A coherence-based approach to moral trade-offs

Aidan Runagall-McNaull (arunagallmcn@student.unimelb.edu.au)
Yoshihisa Kashima (ykashima@unimelb.edu.au)
Simon Laham (slaham@unimelb.edu.au)
School of Psychological Sciences, University of Melbourne, Parkville, Victoria 3010, Australia

Abstract

The present research evaluates a coherence-based network approach to moral trade-off judgements. Under this view, judgement is an outcome of achieving coherence between a network of causally interacting beliefs. Consistent with this, despite similar initial views, participants re-evaluated their beliefs and attitudes in support of their judgement, driving polarisation between individuals reporting competing judgements. Different properties of the dynamic network structure determined metacognitive properties of judgement such as confidence and perceived task difficulty. Whilst the judgement formation process involves revising beliefs and values to achieve a coherent arrangement, the nature of the judgement reached depends on the aggregate weight of these beliefs once the revision process is completed.

Keywords: Moral judgement; moral trade-offs; coherence-based reasoning; belief systems; psychological networks.

Introduction

Our values determine the judgements we make, not the other way around. At least, this has been a widespread assumption over the history of the study of morality. For classical Greek thinkers, good judgements are the product of virtues, stable moral dispositions cultivated over a lifetime (Kamtekar, 2013). Hume and Kant both pointed to different, but similarly stable and enduring aspects of our nature as the bases of moral judgement (Guyer, 2012). Current extant psychological models of moral judgement also share features of these traditions. Agents are often assumed to approach moral problems with an established set of values, rules and expectations that guide judgement on the right course of action. Despite significant variation in the how this judgement is obtained (c.f. Bago & De Neys, 2019; Cushman et al., 2010; Haidt, 2001), determining values are not expected to change over the short period of time taken to form a judgment. Nevertheless, the predictions of these models are not consistently met (McHugh et al., 2022).

In contrast, coherence-based approaches to judgement formation such as constraint satisfaction models (e.g. Glöckner, 2008; Simon & Holyoak, 2002) do not rest on the assumption of unidirectionality. Under this view, judgements are formed by achieving coherence within a network arrangement imposed by the structure of the decision problem. Within these networks, beliefs, attitudes, emotions, or goals (henceforth referred to as 'beliefs' for brevity) are represented as nodes, connected by excitatory or inhibitory links. The nature of these links may depend on logical or causal dependencies between beliefs, as well as whether they favour the same or different judgements (Glöckner & Betsch, 2008). Cognitive consistency processes then proceed to iteratively re-evaluate initial beliefs and values to bring the network into a coherent arrangement that favours one of the available judgement options. For instance, fearing snakes may be causally linked with the belief that snakes are dangerous. In contrast, judging that snakes ought to be conserved may involve aligning the belief that snakes are important for the environment with the view that snakes are beautiful, as they both represent positive evaluations of the judgement object. Importantly, decision making under this view does not simply involve selecting a judgement option. Rather, it involves re-evaluating the constellation of values, beliefs and emotions that relate to the decision problem (Holyoak & Powell, 2016). These revised beliefs and values may subsequently impact other judgements (Horne et al., 2015).

Coherence-based accounts have been successfully applied to a wide range of areas including perception (Mcclelland et al., 2014), analogy (Holyoak & Thagard, 1989), preference ordering (Simon et al., 2004) and legal decision making (Holyoak & Simon, 1999; Simon, 2004). Though several authors have suggested a role for coherence as a mechanism in moral judgement (see Clark et al., 2015; Holyoak & Powell, 2016; Thagard, 1998) this has not been examined empirically. One potential reason is that many of the predictions of this model contradict foundational assumptions in moral philosophy and psychology as mentioned above (Holyoak & Powell, 2016). Additionally, moral judgement is commonly considered to be different in kind to judgements made in perceptual tasks or in legal decision making. In the aforementioned contexts, agents attempt to ascertain a state of the world using ambiguous or uncertain cues. As such, arriving at a judgement on the basis of conflicting information can reasonably be used to update confidence in cues contrary to that judgement, or to reject them entirely. The same may not be true of the determinants of moral judgement such as values, which may not correspond to objective facts about the world in the same way as perceptual cues[1]. At least according to a folk conception of

[1] Though the extent to which this is the case may depend on folk metaethical commitments (Beebe, 2015; Sarkissian et al., 2011)

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

morality, judging in favour of one moral good over another should not count as evidence that our valuing the latter object is wrong or mistaken. When choosing a charity to volunteer for, selecting one option is not typically understood as reducing the worthiness of all other causes, no more than deciding on a cup of tea today ought to reduce your background preference for your usual coffee. Nevertheless, there is some evidence that something akin to this may be occurring, in accordance with predictions of the coherence model (Simon et al., 2015).

Overview of studies

In two studies, we test the predictions of a coherence-based model of moral judgement utilising a moral dilemma concerning the issue of *climate justice*. It is well established that rapidly reducing carbon emissions is vital for individual, social and ecological wellbeing long-term. However, moving away from a carbon intensive economy may be an expensive process. This is most able to be afforded by wealthier, developed nations, who also tend to bear the most responsibility for the current threat of climate change (Wei et al., 2012). Meanwhile, coal or other fossil fuel reserves may represent valuable sources of cheap energy for developing nations, who may be experiencing human welfare issues more immediately threatening than the prospect of climate change. Forming a judgement in these cases involves trading off environmental and social justice concerns.

This dilemma is intended to be complex and realistic, avoiding issues of ecological validity raised with the simple 'trolley-like' moral dilemmas used by previous work (Hofmann et al., 2014; Kahane, 2015; Levitt & List, 2007). Participants take the role of the leader of a developing nation, tasked with approving or denying a proposed coal mine. Before forming a judgement, participants are presented with arguments from two 'advisors', each comprising a range of factual and evaluative claims. This decision task comprised the 'main test' in a three-phase experimental structure adapted from Holyoak and Simon (1999). Bookending the main test was a pre and post-test. These are identical to each other and are used to measure participant endorsement of the same claims present in each advisor's argument, but presented individually. For instance, one advisor claimed that the environment should be the primary concern for your government, whilst the other claimed that this should be human welfare. Similarly, advisors took opposing positions on the number of jobs mining creates, one claiming that the mine will produce many jobs, the other claimed it will create veery few jobs. Participants rated their agreement with each of these 14 claims in the pre and post-tests. Using this approach allows estimation of decision-relevant beliefs, before and then after they are traded off against each other.

Coherence and Consistency

At the pre and post-test stages, two subject-level measures were calculated to quantify the structural relationship between participant beliefs (measured by agreement ratings towards claims made by advisors) and between these beliefs and their reported judgement. Under the proposed model, beliefs favouring a given conclusion become aligned, ultimately strengthening those in favour of a judgement outcome and weakening those contrary. Other coherence-based accounts treat this as a single process (Glöckner, 2008; Simon & Holyoak, 2002), and tend to use the terms 'coherence' and 'consistency' interchangeably. However, here we distinguish between two related but distinct properties of collections of beliefs.

Coherence concerns the alignment between individual beliefs. Two beliefs have a coherent arrangement if they are both endorsed to a similar level and favour the same judgement option.

Consistency concerns the relationship between beliefs and a certain judgement. Beliefs held by an agent are consistent with a given judgement if they favour that judgement option.

There are several theoretical reasons for investigating coherence and consistency independently. For one, it is possible for beliefs to be on average highly consistent with a judgement, but heterogeneously so, resulting in low coherence. Likewise, a collection of beliefs with uniformly low endorsement would be highly coherent but exhibit low consistency if they favoured the chosen judgement option. Beliefs or attitudes relevant to the decision problem may be constrained by other, unrelated (and unmeasured) beliefs if they share some kind of causal connection, such as being part of a network comprising an attitude (Dalege et al., 2016) or ideology (Brandt & Sleegers, 2021). This in turn may influence the behaviour of the decision network as a whole, and may cause coherence and consistency to diverge.

Additionally, considering coherence and consistency independently may be important to understand metacognitive properties of a judgement. For instance, the degree of initial conflict within the network structure dictated by the decision problem in part[2] determines the extent of revision required to achieve coherence (Glöckner, 2008). Greater initial network incoherence, and hence more processing effort required in comparing, weighing, and revising beliefs may be reasonably expected to result in perceptions of higher task difficulty. Conversely, the coherence shifts in judgement antecedents occur mostly prior to, and play a causal role in judgement formation (Simon, 2004). As such, the *final* state (i.e. in the post-test) of relevant beliefs provide agent grounds for their judgement, whilst the extent to which this state is consistent with one's judgement may reflect judgement confidence. Taking confidence into account is important for understanding behavioural outcomes of a judgement or the likelihood of revision (Yeung & Summerfield, 2012). Task

[2] There may be individual differences in the degree to which conflict is resolved (Dalege & van der Does, 2022)

difficulty may be important for estimating cognitive resource allocation, experiences of mental effort and aversiveness towards the task (Kurzban et al., 2013).

However, it is unclear precisely what feature of the final decision network drives judgement. On one hand, having a collection of beliefs that are highly consistent with a certain judgement option seems a clear predictor of judgement (Holyoak & Simon, 1999; Simon et al., 2015). However, evidence for the positive effect of argument coherence on persuasiveness (e.g. Huntsinger, 2013) suggests a role for coherence as well. That is, given a fixed level of mean endorsement, it is reasonable to suspect that a more coherent argument would be more persuasive.

Assuming that a judgement is made between a and b (a = 1, b = -1) and there are I beliefs favouring a (i,j = 1 to I) and K beliefs favouring b (k = 1 to K), let a(i) and b(k) represent the endorsement for belief i favouring a and the endorsement for belief k favouring b ($-1 \leq ai$ and $bk \leq +1$). We computed coherence and consistency as follows:

$Coherence\ for\ a = \sum_{i=1}^{I}\sum_{j=1}^{I} 1 - |a(i)-a(j)|$, where $i \neq j$ (similarly for b).

$Consistency\ for\ a = \sum_{i=1}^{I} aa(i)$ (similarly for b).

Overall coherence and consistency at each timepoint were calculated by averaging across arguments a and b.

Predictions

If coherence-based reasoning is driving moral judgement formation, the following trends should be observed (adapted in part from Holyoak & Powell, 2016).

1. Revision of beliefs to better support judgement
2. Sharply divided decisions are accompanied by high confidence for each individual decision maker.
3. Revision of beliefs largely takes place prior to commitment to a decision.
4. Revision can be triggered by any task that encourages attention and comprehension (even if no decision is required).

Additionally, the following hypotheses will be tested to examine the influence of coherence and consistency on judgement content and metacognitive properties of judgement such as difficulty and confidence.

5. Coherence and consistency will increase independent of each other over the course of judgement formation.
6. Pre-test coherence will be the strongest predictor of task difficulty
7. Post-test consistency will be the strongest predictor of judgement confidence
8. Post-test consistency and coherence will be significant predictors of judgement

Study 1

Participants

Two hundred and fifty-one participants were recruited from Prolific (M_{Age}=38.6, SD= 13.0; 143 female, 104 male, 3 non-binary and one chose not to disclose gender). All participants were recruited from the UK and were paid £4.75 for a 30 minute survey.

Procedure

Participants volunteered to participate in a study called "Moral Judgement and Personality", administered online via Qualtrics. Participants completed the pre, main and post-tests. Attention checks, each comprising five simple true/false arithmetic problems, were included after the pre and post-tests. The main test included the mine approval vignette, with the advisor arguments presented in a counterbalanced order. Participants reported their judgement (approve vs. disapprove), confidence (6-point unipolar scale), and task difficulty (7-point bipolar scale). Pre and post-tests contained identical items, including agreement ratings towards each of the 14 claims made by advisors, presented in a random order. Seven of these claims comprise the argument against the mine, and seven comprising the argument in favour. The pre-test included a further 16 'dummy' propositions included to ensure that the component propositions bore no plausible relationship to each other.

Results

Distribution of judgements was approximately balanced (132 deny, 118 approve), with the decision task being mostly rated between 'slightly' and 'very' difficult. Confidence ratings were roughly normally distributed, with participants reporting that they were 'somewhat' confident in their decision on average.

Following Holyoak and Simon (1999), an 'm-score' was calculated for each participant at both pre and post-test timepoints by taking an average of all agreement ratings, reverse-scoring those items that did not favour the mine. As such, a positive (negative) m-score indicates net approval (disapproval) of the mine. Pre and post-test m-scores for mine approvers and deniers are plotted in figure 1. A mixed model ANOVA with m-scores as the DV and both 'test' (pre vs post-test) and 'judgement' (i.e. approve vs. deny) as predictors. Results showed significant effects of judgement, $F(1, 248)=91.31$, $p<.001$; test, $F(1, 248)=5.44$, $p=.021$; and the judgement*test interaction term, $F(1, 248)=111.7$, $p<.001$. This suggests that mine approvers and deniers had different beliefs towards the mine overall, and that these beliefs changed between pre and post-tests in a manner dependent on the judgement option selected.

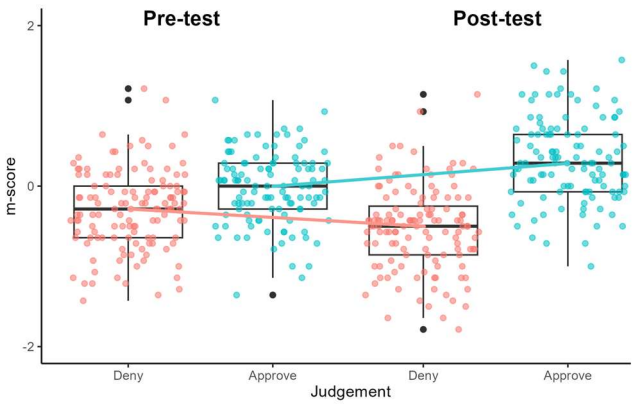

Figure 1: Changes in m-score between pre and post-test split by judgement.

We estimated consistency and coherence at both pre and post-test. Using lme4 (version 1.1.35.1), we performed repeated measures analyses of covariance to examine changes in coherence and consistency for both arguments over judgement formation (see tables 1, 2, supplementary materials). The 'argument' variable indicated whether it was 'aligned' or 'unaligned' with each participant's judgement (i.e., the pro-mine argument is aligned with an approve judgement). Both coherence (β = .033, SE= .0055, p< .001) and consistency (β = .05, SE= .0061, p< .001) increased independently of each other from pre to post-test.

To predict Judgement, we used m-scores rather than consistency. The two measures are closely related to each other, but whilst consistency as we have calculated it quantifies the extent to which the aggregate weight of pre or post-test agreement ratings are aligned with the judgement reported, m-scores quantify how pro or anti-mine the same agreement ratings are in general. Additionally, a coherence-difference measure was calculated to quantify the extent to which one argument was more coherent than the other by subtracting the coherence of the anti-mine argument from the coherence of the pro-mine argument. As such, for both coherence-difference and m-score measures, a negative value indicated that agreement ratings towards propositions comprising the anti-mine argument were respectively more coherent with each other or stronger on average. The dependent variable was computed by multiplying judgement (coded as 1/-1, approve/deny) with confidence. A general linear model (GLM) was run with coherence difference and m-score at pre and post-test as predictors (F(4, 246)=72.85, p<.001). There was a strong positive effect of post-test m-score, β = 1.99, SE= .18, p< .001. There was also a weak negative effect of pre-test m-score; however, when post-test m-score was removed from the model, the effect of pre-test m-score became positive. As pre and post-test m-scores are strongly correlated (r(249)=.74, p<.001), the negative effect of pre-test m-scores is likely to be a suppression effect.

A multiple linear regression (MLR) model was run to examine the effect of coherence and consistency on judgement confidence with pre and post-test coherence and consistency as predictors and judgement difficulty included as a covariate (F(5,245)=46.35, p<.001). Post-test consistency was the only significant predictor of judgement confidence. Similarly, a MLR was run with judgement difficulty as DV, pre/post-test mean coherence and consistency as predictors and judgement confidence as a covariate (F(5,245)=39.1, p<.001). Results showed a significant, negative effect of pre-test coherence, as well as an effect of post-test consistency that was marginally non-significant. Results for both regression analyses included in table 1.

Study 2

Study 1 showed that moral judgement is associated with a revision of beliefs as well as an increase in coherence and consistency. Study 2 examined whether these changes were causally involved in judgement formation. Study 2 additionally tested predictions 3 and 4.

Participants

Six hundred undergraduate students enrolled at the University of Melbourne participated in return for course credit (M_{Age}=19.28, SD=2.63; 461 female, 128 male, six non-binary and five participants who chose not to disclose their gender).

Table 1: Regression results for judgement confidence and difficulty

Study	Timepoint	Predictor	Confidence		Difficulty	
			β	Standard error	β	Standard error
Study 1	Pre-test	coherence	-.28	0.20	-.91***	0.25
		consistency	0.07	0.20	0.17	0.26
	Post-test	coherence	-0.37	0.21	0.09	0.27
		consistency	0.68***	0.19	-0.47	0.25
Study 2	Pre-test	coherence	-0.14	0.16	-0.63**	0.22
		consistency	-0.33	0.17	-0.48*	0.23
	Post-test	coherence	-0.25	0.18	0.05	0.25
		consistency	0.35*	0.15	-0.21	0.21

*Note.*** p<.001, ** p< 0.01, * p< 0.05

Procedure

Participants volunteered to participate in a study called "Moral Judgement and Personality", administered online via Qualtrics. Participants were presented with one of four different experimental conditions (N=150 per condition).

'Replication' was identical to study 1. 'No judgement' was identical to Replication, except that participants were not asked to report their judgement and rate confidence and difficulty until after the post-test. 'No trade-off' used an alternative vignette in the main test, semantically similar to the original stimulus. This detailed a developing nation, including details of unutilised coal deposits as well as environmental and human welfare concerns. Importantly, these concerns were not presented in the context of a trade-off. After reading the vignette, participants were presented with post-test items, and then the original mine approval vignette and reported their judgement, confidence, and difficulty as in the other conditions. 'Unrelated' was identical to the 'no trade-off' condition but used an unrelated narrative about the history of pasta in the main test.

Results

Two hundred and seventy-seven chose to deny the mine; 323 chose approval. On average, participants were 'somewhat confident' about this judgement, and found the task 'slightly difficult' on average.

We performed a mixed model ANOVA with m-scores as the DV and test (that is, pre or post-test) judgement and condition as predictors. Main effects for all predictors except condition were significant (see table 3, supplementary materials). Due to a significant condition*judgement*test interaction term ($F(3, 592)=9.8$, $p<.001$), ANOVAs were performed for each condition separately. Analyses for both 'replication' and 'no-judgement' revealed significant effects of judgement, test and the judgement*test interaction. The 'no-trade' condition had significant judgement and test main effects, but no significant interaction. M-scores in this condition changed between pre and post-tests, but this change did not vary depending on the judgement reported. Analyses for the 'unrelated' condition showed no significant effects. Results for these analyses available in table 4, supplementary materials. Figure 2 shows m-score changes in the four conditions.

As the polarisation effect observed in study 1 was not replicated in the 'no trade-off' and 'unrelated' conditions, these were excluded from subsequent analyses. The 'replication' and 'no judgement' conditions were combined in order to achieve a similar sample size to study 1.

As in study 1, we estimated coherence and consistency in both pre and post-tests. We then fitted multi-level models to examine changes in coherence and consistency over judgement formation (see table 5, 6, supplementary materials). Both coherence ($\beta = .035$, $SE= .0044$, $p< .001$) and consistency ($\beta = .032$, $SE= .006$, $p< .001$) increased independently of each other from pre to post-test.

Analyses to predict judgement were performed as in Study 1, using judgement as the DV and pre/post-test m-score and coherence difference as predictors ($F(4, 295)= 42.56$, $p<.001$). There was a significant, positive effect of post-test m-score ($\beta =1.34$, $SE= .2$, $p< .001$), whilst pre ($\beta =-.45$, $SE= .17$, $p= .008$) and post-test coherence ($\beta =-.33$, $SE= .15$, $p= .03$) difference showed significant but weaker contributions in opposing directions.

Predictors of judgement confidence and difficulty were also analysed in an identical manner to study 1. Post-test consistency showed a significant, positive effect on judgement confidence, whilst both pre-test consistency and pre-test mean coherence were significant, negative predictors of judgement difficulty. Results for these analyses are detailed in table 1.

Discussion

We found significant shifts in participant beliefs between the pre and post-test in study 1, consistent with model predictions. These beliefs shifted in the direction of their judgement their judgement, as evidenced by a significant judgement*test interaction term in the analysis of m-scores in both studies. This has the effect of polarising individuals reaching different conclusions, despite similar initial commitments (see figures 1, 2). Whilst similar patterns have been observed in non-moral decision tasks (Holyoak & Simon, 1999; Simon et al., 2001), this research is the first to demonstrate these results in a moral judgement context.

Our results suggest that these shifts in belief could be triggered by simply reflecting on the dilemma, without having to register a judgement (study 2, 'no judgement'). As such, the results of study 1 are unlikely to driven by post-hoc justification. However, reflecting on the semantic content in the absence of a trade-off did not produce similar results, in violation of prediction 4 above. Whilst Holyoak and Powell (2016) suggest that any task requiring attention and comprehension may be sufficient to trigger re-evaluation, it appears that a trade-off structure may be necessary. Likewise, being exposed to the same items twice (i.e. in the pre and post-test, study 2 'unrelated') did not result in belief revision. Overall, this pattern of findings suggests that shifts in agent beliefs are causally involved in judgement formation in moral trade-offs.

The significant influence of post-test consistency on judgement confidence mirrored the role of m-scores on judgement content. Whilst a net-positive view on the mine tended to be associated with an approve judgement, having a net-positive view in *addition* to an approve judgement (i.e. having

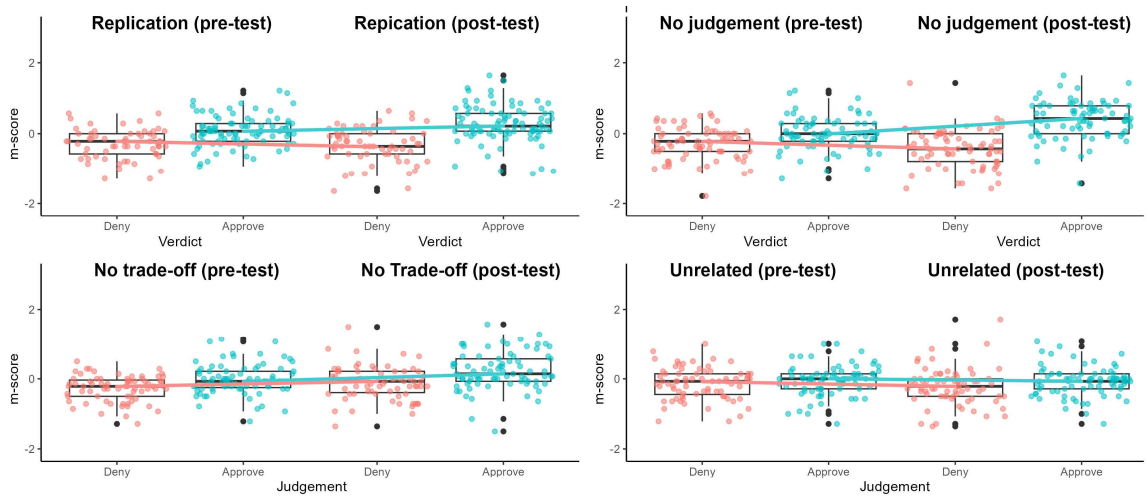

Figure 2: Changes in m-score split by judgement across four conditions.

beliefs consistent with your judgement) was associated with higher confidence ratings in both studies. Post-test measures that were best predictors of judgement, consistent with judgement being based on the final network arrangement after the revision and restructuring process had been completed.

The extent to which beliefs were revised was associated with perceptions of task difficulty. In both studies, pre-test coherence negatively predicted difficulty ratings. Participants who started the study with beliefs that were more coherent with each other and as such required less revision, found the decision task easier. However, note that this is a *task specific* measure of coherence. In different decision problems, the beliefs that jointly favoured mine approval here may favour competing conclusions. As such, the same beliefs measured in the present studies may be more or less coherent with respect to a different decision task.

Whilst most coherence-based accounts of judgement treat maximisation of coherence and consistency as a single process (Glöckner, 2008; Simon & Holyoak, 2002), the fact that they increase independently of one another and predict separate judgement outcomes make a good case for their being evaluated separately. Overall, the effect of pre-test coherence on task difficulty suggests that moral judgement formation is a process of coherence maximisation. However, the content of the judgement reached depends on the consistency of beliefs and attitudes after the coherence maximisation process has concluded.

The results obtained here may help explain why current models of moral judgement often fail to make accurate predictions. These approaches tend to measure participant values and then use these measurements to predict judgement (e.g. Bago & De Neys, 2019). This only takes into account initial evaluative tendencies which, as we have shown, are revised over the course of judgement formation. The *final* arrangement is a better predictor of judgement and may have a more causal role in judgement formation than initial belief-states. Failing to take into account the dynamic nature of judgement antecedents also neglects the causal role of coherence maximisation in judgement formation.

These findings also have significance for the study of political polarisation. Our results suggest a potential role for decision contexts in driving polarisation, which may have been neglected thus far. Moreover, they suggest that polarisation can occur even in the presence of diverse information, in contrast to echo-chamber style explanations of this phenomenon which foreground information uniformity as a key factor in polarisation (Arguedas et al., 2021). A key open question in this interpretation is the stability of the shifts observed here. Whilst there is some evidence that these revised beliefs can continue to influence other judgements for a period of several hours (Horne et al., 2015), further research is necessary to understand the conditions under which this can occur. In particular, it will be important to bridge the gap between coherence maximisation within decision contexts and work on long-term belief change driven by dissonance reduction (Dalege & van der Does, 2022).

Conclusion

Our findings provide good evidence for the coherence model of moral judgement. Moral judgement appears to be a process of re-evaluating initial beliefs to achieve coherence within a network structure dictated by the decision problem. Judgement content is determined by the aggregate weight of beliefs after coherence maximisation has completed. This process also drove polarisation between participants reaching competing judgement, whilst separate aspects of the initial and final belief network determine metacognitive properties of judgement. Our results underscore the importance of accounting for a range of potential determinants of judgement, as well as dynamic relationships between them over the course of judgement formation.

References

Arguedas, A. R., Robertson, C. T., Fletcher, R., & Nielsen, R. K. (2021). Echo chambers, filter bubbles, and polarisation: A literature review. *Reuters Institute for the Study of Journalism*, 1–42.

Bago, B., & De Neys, W. (2019). The intuitive greater good: Testing the corrective dual process model of moral cognition. *Journal of Experimental Psychology: General*, *148*(10), 1782–1801.

Beebe, J. (2015). The Empirical Study of Folk Metaethics. *Etyka*, *50*, 11–28.

Brandt, M. J., & Sleegers, W. W. A. (2021). Evaluating Belief System Networks as a Theory of Political Belief System Dynamics. *Personality and Social Psychology Review*, *25*(2), 159–185.

Clark, C. J., Chen, E. E., & Ditto, P. H. (2015). Moral coherence processes: Constructing culpability and consequences. *Current Opinion in Psychology*, *6*, 123–128. https://doi.org/10.1016/j.copsyc.2015.07.016

Cushman, F., Young, L., & Greene, J. D. (2010). Multi-System Moral Psychology. *The Moral Psychology Handbook*, 48–71.

Dalege, J., Borsboom, D., Van Harreveld, F., Van den Berg, H., Conner, M., & Van der Maas, H. L. J. (2016). Toward a formalized account of attitudes: The Causal Attitude Network (CAN) Model. *Psychological Review*, *123*(1), 2–22.

Dalege, J., & van der Does, T. (2022). Using a cognitive network model of moral and social beliefs to explain belief change. *Science Advances*, *8*(33), eabm0137.

Glöckner, A. (2008). How Evolution Outwits Bounded Rationality. The Efficient Interaction of Automatic and Deliberate Processes in Decision Making and Implications for Institutions. *Preprints of the Max Planck Institute for Research on Collective Goods*, *2008/2*.

Glöckner, A., & Betsch, T. (2008). Modeling option and strategy choices with connectionist networks : towards an integrative model of automatic and deliberate decision making Modeling Option and Strategy Choices with Connectionist Networks : Towards an Integrative Model of Automatic and De. *Preprints of the Max Planck Institute for Research on Collective Goods*, *2008/2*.

Guyer, P. (2012). Passion for Reason: Hume, Kant and the Motivation for Morality. *Proceedings and Addresses of the American Philosophical Association*, *86*(2), 1–25.

Haidt, J. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgement. *Psychological Review*, *108*(4), 814–834.

Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science*, *345*(6202), 1340–1343.

Holyoak, K. J., & Powell, D. (2016). Deontological coherence: A framework for commonsense moral reasoning. *Psychological Bulletin*, *142*(11), 1179–

Holyoak, K. J., & Simon, D. (1999). Bidirectional Reasoning in Decision Making by Constraint Satisfaction.pdf. *Journal of Experimental Psychology: General*, *128*(1), 3–31.

Holyoak, K. J., & Thagard, P. (1989). Analogical Mapping by Constraint Satisfaction. *Cognitive Science*, *13*(3), 295–355.

Horne, Z., Powell, D., & Hummel, J. (2015). A Single Counterexample Leads to Moral Belief Revision. *Cognitive Science*, *39*(8), 1950–1964.

Huntsinger, J. R. (2013). Incidental Experiences of Affective Coherence and Incoherence Influence Persuasion. *Personality and Social Psychology Bulletin*, *39*(6), 792–802.

Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social Neuroscience*, *10*(5), 551–560.

Kamtekar, R. (2013). Ancient Virtue Ethics: An overview with an emphasis on practical wisdom. In *The Cambridge companion to virtue ethics* (pp. 29–48).

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences*, *36*(6), 661–679.

Levitt, S. D., & List, J. A. (2007). About the Real World ? *Journal of Economic Perspectives*, *21*(2), 153–174.

Mcclelland, J. L., Mirman, D., Bolger, D. J., & Khaitan, P. (2014). Interactive activation and mutual constraint satisfaction in perception and cognition. *Cognitive Science*, *38*(6), 1139–1189.

McHugh, C., McGann, M., Igou, E. R., & Kinsella, E. L. (2022). Moral Judgment as Categorization (MJAC). *Perspectives on Psychological Science*, *17*(1), 131–152.

Russo, J. E., Carlson, K., Meloy, M. G., & Yong, K. (2008). The Goal of Consistency as a Cause of Information Distortion. *Journal of Experimental Psychology: General*, *137*(3), 456–470.

Sarkissian, H., Park, J., Tien, D., Wright, J. C., & Knobe, J. (2011). Folk moral relativism. *Mind and Language*, *26*(4), 482–505.

Simon, D. (2004). A Third View of the Black Box: Cognitive Coherence in Legal Decision Making. *The University of Chicago Law Review*, *71*(2), 511–586.

Simon, D., & Holyoak, K. J. (2002). Structural Dynamics of Cognition: From Consistency Theories to Constraint Satisfaction. *Personality and Social Psychology Review*, *6*(4), 283–294.

Simon, D., Krawczyk, D. C., & Holyoak, K. J. (2004). Construction of Preferences by Constraint Satisfaction. *Psychological Science*, *15*(5), 331–336.

Simon, D., Pham, L. B., Le, Q. A., & Holyoak, K. J. (2001). The Emergence of Coherence over the Course of Decision Making. *Journal of Experimental Psychology: Learning Memory and Cognition*, *27*(5),

1250–1260.
Simon, D., Stenstrom, D. M., & Read, S. J. (2015). The coherence effect: Blending cold and hot cognitions. *Journal of Personality and Social Psychology*, *109*(3), 369–394.
Thagard, P. (1998). Ethical coherence. *Philosophical Psychology*, *11*(4), 405–422.
Wei, T., Yang, S., Moore, J. C., Shi, P., Cui, X., Duan, Q., Xu, B., Dai, Y., Yuan, W., Wei, X., Yang, Z., Wen, T., Teng, F., Gao, Y., Chou, J., Yan, X., Wei, Z., Guo, Y., Jiang, Y., … Dong, W. (2012). Developed and developing world responsibilities for historical climate change and CO2 mitigation. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(32), 12911–12915.
Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1310–1321.