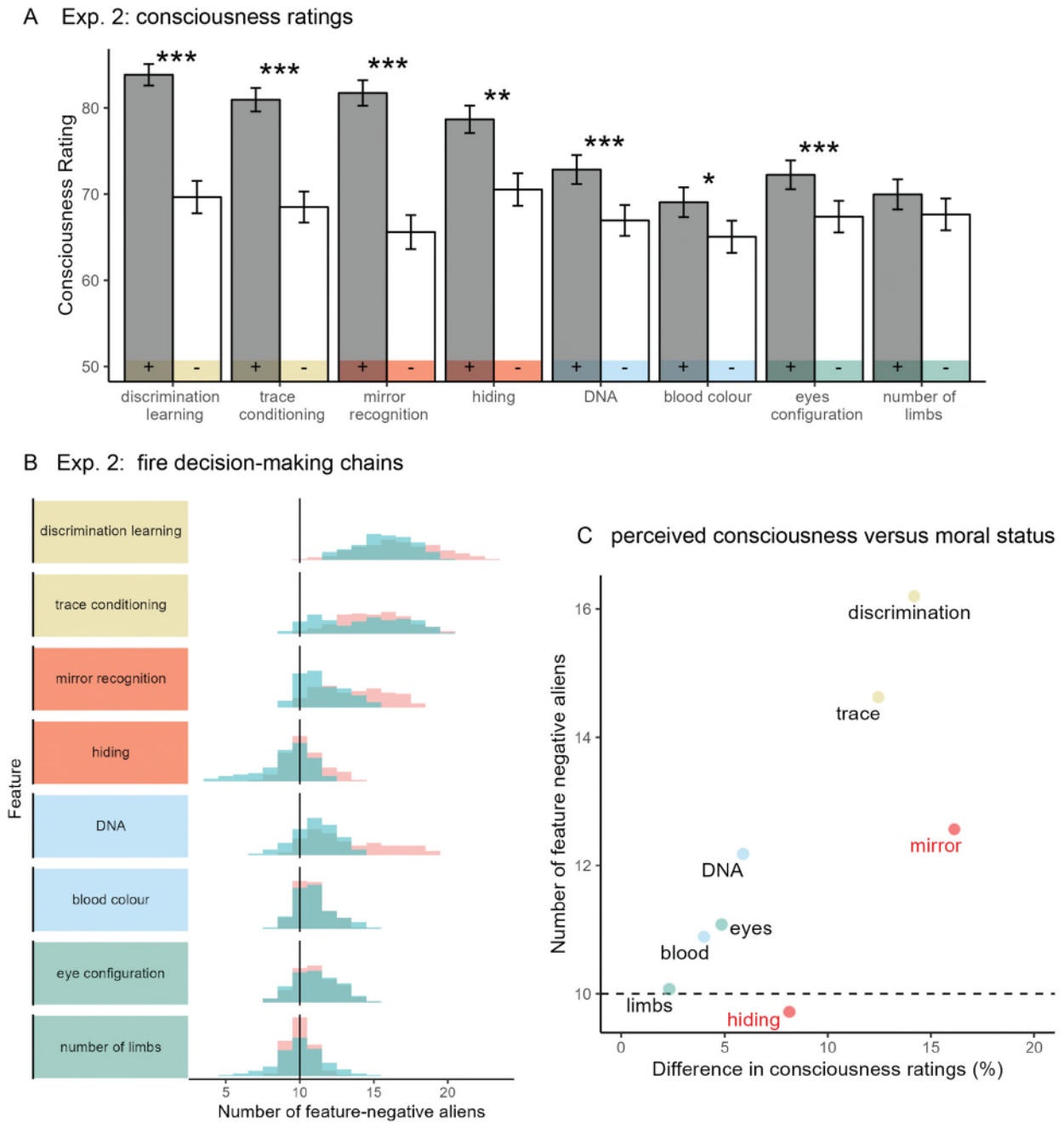

Figure 4. Exp. 2: results. A: mean consciousness rating per alien species. B: histograms of N- values across the two chains (presented in cyan and pink), excluding the first 10 steps. These values represent the number of feature-negative aliens that was equal in moral weight to 10 feature-positive aliens. C: mean number of feature-negative aliens at the end of the chain, plotted against the difference in consciousness ratings per individual dimensions. Self-awareness items are marked in red.

Moral judgments. Similar to experiment 1, participants decided whether they would rather save aliens from the feature-negative or the feature-positive species. Here also, the number of feature-positive aliens (N^+) was always set to 10, and the number of feature-negative aliens (N^-) followed the moral judgments of previous participants. Longer chains allowed us to estimate the conversion rate between feature-positive and feature-negative aliens for each of our eight features. We discarded the first 10 N^- values as a burn-in period, and took the mean of the remaining values as our estimate for N^-. Generally, we observed high levels of agreement between the two chains (see Fig. 4B). One exception was mirror recognition, with mean N^- of 13.62 and 11.52 for chains number 1 and 2.

In line with the results from Experiment 1, N^- was highest for discrimination learning (16.20) and trace conditioning (14.63), both operationalizations of the phenomenal richness dimension from Exp. 1. In other words, participants valued the lives of aliens who showed signs of visual phenomenal richness about 1.5 times more than the lives of aliens who did not show these signs. This is particularly striking for discrimination learning, where this difference in moral status was driven by the subtle fact that feature-positive, but not feature-negative aliens could "tell the difference between very similar objects (for example, very similar shades of red)".

Next, mirror self-recognition had moral value ($N^- = 12.57$, but see above caveat about convergence), whereas the tendency to hide after a shameful act (stealing from others), also typically taken as a sign of self-awareness, had a neutral moral value ($N^- = 9.72$). This was the case even though participants saw it as a reliable sign of consciousness (mean difference in consciousness ratings for hiders and non-hiders: $M = 8.18$, 95% CI $[4.04, 12.31]$). Participants also attributed greater value to the lives of aliens whose DNA was composed of the same DNA bases as human DNA ($N^- = 12.18$). Finally, eye configuration ($N^- = 11.08$), the colour of the

blood ($N^- = 10.89$), and the number of limbs ($N^- = 10.07$), all had only small to negligible effects on moral judgments.

The relation between moral judgments and consciousness ratings

Consciousness ratings were strongly aligned with N^- across features (see Fig. 5C). Similar to Experiment 1, this linear alignment was stronger when excluding the self-related mirror self-recognition and hiding after a shameful act (correlation without selfhood-related items: $r = .99$, 95% CI $[.93, > .99]$, $t(4) = 15.97$, $p < .001$; correlation with selfhood-related items: $r = .73$, 95% CI $[.05, .95]$, $t(6) = 2.61$, $p = .040$).

Experiment 3

Experiment 3 was designed to replicate the results of Exp. 2 in a pre-registered sample.

Methods

A total of 390 participants took part in Exp. 3, in three chains of 130 participants. Apart from the number of participants, it was identical to Exp. 2. A pre-registration document is available at https://doi.org/10.17605/OSF.IO/DYSQZ.

Results

Consciousness ratings. A significant difference in consciousness ratings between feature positive and feature negative aliens was observed for all alien pairs. All effects apart from the number of limbs survived a Bonferroni correction for multiple comparisons. Furthermore, an analysis of variance revealed that a difference in consciousness ratings between feature-positive and feature-negative aliens was significantly different for the 8 features of interest ($F(7,6224) =$

4.52, $MSE = 604.80$, $p < .001$). A post-hoc t-test confirmed that the difference in consciousness ratings was significantly higher in the four psychological items than in the physical ones ($M = 6.23$, 95% CI [4.87,7.59], $t(384) = 9.01$, $p < .001$).

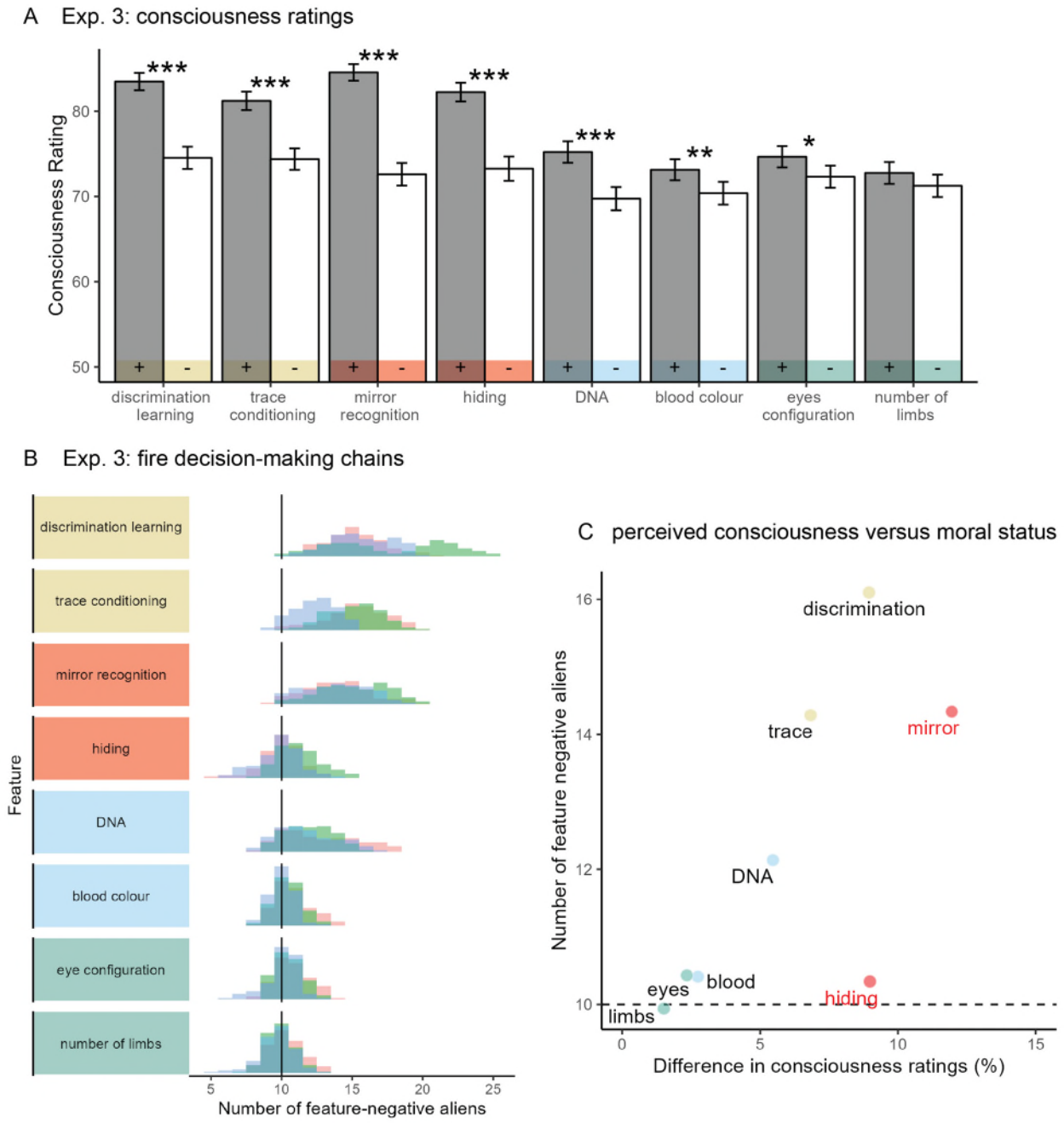

Figure 5. Exp. 3: results. A: mean consciousness rating per alien species. B: histograms of N- values across the two chains, excluding the first 30 steps. C: mean number of feature-negative aliens at the end of the chain, plotted against the difference in consciousness ratings per individual dimensions. Self-awareness items are marked in red.

Moral judgments.

To test for the effect of alien descriptions on moral status, we examined the distribution of N^- values for each alien pair. In order not to inflate our false-positive rate due to the temporally autocorrelated structure of single chains, we revised our pre-registered plan and extracted the effective sample size using Neal's approximation (Kass, Carlin, Gelman, & Neal, 1998) before running a t-test against 10. Even using this conservative method, N^- values were significantly higher than 10 for all items apart from the number of limbs and experience projection (hiding).

The relation between moral judgments and consciousness ratings.

Consciousness ratings were strongly aligned with N^- across features (see Fig. 4). Similar to Experiments 1 and 2, this linear alignment was weaker for selfhood-related features: mirror self-recognition and hiding (correlation without selfhood-related items: $r = .99$, 95% CI $[.87, > .99]$, $t(4) = 11.47$, $p < .001$; correlation with selfhood-related items: $r = .69$, 95% CI $[-.02, .94]$, $t(6) = 2.35$, $p = .057$). In a pre-registered test, we extracted the proportion of explained variance for the linear alignment between consciousness ratings and N^- values for all possible subsets of 6 items. The removal of the two self-related items, hiding and mirror self-recognition, produced the best model fit (p = .029), confirming that these two items deviated most from the linear relationship.

The value of human-like biology

Across the three experiments, we found that biological resemblance to humans had an effect on the perceived consciousness of aliens, as well as on the moral status that participants attached to them. One possibility is that these two effects have unique causes: we attach moral significance to human-like biology independent of the effects that this has on consciousness ascriptions. But an alternative is that the two effects are causally related. For example, we may

attach moral significance to human-like biology because learning that someone is biologically similar to us increases the probability that they are conscious.

In order to test if the two effects were related, we contrasted the difference in consciousness ratings between human-like and non-human-like aliens as a function of participants' decision in the fire dilemma. In all three experiments, we find that those participants who saved the biologically human-like group also thought they tended to be more conscious. (see Fig. 6). In Exp. 1, participants who chose to save aliens that resembled humans in their biological makeup also saw these aliens as more conscious ($t(37) = 4.22$, $p < .001$), but this was not true for participants who chose to save the feature-negative group ($t(16) = 1.73$, $p = .104$). The difference between the two groups was in itself significant ($\Delta M = 14.11$, 95% CI [6.95,21.26], $t(38.00) = 3.99$, $p < .001$).

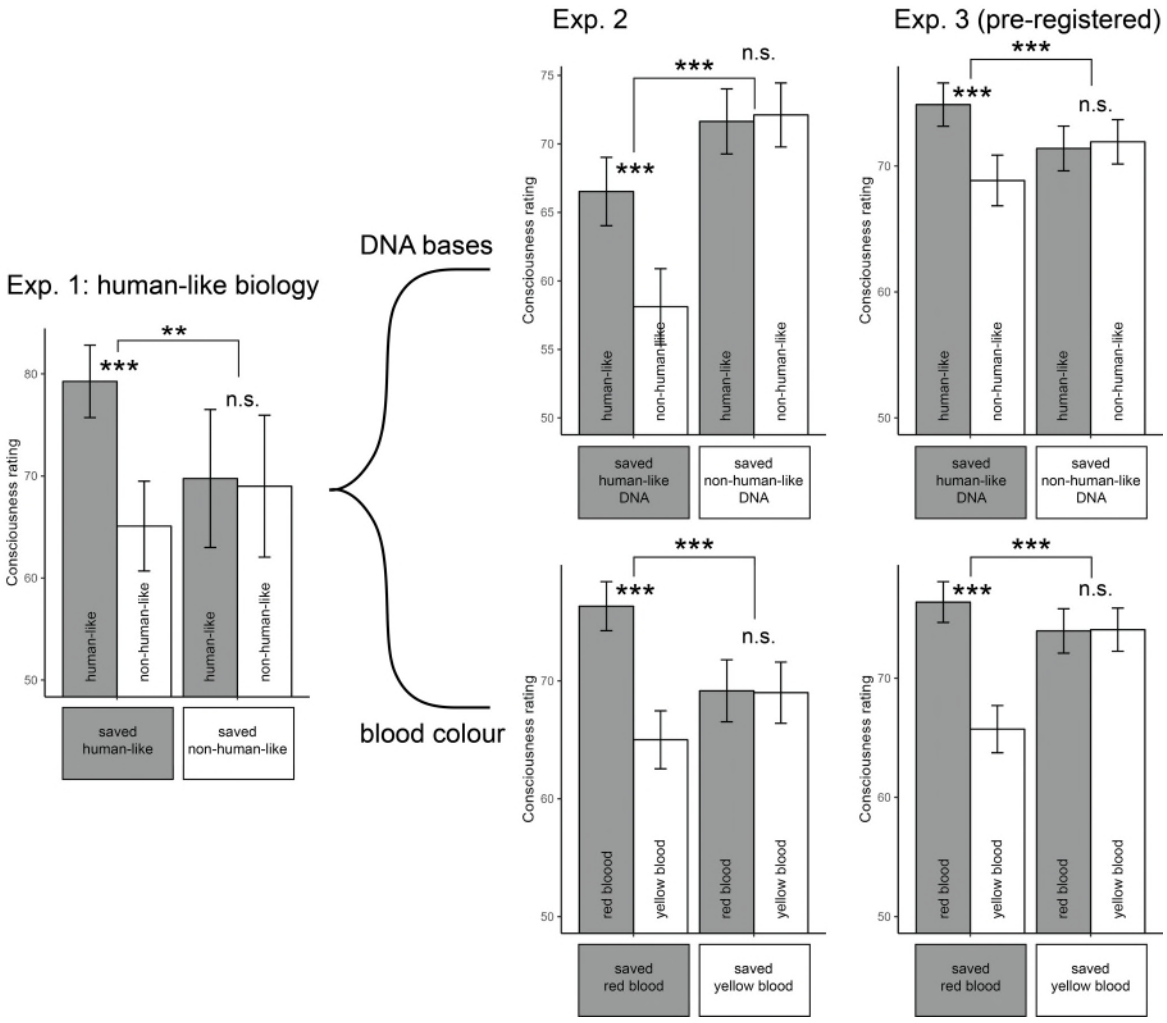

Figure 6. The value of human-like biology. Plotting consciousness ratings for the feature-positive (grey) and feature-negative (white) groups as a function of participants' decisions in the fire dilemma. A difference in consciousness ratings between human-like and non-human-like aliens appears only in the group of participants who saved the human-like aliens. Error bars represent the standard error of the mean.

The same pattern was observed when focusing on individual markers of human-like biology. Participants who chose to save aliens whose DNA bases were chemically similar to those of humans perceived them as more conscious (Exp. 2: $t(111) = 5.95, p < .001$; Exp. 3: $t(201) = 6.73, p < .001$), whereas participants who chose to save the other group showed no difference (Exp. 2: $t(105) = 0.09, p = .927$; Exp. 3: $t(187) = -0.14, p = .892$), with a

significant interaction pattern (Exp. 2: $t(215.37) = -4.31$, $p < .001$; Exp. 3: $t(291.78) = -6.10$, $p < .001$). Similarly, participants who chose to save red-blooded aliens thought they were more likely to be conscious than yellow-blooded aliens (Exp. 2: $t(109) = 3.69$, $p < .001$; Exp. 3: $t(194) = 4.66$, $p < .001$), this was not the case for participants who saved the yellow-blooded aliens (Exp. 2: $t(107) = -0.54$, $p = .592$; Exp. 3: $t(194) = -0.55$, $p = .581$), and the interaction was again significant (Exp. 2: $t(140.59) = -3.64$, $p < .001$; Exp. 3: $t(357.21) = -4.08$, $p < .001$). Together, we find a strong coupling between the effects of human-like biology on consciousness ratings and on moral decision-making.

Discussion

In three experiments, participants made moral judgments about the life and death of imaginary aliens. Using imaginary aliens allowed us to experimentally manipulate beliefs about abstract dimensions (Experiment 1) or specific features (Experiments 2 and 3), and measure their causal effect on moral decision making and beliefs about conscious experience. Both consciousness ratings and moral decisions were sensitive to our manipulation. Specifically, consciousness ratings were affected more by behavioural indicators of our five dimensions of consciousness (phenomenal and evaluative richness, unity, temporality, and selfhood, based on a taxonomy by Birch et al., 2020) than by physical attributes such as biological or physical similarity to humans. Furthermore, we found that dimensions of conscious experience had a substantial effect on participants' moral decisions. Interestingly, in all three experiments markers of low-level visual consciousness had stronger effects on moral decision-making than markers of self-awareness, despite both exerting similar effects on consciousness ratings.

Previous work has established a strong relation between moral decision making and the ascription of mental attributes such as intelligence, experience, or agency. For example, beliefs about the capacity for subjective experience were associated with a desire to avoid harm (H. M. Gray et al., 2007). Similarly, being told that an animal or an alien was more intelligent made participants more likely to report that it was not OK to eat them (Piazza & Loughnan, 2016), and this association between perceived intelligence and moral worth was apparent already in young children (Wilks et al., 2021). We build and expand on these findings in two ways. First, adopting a fine-grained taxonomy of dimensions of conscious experience (Birch et al., 2020) revealed different effects for different dimensions on moral status, with a surprisingly strong effect for visual perceptual capacities. Second, Experiments 2 and 3 provided evidence that people extract information about conscious experience from behavioural observations, similar to the interpretation of behavioural findings by comparative psychologists, and that they use this information to guide their moral decision making. This second finding is especially important in light of the debates over the moral significance of functional versus phenomenal aspects of consciousness (Carruthers, 2019; Danaher, 2020; Levy, 2014), and in the context of ongoing debates about the possibility of AI consciousness (Birch, 2024; Colombatto & Fleming, 2024; Danaher, 2020).

In both experiments we found a strong alignment between consciousness ratings and moral status, with the single exception of selfhood and its constituent operationalizations, which contributed to consciousness ratings much more than to moral status. Specifically, mirror recognition had the strongest effect on perceived consciousness of all items in Experiments 2 and 3, but its effect on moral decision-making was consistently lower than that of visual discrimination learning, that is, the capacity to tell between very similar objects. While we are not aware of any study that directly examined the effects of perceived self awareness on moral status,

the idea that moral worth is based on self-awareness dates back at least to Emmanuel Kant (1785). In a striking contrast with this notion, here we find that beliefs about other aspects of consciousness such as visual awareness and working memory are much more influential for moral decision-making than beliefs about selfhood. Such findings are more in line with the utilitarian views of Mill (1861) and Bentham (1789), and more recently with the view of Shepherd (2018).

Our participants placed moral significance not only on behavioural markers of private mental states, but also on biological resemblance to humans, both in blood colour and in the chemical composition of DNA. Participants also rated biologically human-like aliens as more conscious than ones that were biologically dissimilar to humans. These two effects were not observed for surface-level differences in size and eye configuration. The ethical significance people place on biological similarity to humans echoes previous findings of a common speciesism bias (Caviola, Everett, & Faber, 2019), and may suggest a hard boundary on the moral status that people would be willing to attribute to artificial intelligences whose physical implementation is substantially different from our own.

Across participants, the effects of biological similarity to humans on perceived consciousness and on moral standing were strongly coupled. This coupling can reflect a mediating role for beliefs about consciousness in the effects of human-like biology on moral decision-making, such that human-like biology contributes to moral status only to the extent that it affects beliefs about conscious experience (in line with perceived consciousness being the essence of morality, K. Gray et al., 2012). Alternatively, it may be driven by a third factor that contributes to both beliefs about consciousness and moral status. For example, participants with specific religious convictions may prioritise the lives of more human-like individuals, and hold

an independent belief that human-like biology correlates with consciousness. Finally, it may be that participants who sacrificed the more human-like aliens (due to their smaller number) were motivated to report that there is no difference in consciousness between the two groups (Piazza & Loughnan, 2016). For this last option, it is the moral decision that affects beliefs about the presence of consciousness, rather than the other way around. In the Supplementary Materials we report the result of a pre-registered experiment in which we failed to find evidence for a mediating role for beliefs about consciousness in the effects of human-like biology on moral decision-making. We leave it for future work to determine the mechanisms underlying the three-way interplay between biological similarity to humans, beliefs about the presence of consciousness and moral status.

Several limitations of the present work deserve mention. First, the phenomenal richness dimension focused solely on visual perception, rather than other modalities such as touch, smell, or hearing. Whether the surprisingly strong effect of phenomenal richness on moral status would generalize to these other modalities remains an open question. Second, despite the strong coupling between the effects of biological similarity to humans on consciousness ratings and on moral decision-making, a pre-registered experiment (described in the Supplementary Materials) failed to find evidence that consciousness ascriptions causally mediate the effect of biological similarity to humans on moral status, leaving the underlying mechanism an open question for future work. Finally, our participants were English-speaking adults recruited via Prolific, limiting our ability to draw strong inferences about the universality of these effects.

We would like to end by pointing out that our findings speak to the ethical significance of cognitive science, and consciousness science specifically (Mazor et al., 2023). Even seemingly neutral findings such as the capacity to learn associations between events separated in time, or the

ability to tell apart similar stimuli, had strong effects on people's decisions about hypothetical questions of life and death. Indeed, integration of empirical findings from research on invertebrates was recently used to argue for a change to their legal status (Birch, Burn, Schnell, Browning, & Crump, 2021), and a recent report about the possibility of AI consciousness, signed by leading consciousness scientists, discussed the moral risks of over- or under-attributing consciousness to AI (Butlin et al., 2023). As scientists continue to unravel the mysteries of minds, brains and machines, their findings will inevitably inform and change whose interests we value as a society.

References

Balasko, M., & Cabanac, M. (1998). Motivational conflict among water need, palatability, and cold discomfort in rats. *Physiology & Behavior*, *65*(1), 35–41. https://doi.org/10.1016/S0031-9384(98)00090-0

Bentham, J. (1789). *The Principles of Morals and Legislation.*

Birch, J. (2024). *The edge of sentience: Risk and precaution in humans, other animals, and ai.* Oxford: Oxford University Press.

Birch, J., Burn, C., Schnell, A., Browning, H., & Crump, A. (2021). Review of the evidence of sentience in cephalopod molluscs and decapod crustaceans. *General - Animal Feeling*. Retrieved from https://www.wellbeingintlstudiesrepository.org/af_gen/2

Birch, J., Schnell, A. K., & Clayton, N. S. (2020). Dimensions of Animal Consciousness. *Trends in Cognitive Sciences*, *24*(10), 789–801. https://doi.org/10.1016/j.tics.2020.07.007

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., … VanRullen, R. (2023.). *Consciousness in artificial intelligence: Insights from the science of consciousness*. https://doi.org/10.48550/arXiv.2308.08708

Carruthers, P. (2019). *Human and Animal Minds: The Consciousness Questions Laid to Rest* (1st ed.). Oxford University Press. https://doi.org/10.1093/oso/9780198843702.001.0001

Caviola, L., Everett, J. A. C., & Faber, N. S. (2019). The moral standing of animals: Towards a psychology of speciesism. *Journal of Personality and Social Psychology*, *116*(6), 1011–1029. https://doi.org/10.1037/pspp0000182

Clark, R. E., Manns, J. R., & Squire, L. R. (2002). Classical conditioning, awareness, and brain systems. *Trends in Cognitive Sciences*, *6*(12), 524–531. https://doi.org/10.1016/S1364-6613(02)02041-7

Clayton, N. S., & Dickinson, A. (1998). Episodic-like memory during cache recovery by scrub jays. *Nature*, *395*(6699), 272–274. https://doi.org/10.1038/26216

Colombatto, C., & Fleming, S. M. (2024). Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness*, *2024*(1), niae013. https://doi.org/10.1093/nc/niae013

Danaher, J. (2020). Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism. *Science and Engineering Ethics*, *26*(4), 2023–2049. https://doi.org/10.1007/s11948-019-00119-x

Dawkins, M. S. (2003). Behaviour as a tool in the assessment of animal welfare1. *Zoology*, *106*(4), 383–387. https://doi.org/10.1078/0944-2006-00122

De Waal, F. (1986). Deception in the natural communication of chimpanzees. *Deception: Perspectives on Human and Nonhuman Deceit*, 221244.

Gallup, G. G. (1970). Chimpanzees: Self-recognition. *Science*, *167*(3914), 86–87. https://doi.org/10.1126/science.167.3914.86

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of Mind Perception. *Science*, *315*(5812), 619–619. https://doi.org/10.1126/science.1134475

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*(2), 101–124. https://doi.org/10.1080/1047840X.2012.651387

Hillemann, F., Bugnyar, T., Kotrschal, K., & Wascher, C. A. F. (2014). Waiting for better, not for more: Corvids respond to quality in two delay maintenance tasks. *Animal Behaviour*, *90*, 1–10. https://doi.org/10.1016/j.anbehav.2014.01.007

Jaworska, A., & Tannenbaum, J. (2013). *The Grounds of Moral Status*. Retrieved from https://plato.stanford.edu/eNtRIeS/grounds-moral-status/

Kabadayi, C., & Osvath, M. (2017). Ravens parallel great apes in flexible planning for tool-use and bartering. *Science*, *357*(6347), 202–204. https://doi.org/10.1126/science.aam8138

Kant, I. (1785). Groundwork of the metaphysics of morals. *The Classics of Western Philosophy: A Reader's Guide*, 346.

Kass, R. E., Carlin, B. P., Gelman, A., & Neal, R. M. (1998). Markov chain monte carlo in practice: A roundtable discussion. *The American Statistician*, *52*(2), 93–100. https://doi.org/10.1080/00031305.1998.10480547

Levy, N. (2014). The Value of Consciousness. *Journal of Consciousness Studies : Controversies in Science & the Humanities*, *21*(1-2), 127. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4001209/

Mazor, M., Brown, S., Ciaunica, A., Demertzi, A., Fahrenfort, J., Faivre, N., … Lubianiker, N. (2023). The Scientific Study of Consciousness Cannot and Should Not Be Morally Neutral. *Perspectives on Psychological Science*, *18*(3), 535–543. https://doi.org/10.1177/17456916221110222

McMahan, J. (2002). *The ethics of killing: Problems at the margins of life*. Oxford University Press, USA.

Mill, J. S. (1861). 1969. utilitarianism. *Collected Works of John Stuart Mill*, *10*, 203259.

Narins, P. M., Grabul, D. S., Soma, K. K., Gaucher, P., & Hödl, W. (2005). Cross-modal integration in a dart-poison frog. *Proceedings of the National Academy of Sciences*, *102*(7), 2425–2429. https://doi.org/10.1073/pnas.0406407102

Ortega, L. J., Stoppa, K., Güntürkün, O., & Troje, N. F. (2008). Limits of intraocular and interocular transfer in pigeons. *Behavioural Brain Research*, *193*(1), 69–78. https://doi.org/10.1016/j.bbr.2008.04.022

Pearce, J. M., Esber, G. R., George, D. N., & Haselgrove, M. (2008). The nature of discrimination learning in pigeons. *Learning & Behavior*, *36*(3), 188–199. https://doi.org/10.3758/LB.36.3.188

Piazza, J., & Loughnan, S. (2016). When Meat Gets Personal, Animals' Minds Matter Less: Motivated Use of Intelligence Information in Judgments of Moral Standing. *Social*

Psychological and Personality Science, *7*(8), 867–874. https://doi.org/10.1177/1948550616660159

Quinn, W. (1984). Abortion: Identity and loss. *Philosophy & Public Affairs*, 2454.

Reimert, I., Fong, S., Rodenburg, T. B., & Bolhuis, J. E. (2017). Emotional states and emotional contagion in pigs after exposure to a positive and negative treatment. *Applied Animal Behaviour Science*, *193*, 37–42. https://doi.org/10.1016/j.applanim.2017.03.009

Sanborn, A. N., Griffiths, T. L., & Shiffrin, R. M. (2010). Uncovering mental representations with markov chain monte carlo. *Cognitive Psychology*, *60*(2), 63–106. https://doi.org/10.1016/j.cogpsych.2009.07.001

Shepherd, J. (2018). *Consciousness and Moral Status* (1st ed.). Routledge. https://doi.org/10.4324/9781315396347

Tooley, M. (1972). Abortion and infanticide. *Philosophy & Public Affairs*, 3765.

Wilks, M., Caviola, L., Kahane, G., & Bloom, P. (2021). Children Prioritize Humans Over Animals Less Than Adults Do. *Psychological Science*, *32*(1), 27–38. https://doi.org/10.1177/0956797620960398

Supplementary Materials

Materials and procedure: Exp. 1

After giving their consent to participate, participants were given the following instructions:

> In this experiment, you will read about imaginary aliens and make some decisions about them.

In the distant future, scientists go on an expedition to another planet and discover several new species of aliens. All aliens on the planet were found to have two eyes and a set of hand-like limbs. The aliens also feed on the sweet space-berries that grow on the planet. The scientists spend several months studying the different aliens, letting them solve different games and tasks, measuring their bodies, and comparing their characteristics. The scientists sort the alien species into 8 pairs. Within each such pair, the scientists find the two species to be identical in all characteristics except for two differences. In the rest of this experiment, we will describe these pairs to you and the differences between them, and ask you to answer questions about them.

Please read the descriptions carefully and answer the questions as accurately as possible.

Participants were then presented with eight pairs of alien species, in random order. Each pair was described in two screens of text and visual illustrations (described below in square brackets). Species were given one-syllable names (for example, Geks and Rubs), fully randomised across participants. For clarity, we refer to the two species as "Positives" and "Negatives":

1. Phenomenal Richness (in vision)

Positives were found to be almost identical Negatives except for two facts. Positives are able to see things and to learn simple rules. They can also tell the difference between very similar objects (for example, very similar shades of red). Negatives are able to see things and to learn simple rules. They are however unable to tell the difference between very similar objects [illustration of two similar shades of red].

Also, in one experiment the scientists trained the aliens to associate a flash of light with a treat. The scientists varied the time the aliens had to wait between the light and the treat. Positives were able to make the association between the light and the treat, even when the treat was given as long as one second after the light. Negatives were also able to make the association between the light and the treat, but only when the treat was given immediately. Researchers use this test as a measure of the ability to hold information in memory for short periods of time [a bar plot showing "Learned association" for the two alien species and with or without a 1 second delay].

2) Evaluative richness

Positives were found to be almost identical to Negatives except for two facts. Positives are able to have good and bad moods. When they are shown an angry face, they keep looking at the face if they are in a good mood, but they look away from the face if they are in a bad mood. Negatives are also able to have good and bad moods. When they are shown an angry face, they keep looking at the face regardless of their current mood. Researchers use this test as a measure of the effect of emotions on the aliens' behaviour [a bar plot showing "Time looking at angry face" in the two moods and alien types].

Also, in one experiment, the scientists placed the aliens in a room and gave them space berries to eat, while slowly reducing the temperature of the room. In one condition, the space-berries were very sweet, and in a second condition they were not sweet at all. Positives like sweet space berries and dislike the cold. They left the room when it reached a temperature of 12 degrees in the unsweet berries condition, but left later, when the room reached a temperature of 8 degrees in the sweet berries condition. Negatives also like sweet space berries and dislike the cold. They left the room when temperature reached 12 degrees, regardless of the taste of the berries. Researchers use this test as a measure of the ability to weigh different needs against each other. [a line plot showing "temp. in room" decreasing as a function of time, and a bar plot showing "temp. when leaving room" for the two species and berry types].

3) Unity of experience

Positives were found to be almost identical to Negatives except for two facts. When Positives are shown a printed image of a new fruit, they can later recognize that fruit by touch alone, even if they haven't touched it before. When Negatives are shown a printed image of a new fruit, they can't later recognize that fruit by touch alone unless they have touched it before. Researchers use this test as a measure of the ability to combine information from different senses [an illustration of a line drawing of a pineapple accompanied by "sight, no touch" icons, and a bowl of fruit, including a pineapple, accompanied by "touch, no sight" icons].

Also, in one experiment the scientists covered the right eye of the aliens, and then trained them to make a sound when they see a picture of a flower. The aliens were then shown pictures of flowers to test their learning. Positives made a squeak sound when a picture of a flower was presented to their trained left eye or to their untrained right eye. Negatives made a squeak sound when a picture of a flower was presented to their trained left

eye, but not when a picture of a flower was presented to their untrained right eye. Researchers used this task as a measure of the ability to integrate information between both eyes [a schematic illustration of the experiment].

4) Unity in time

Positives were found to be almost identical to Negatives except for two facts. In one experiment, the researchers showed the aliens a locked glass box, with a triangle-shaped keyhole. Inside the box were space berries that the aliens could see and smell. The following day, the researchers hid the box and allowed the aliens to choose between one of 3 different shaped keys: triangular, square and circular. Positives were able to see the space berries through the glass and understand that they needed a triangular key to open the box. On the following day, Positives mostly favoured picking the triangular key, as if they planned to use it for opening the box later. Negatives were also able to see the space berries through the glass and understand that they needed a triangular key to open the box. However, on the following day, Negatives chose indiscriminately between the three keys [an illustration of a transparent box with a triangle-shaped keyhole filled with berries labelled “Day 1”, and of a triangle, a circle and a square, labelled “Day 2”].

In another experiment, the aliens were presented with the option of picking and eating space berries immediately, or waiting 2 days to eat the berries, when they are ripe and sweeter. Negatives prefer the larger, sweeter space berries. However, when given unripe berries, Negatives tended to eat the fruit immediately, and they don’t wait for the berries to ripen. Positives also prefer the larger, sweeter space berries. When given unripe berries, Positives save them for later, until they are ripe and ready to eat. The scientists interpreted these experiments as measuring the capacity to make plans for the future [an illustration of a tree with three small red fruit labelled “Day 1”, and a tree with three large fruit labelled “Day 2”].

5) Selfhood

Positives were found to be almost identical to Negatives except for two facts. Positives like to keep themselves clean. When Positives see a fleck of dirt on their forehead in a mirror, they act to remove it by rubbing their face. Negatives also like to keep themselves clean. However, when Negatives see a fleck of dirt on the forehead in a mirror, they ignore the mirror [an illustration of a mirror]

Positives live in small groups, and they sometimes steal food from their neighbours. When caught doing so, they attempt to hide themselves. Negatives also live in small groups, and they also steal food from their neighbours. However, they do not attempt to hide themselves when caught [an illustration of two eyes hiding in a dark hole].

6) Eye configuration

Positives were found to be almost identical to Negatives except for two facts. Positives have right and left eyes. Negatives also have two eyes, but they are situated one above the other, rather than side by side [an illustration of two minimal faces, consisting of an oval shape and two eyes, configured side by side for one and one above the other for the other].

Positives have two arm-like limbs, whereas Negatives have five [an illustration of the same two faces, the first with two hands and the second with 5].

7) Size

Positives were found to be almost identical to Negatives except for two facts. Negatives weigh around 1 gram, about the same as a paperclip, whereas Positives weigh around 45kg (7 stone), around the average weight of a punching bag [an illustration of a paperclip and a punching bag].

Negatives are around 1 cm tall (0.4 inch), whereas Positives are around 120 cm tall (4 feet) [a bar plot showing height (in cm.) for the two species].

8) Biological similarity to humans

Positives were found to be almost identical to Negatives except for two facts. Positive genes are made up of the same four bases that make up human DNA, ACTG. The genes of Negatives are made up of bases that bear no chemical resemblance to human DNA [an illustration of four coloured shapes labelled “human bases”, the same four shapes labelled with the name of the feature-positive species, and of other shapes labelled with the name of the feature-negative species].

Positive blood is red, whereas Negative blood is yellow [an illustration of two pippettes: one filled with a red liquid and the other with yellow liquid].

After reading the descriptions, participants were asked to use their own words to describe the difference between the two species. They were then presented with the following dilemma:

A fire started on the planet. A group of 10 Positives was caught in the fire, as well as a group of 11 Negatives. The scientists can only save one group.

(1) Save the 10 Positives, let the 11 Negatives die.
(2) Save the 11 Negatives, let the 10 Positives die.

Crucially, the number of feature-negative aliens (here, 11) was determined in a between-subject staircasing procedure as described in the main text.

In the next screen, participants were asked to rate the accuracy of two statements, going from “Not at all true” to “Completely true”, using a slider:

Positives are conscious

and

Negatives are conscious

Within each pair, participants were allowed to move between screens and change their previous decisions. Once finished, they submitted their responses and moved to the next pair.

At the end of the experiment, participants were asked to share any thoughts they had about the experiment.

Materials and procedure: Exp. 2 and 3

Experiments 2 and 3 were identical in structure to Exp. 1, except for the alien descriptions. First, alien descriptions consisted of text alone, without illustrations. Second, each alien pair corresponded to a single observation rather than two. Finally, we did not remark on the scientists’

interpretation of the difference between the two aliens (for example, as revealing differences in the effect of emotion on decision-making):

1. Discrimination learning

Positives were found to be almost identical Negatives except for one fact. Positives are able to see things and to learn simple rules. They can also tell the difference between very similar objects (for example, very similar shades of red). Negatives are able to see things and to learn simple rules. They are however unable to tell the difference between very similar objects.

2. Trace conditioning

Positives were found to be almost identical Negatives except for one fact. In one experiment the scientists trained the aliens to associate a flash of light with a treat. Positives were able to make the association between the light and the treat, even when the treat was given as long as one second after the light. Negatives were also able to make the association between the light and the treat, but only when the treat was given immediately.

3. Mirror self-recognition

Positives were found to be almost identical to Negatives except for one fact. Positives like to keep themselves clean. When Positives see a fleck of dirt on their forehead in a mirror, they act to remove it by rubbing their face. Negatives also like to keep themselves clean. However, when Negatives see a fleck of dirt on the forehead in a mirror, they ignore the mirror.

4. Hiding

Positives were found to be almost identical Negatives except for one fact. Positives live in small groups, and they sometimes steal food from their neighbours. When caught doing so, they attempt to hide themselves. Negatives also live in small groups, and they also steal food from their neighbours. However, they do not attempt to hide themselves when caught.

5. Eye configuration

Positives were found to be almost identical to Negatives except for one fact. Positives have right and left eyes. Negatives also have two eyes, but they are situated one above the other, rather than side by side.

6. Number of limbs

Positives were found to be almost identical Negatives except for one fact. Positives have two arm-like limbs, whereas Negatives have five.

7. DNA building blocks

Positives were found to be almost identical to Negatives except for one fact. Positive genes bear no resemblance to human genes. They are however made up of the same building blocks that make up human DNA. The genes of Negatives also bear no resemblance to human genes. They are made up of building blocks that bear no chemical resemblance to human DNA.

8. Blood colour

Positives were found to be almost identical Negatives except for one fact. Positive blood is red, whereas Negative blood is yellow.

Additional Experiment: testing the causal role of beliefs about consciousness

In Exp 1-3, we found that participants were more likely to save aliens that are biologically similar to humans, for example ones with a similar blood colour. We also found that participants associated biological similarity to humans with higher levels of consciousness, and that the two effects were correlated: those participants who saved biologically-similar aliens judged them to be more conscious. In this experiment we asked whether the effect of biological similarity on moral status is causally dependent on the effect of biological similarity on the ascription of consciousness. A pre-registration document is available at https://doi.org/10.17605/OSF.IO/AUCYP.

The research complied with all relevant ethical regulations, and was approved by the UCL Research Ethics Committee (Project ID Number: EP_2024_004). Participants were recruited via Prolific, and gave their informed consent prior to their participation. To be eligible to take part in this study, their Prolific approval rate had to be 95% or higher, their reported first language

English, and their age between 18 and 60. We collected data from 517 participants, out of which 502 passed the comprehension check (251 in each group). The entire experiment took around 2.5 minutes to complete. Participants were paid £0.4 for their participation, equivalent to an hourly wage of £8.

Procedure. Participants were presented with the following instructions:

In this experiment we will describe a hypothetical scenario, test your understanding, and then ask you to make one decision.

In the distant future, scientists go on an expedition to another planet and discover three new species of aliens.

All aliens on the planet were found to have two eyes and a set of hand-like limbs. The aliens are not harmful to humans, and there are many of them.

The scientists spend several months studying the different aliens, letting them solve different games and tasks, measuring their bodies, and comparing their characteristics.

They identify three species of aliens, which are exactly identical, except for one difference: the colour of their blood.

Aliens of the first species have green blood, aliens of the second species have yellow blood, and aliens of the third species have red blood.

In addition, one in every two participants (the "temperature" group) has been given the following additional information:

Importantly, despite this difference in blood colour, the scientists found that the temperature of their blood is exactly the same.

Whereas the other group (the "consciousness" group) has been given the following information:

Importantly, despite this difference in blood colour, the scientists found that the level of consiousness of the three groups of aliens is exactly the same. They experience the world in the exact same way.

Comprehension question. To ensure comprehension, participants were asked the following question:

What is the difference between the three species of aliens that were found on the planet?

The correct answer is “only the colour of their blood”.

Fire dilemma. Participants were presented with the following dilemma:

A fire started on the planet. Three groups of aliens were caught in the fire: a group of 10 yellow-blooded aliens, a group of 11 green-blooded aliens, and a group of 10 red-blooded aliens.

Unfortunately, the scientists can only save one group of aliens from the fire. What should they do?

Participants were given the choice between saving 11 green-blood aliens, 10 red-blooded aliens, or 10 yellow-blooded aliens.

Finally,they were asked to explain their decision.

Results. This study was designed to test the hypothesis that the moral worth of red-bloodedness is mediated by beliefs about consciousness.

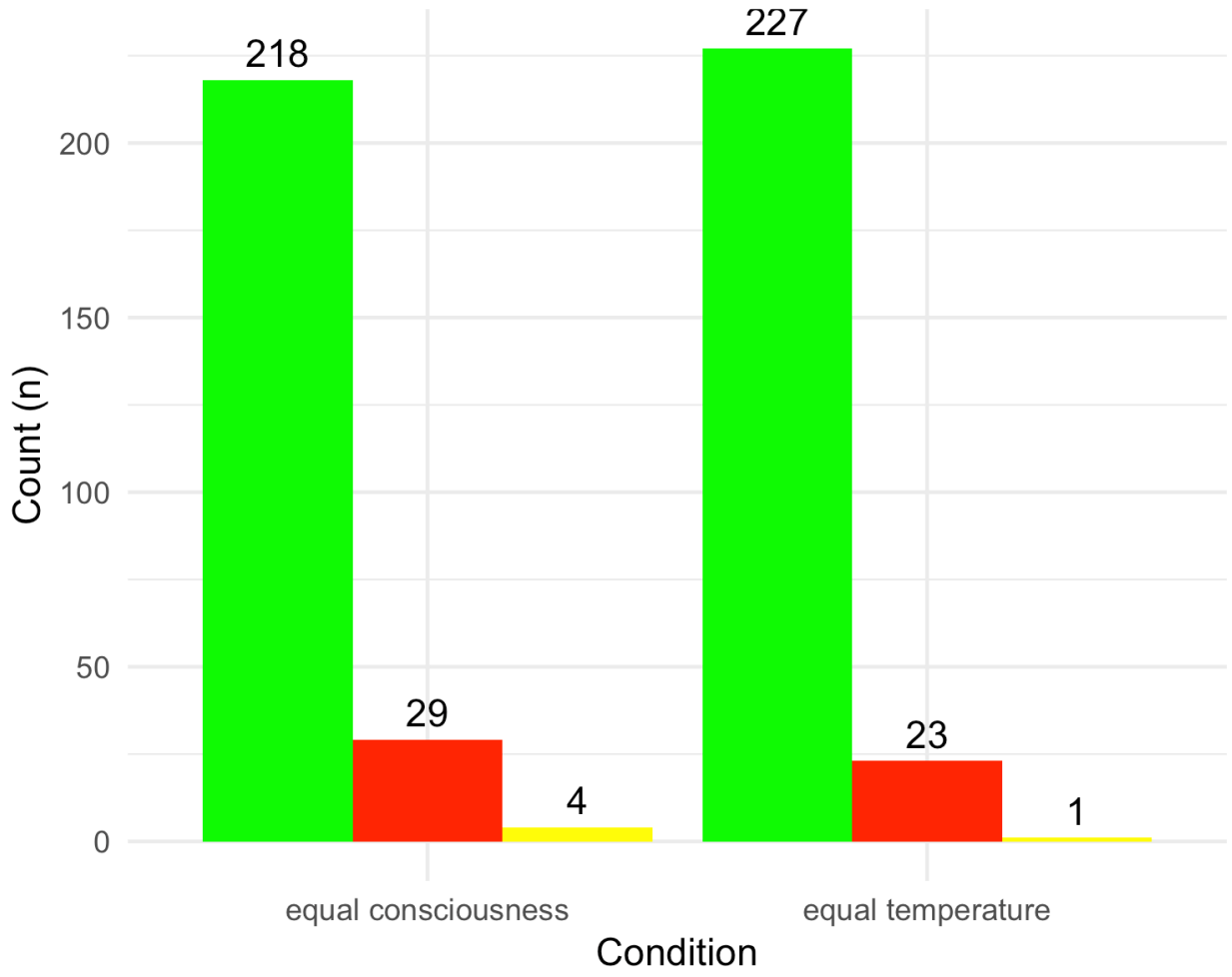

Figure 7. Number of participants who chose to save the green, red and yellow-blooded groups as a function of condition.

Hypothesis 1 (A red-blood bias): We tested whether participants chose to save the red-blooded aliens more often than the yellow-blooded aliens. Out of 57 participants who saved the red- or the yellow-blooded groups, 52 participants chose to save the red-blooded group, indicating a strong preference for human-like blood-colour ($p < .001$).

Hypothesis 2 (red-blood bias by group): We tested whether participants in the "equal temperature" group chose to save the red-blooded aliens more often than those in the "equal

consciousness" group. Out of 251 participants in the "equal consciousness" group, 29 saved the red-blooded aliens. Out of 251 participants in the "equal temperature" group, 23 saved the red-blooded aliens. The difference between the two proportions was not significant in a chi squared test (p = .464). Descriptively, the results are trending in the opposite direction to our hypothesis, such that the preference for red-blooded aliens was stronger in the "equal consciousness" group.

Hypothesis 3 (red-blood bias in consciousness group): We tested whether participants in the "equal consciousness" group showed no red-blood bias. Out of 33 participants who saved the red- or the yellow-blooded groups, 29 participants chose to save the red-blooded group, indicating a strong preference for human-like blood-colour in the equal consciousness group (p < .001). This indicates that participants prioritised the lives of red-blooded aliens even when they were explicitly told that blood colour was unrelated to consciousness.