Dimensions of Moral Status

Matan Mazor1, Ariana Risoli2, Anna Eberhardt2, & Stephen M. Fleming2,3,4

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

2 Department of Experimental Psychology, University College London

3 Wellcome Centre for Human Neuroimaging, University College London

4 Max Planck UCL Centre for Computational Psychiatry and Aging Research

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](mailto:matan.mazor@psy.ox.ac.uk)

Abstract

Whether someone is conscious or not has important ethical and legal implications. For example, animal welfare regulations seek to minimize subjective suffering in sentient animals, and discussions about AI ethics focus on the possibility of artificial consciousness as a key concern. However, the presence of consicousness cannot be directly perceived, but must be inferred from external cues such as behaviour and biological makeup. Furthermore, consciousness can vary along more than one dimension, with some potentially more ethically relevant than others. In three experiments (one pre-registered, total N=652), we find that people form beliefs about consciousness based on behaviours that are considered markers of consciousness in the scientific literature. Using a between-subject staircasing procedure, we then show that beliefs about consciousness contribute to the moral status of imagined aliens, and that such beliefs partly explain the moral significance people attach to biological similarity to humans. Finally, we find that markers of visual awareness contribute more to moral status than markers of self-awarensss, 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

*Word count:* 5,293

Dimensions of Moral Status

# Introduction

An entity has *Moral Status* if and only if “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 agents. 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 based 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 behaviour of other agents, and need to infer their internal states from these third-person observations. In other words, regardless of whether moral status *should* be based on behaviour or experience, in practice it *must* be based on behaviour, because we can never directly perceive the subjective experience of others.

In psychology, behavioural observations are commonly used to learn about internal mental processes, which are often assumed to have an experiential nature for the subject. In comparative psychology, 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 known examples are the study of caching behaviour as a measure of episodic memory (Clayton & Dickinson, 1998) and the use of trace-conditioning 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.

Here we asked whether non-scientists would also use these and similar behaviours as indicating the presence of private mental states. We then asked whether these inferences about consciousness would bias participants’ moral decision making, and whether some 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, ndefinedwe followed Piazza and Loughnan (2016) and desgcribed 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 such pair the two species were identical except for two differences. The experiment then consisted of descriptions of the 8 animal pairs, followed by qu

For each pair, participants were asked to describe in their own words the main difference between the two alien species. This allowed us

eewilks2021]. Lastly, participants were asked to use two sliders to indicate the extent to which they thought each species was conscious.

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 5 dimensions are based on a taxonomy of animal consciousness, described in Birch et al. (2020). The *feature-positive* species had more of the mental capacity of interest, or resembled humans more, compared with the *feature-negative* species. Each dimension was presented as two scientific findings, followed by their interpretation by the scientists, and was accompanied by cartoon figures of the experimental design. The findings and their interpretations were based on real animal studies. Alien species were given one-syllable gibberish names, fully counterbalanced across participants. Below we describe the 8 dimensions and their experimental operationzaliations, based on animal studies. The full descriptions as presented to our subjects are available at [github.com/matanmazor/dimensions\_of\_moral\_status](https://github.com/matanmazor/dimensions_of_moral_status).

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 between 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. 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 aliens 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.

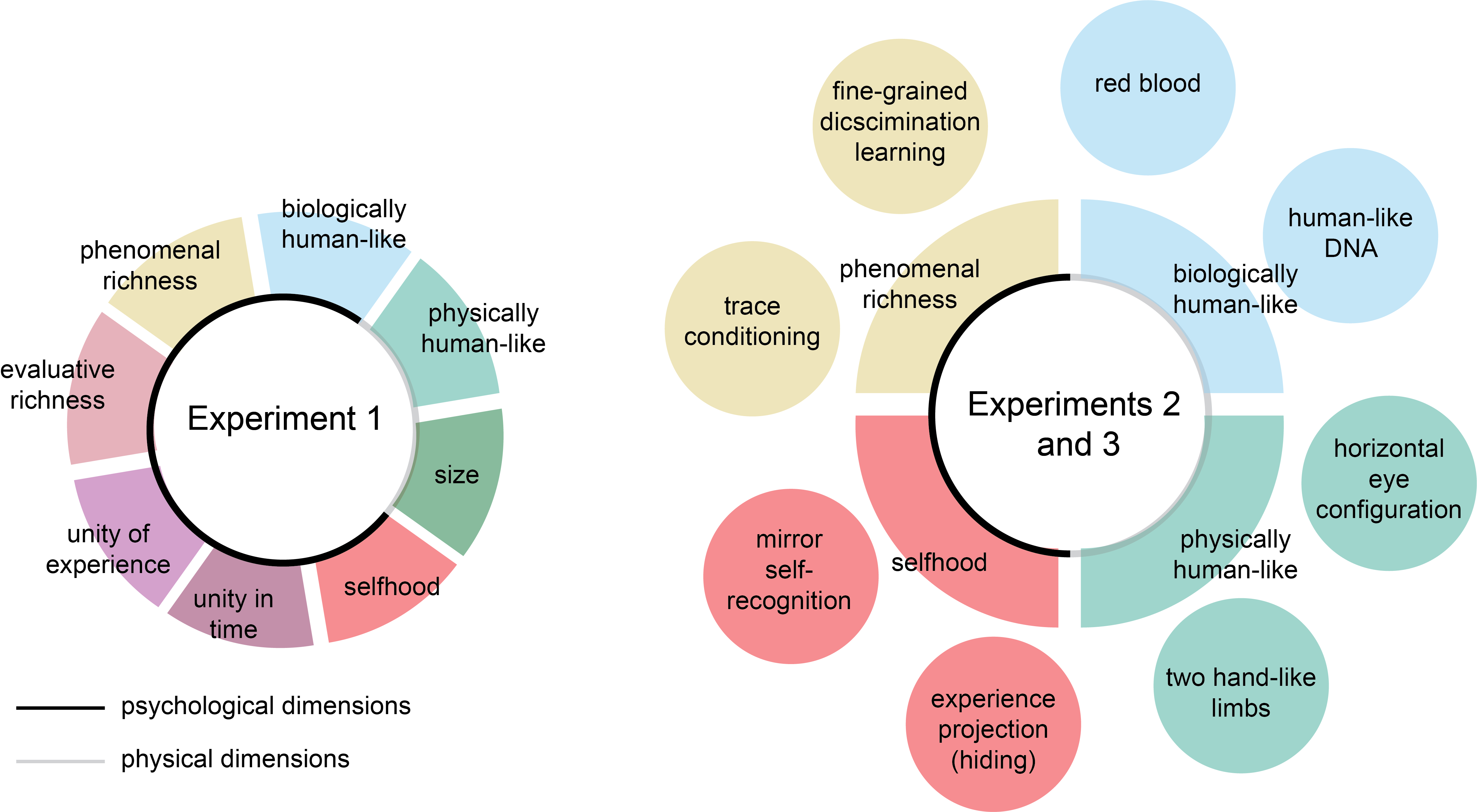


Figure 1: Manipulated dimensions. In Exp. 1, we manipulated eight dimensions: five psychological and five physical. In Exp. 2 and 3, we focused on a subset of four dimensions: phenomenal richness in vision, selhood, physical resemblance to humans, and biological resemblance to humans. Each dimension was broken down 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](https://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 (, 95% CI , , ; see Fig. 2). This difference in consciousness ratings was significant for the dimensions phenomenal richness (, 95% CI , , ), unity (, 95% CI , , ), temporality (, 95% CI , , ), and selfhood (, 95% CI , , ). Interestingly, perceived consciousness was also higher for aliens that are more biologically similar to humans (, 95% CI , , ). 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.

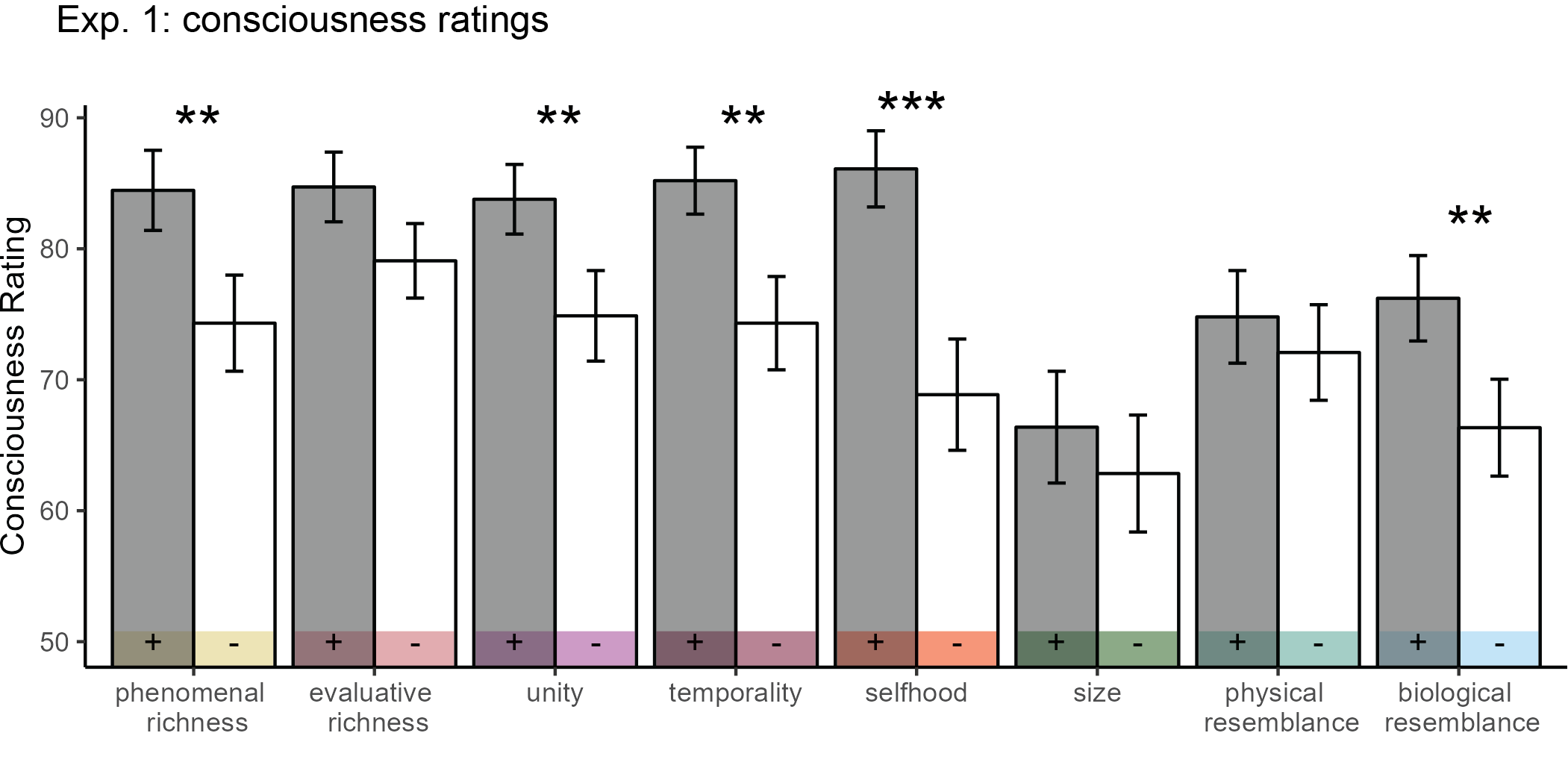


Figure 2: Mean consciousness ratings in Exp. 1. The feature-positive and feature-negative aliens are presented in grey and white, respectively. Error bars represent the standard error of the mean. Starts represent statistical significance after a Bonferroni correction for multiple comparisons. p<0.01: \*\*, 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 () was always set to 10, and the number of feature-negative aliens () 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.

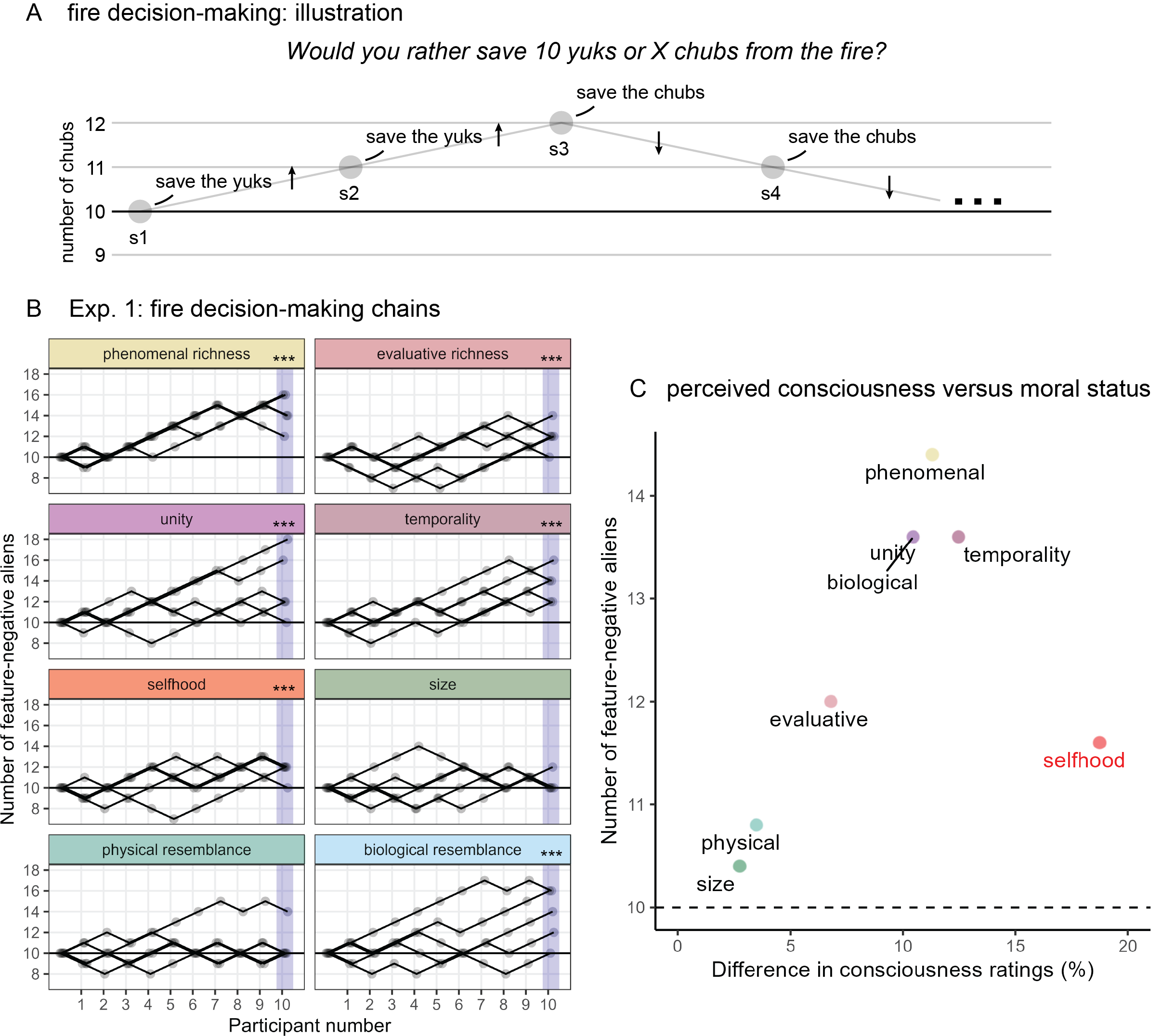


Figure 3: Exp. 1: results. A: the between-subject staircase procedure. Each circle represents one participant. The number of feature-negative aliens increased following decisions to save the feature-positive group, and decreased following decisions to save the feature-negative group. B: the number of 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 per individual dimensions.

We focused on the average value of at the end of the chain. This value 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 (=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 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 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. In the last section of our analysis, we asked whether beliefs about consciousness systematically covary with biases in moral decision making.

### The relation between moral judgments and consciousness ratings.

So far we have shown that in our data, 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 the 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, high values of were associated with a more pronounced difference between consciousness ratings for the feature-positive and feature-negative alien species (see Fig. 3C). One exception was selfhood, scoring highest in consciousness ratings (with a mean difference of 18.75 between feature-positive and feature-negative aliens), but having only moderate effects on moral status judgments (=11.6). Indeed, the effect on perceived consciousness correlated with only when excluding the selfhood dimension (, 95% CI , , ), but not when including it (, 95% CI , , ). In Experiment 2 and 3 we look closer at the two operationalizations of selfhood in our study, mirror self recognition and a tedency to hide after a shameful act, and ask whether this mismatch between consciousness ratings and moral worth judgments is common to both.

# Experiment 2

In Experiment 1, we found that short descriptions of scientific findings affected participants’ attribution 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). 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](https://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 (, 95% CI , , ). 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.

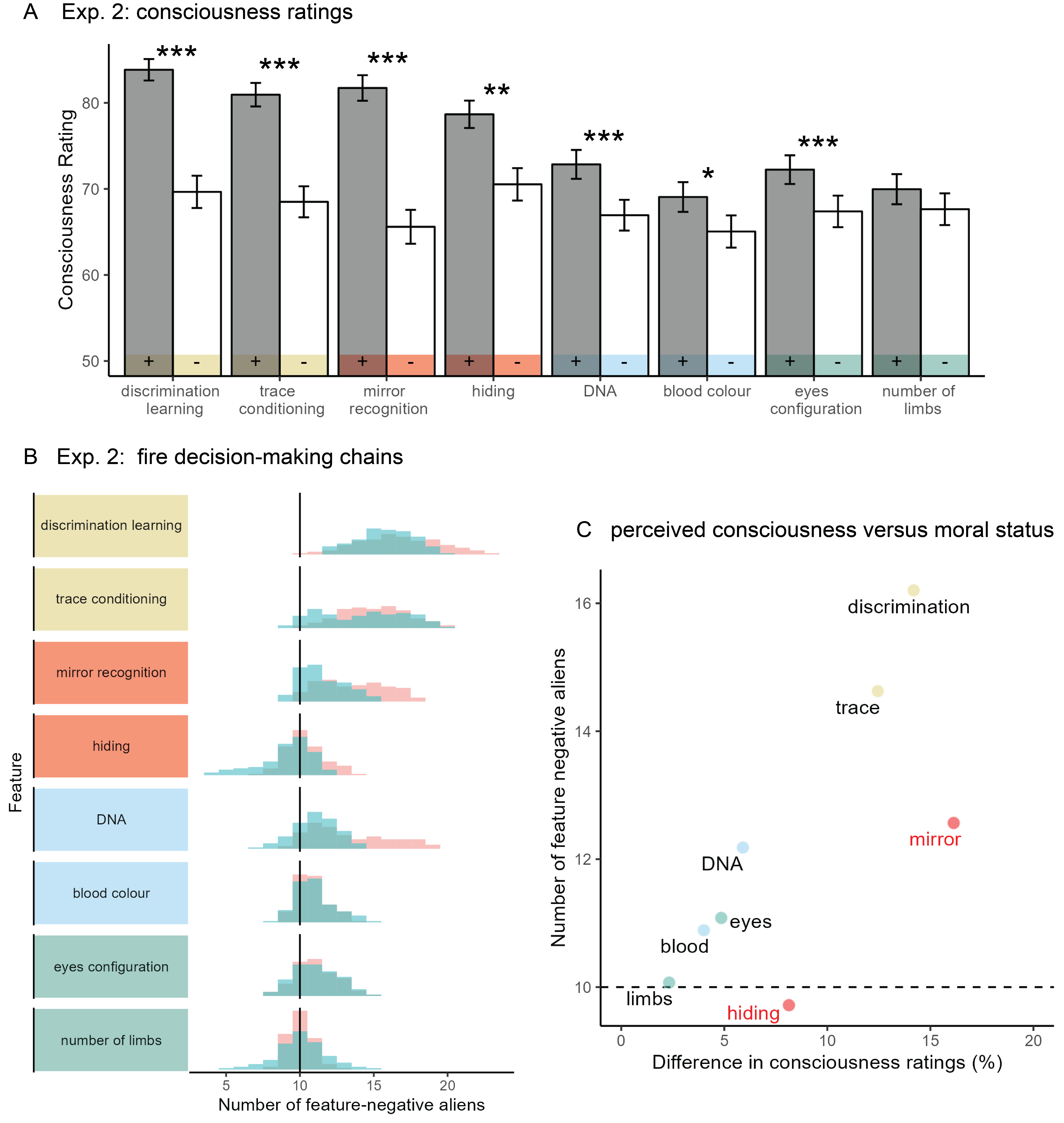


Figure 4: Exp. 2: results. A: mean consciousness rating per alien species. B: histograms of N- values across the two chains, excluding the first 10 steps. C: mean number of feature-negative aliens at the end of the chain, plotted against the difference in consciousness ratings per individual dimensions. Self-awareness items are marked in red.

### Moral judgments.

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 () was always set to 10, and the number of feature-negative aliens () 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 values as a burn-in period, and took the mean of the remaining values as our estimate for . Generally, we observed high levels of agreement between the two chains (see Fig. 4B). One exception was mirror recognition, with mean of 13.62 and 11.52 for chains number 1 and 2.

In line with the results from Experiment 1, was highest for discrimination learning (16.20) and trace conditioning (14.63), both operationalisations 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 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 ( 12.57, but see above comment 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 ( 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: , 95% CI ). Participants also attributed greater value to the lives of aliens whose DNA was composed of the same DNA bases as human DNA ( 12.18). Finally, eye configuration ( 11.08), the colour of the blood ( 10.89), and the number of limbs ( 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 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: , 95% CI , , ; correlation with selfhood-related items: , 95% CI , , ).

# Experiment 3

Experiment 3 was designed to replicate 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 (, , ). 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 (, 95% CI , , ).

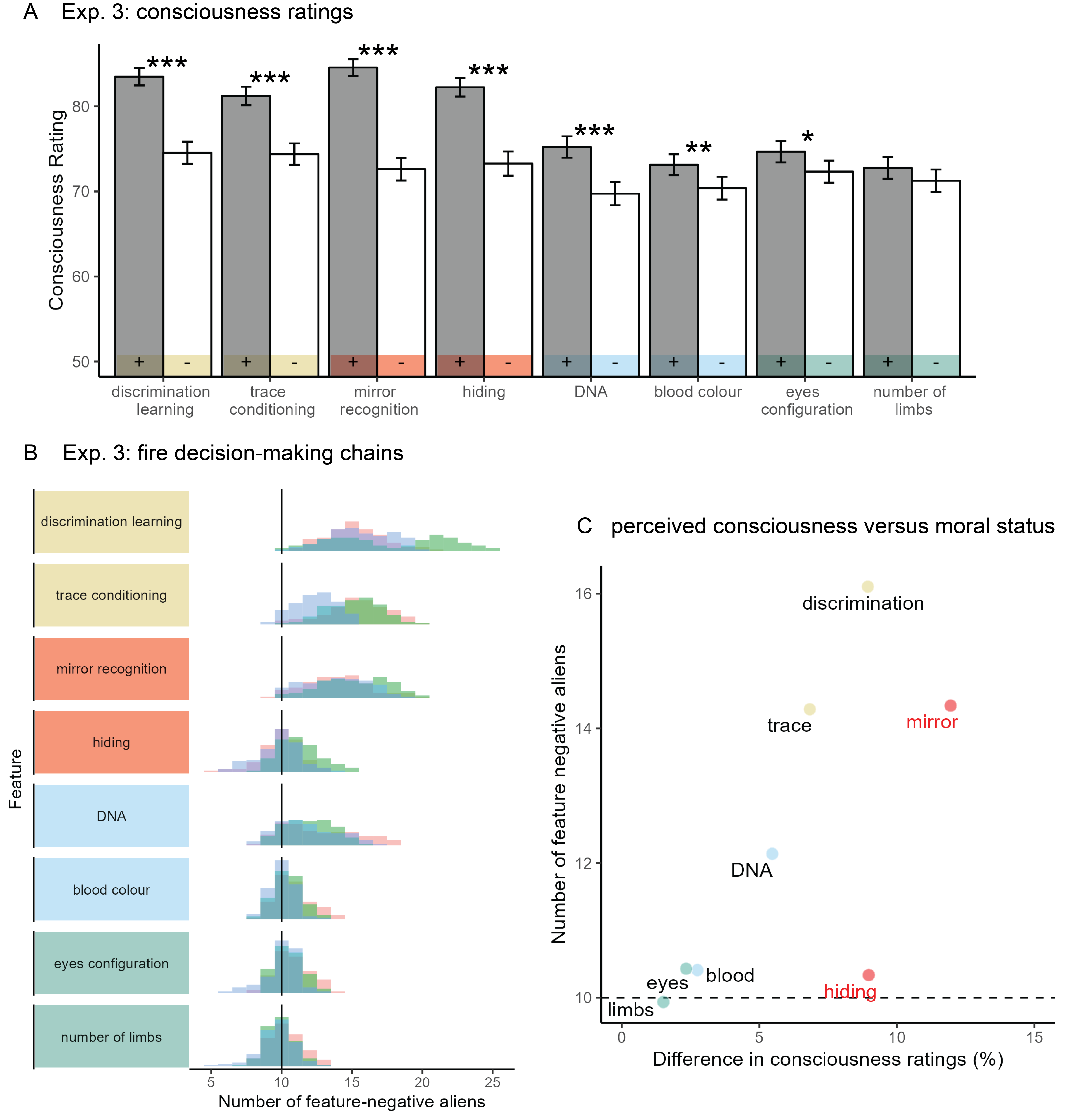


Figure 5: Exp. 3: results. A: mean consciousness rating per alien species. B: histograms of N- values across the two chains, excluding the first 30 steps. C: mean number of feature-negative aliens at the end of the chain, plotted against the difference in consciousness ratings per individual dimensions. 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 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, values were significantly higher than 10 in 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 across features (see Fig. 4). Similar to Experiment 1 and 2, this linear alignment held weaker for selfhood-related features: mirror self-recognition and hiding (correlation without selfhood-related items: , 95% CI , , ; correlation with selfhood-related items: , 95% CI , , ). In a pre-registered test, we extracted the proportion of explained variance for the linear alignment between consciousness ratings and 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 ( = .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 (, ), but this was not true for participants who chose to save the feature-negative group (, ). The difference between the two groups was in itself significant (, 95% CI , , ).

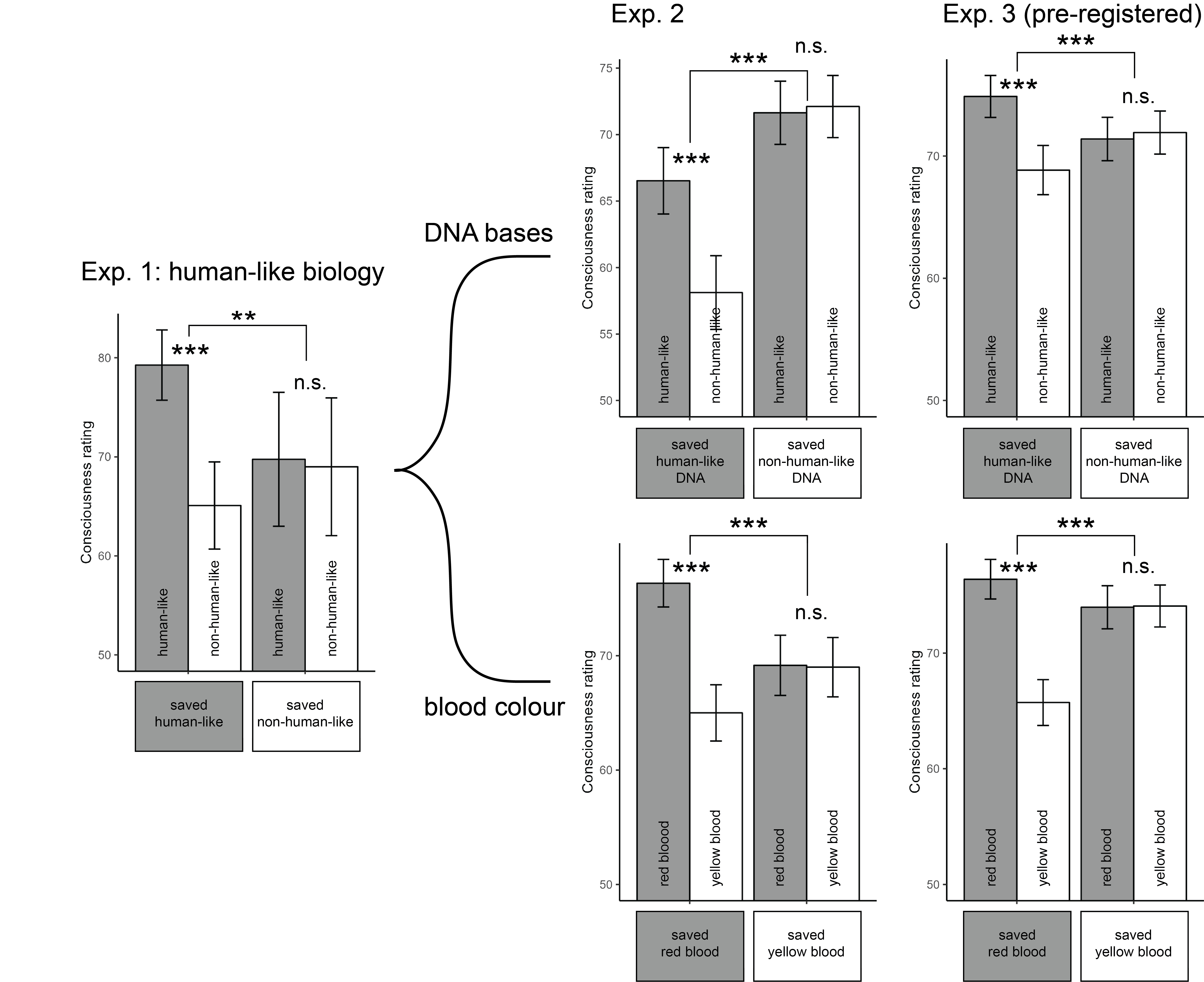


Figure 6: The value of human-like biology. Plotting consciousness ratings for the feature-positive (grey) and feature-negative (white) groups as a function of participants’ decisions in the fire dilemma. A difference in consciousness ratings between human-like and non-human-like aliens appears only in 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: , ; Exp. 3: , ), whereas participants who chose to save the other group showed no difference (Exp. 2: , ; Exp. 3: , ), with a significant interaction pattern (Exp. 2: , ; Exp. 3: , ). Similarly, participants who chose to save red-blooded aliens thought they were more likely to be conscious than yellow-blooded aliens (Exp. 2: , ; Exp. 3: , ), this was not the case for participants who saved the yellow-blooded aliens (Exp. 2: , ; Exp. 3: , ), and the interaction was again significant (Exp. 2: , ; Exp. 3: , ). 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 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 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 (Gray, Gray, & Wegner, 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, Caviola, Kahane, & Bloom, 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 low-level visual 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), also in the context of ongoing discussions about the possibility of AI consciousness [Danaher (2020); Birch (2024); ].

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 lower-level aspects of consciousness such as visual awareness and working memory are much more influential for moral decision-making than beliefs about selfhood, more in line with the utilitarian views of Mill (1861) and Bentham (1789), and more recently with the view of Shepherd (2018).

Finally, our findings speak to the ethical significance of consciousness science findings (Mazor et al., n.d.). 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 in hypothetical questions of life and death. As scientists continue to unravel the mysteries of minds, brains and machines, they inevitably inform and change whose interests we value as a society.

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., 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>

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>

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>

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>

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. (n.d.). The scientific study of consciousness cannot and should not be morally neutral. *Perspectives on Psychological Science*, *0*(0), 17456916221110222. <https://doi.org/10.1177/17456916221110222>

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>

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>

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>