### Moral dimensionality

### Introduction

In modern, culturally-heterogeneous societies, inefficiency of communication of important moral concepts is often evidenced by asymmetrical moral judgements and hypocritical behaviour, especially in our increasingly compartmentalised social landscapes [Rozuel 2011]. This raises the question of how to present target audiences with some (perhaps novel) moral concept, like an ethical dimension of one’s ecological attitude, in a way which would resonate with them, and be conducive to a coherent moral stance, decreasing action-observer biases. We analyse this problem below by introducing a formal moral inconsistency framework.

### Problem outline

We propose to consider the following moral inconsistency problem. Given a moral concept C, find its representation R(C) such that upon presenting a random adult subject with R(C), the chances of them implementing it in their lives in a consistent manner are highest. A few remarks are in order - we do not impose any normative ethical framework (in a sense one might think of R as being induced by it: for example one might wish to explain why acting in accordance with C optimises the utility function, or one might present it in a deontological manner), and further, we would ideally like to understand how such an efficacious R would look like in a culture-independent manner. The following work is structured as follows: firstly, we consider the ways in which the aforementioned problem might be tackled in existing cognitive science of morality (CSM) nativist frameworks, secondly, we investigate it from the perspective of moral learning, and, finally, propose the notion of moral dimensionality of a representation in the context of model-based reinforcement learning, as a useful conceptual tool, although one that requires further research, in particular experimental verification.

### UMG and social intuitionist approaches

We begin by looking at how one might approach explaining moral inconsistency in terms of Universal Moral Grammar theory (see eg. [Mikhail 2006] for its summary), which assumes that a moral judgement is arrived at (unconsciously) through i) deriving a structural description of the given situation ii) performing a computation on the description, which incorporates deontic rules implicit in the subject’s ethics. Now, given a Moral Grammar (ie. a set of derivation rules, structural descriptions, and deontic rules characteristic for a given socio-cultural context) the subject should always arrive at the same moral judgement when presented with stimuli whose structural descriptions coincide. It is therefore difficult to reconcile it with the widespread phenomenon of hypocrisy (see eg. [Hale and Pillow 2015] for a recent psychological exploration of asymmetries in moral judgements of oneself vs others), as it often amounts to espousing contradictory moral stances in situations which differ simply in whether the subject is the judgement-maker of someone else. Such situations, on UMG account, should yield selfsame structural descriptions, leading to matching moral judgments. We therefore claim that one cannot provide a satisfactory UMG-based explanation of moral inconsistency.

Hence we turn our attention to a different CSM framework, social intuitionist theory developed by Haidt ([Graham et al. 2013] provides an introduction to Moral Foundations Theory, building on, and expanding initial social intuitionist claims). Following [Graham et al. 2013] we can briefly summarise this theory, as claiming that moral judgements are products of automatic moral intuitions, not subject to any verification through logic, which on this account is seen rather as a tool serving post-hoc rationalisation. [Demaree-Cotton and Kahane 2018] argue that such an approach introducing a problematic separation of emotions and reasoning which does not find an experimental neurobiological verification, and [Railton 2017] raises additional ‘application’ challenge to the theory - namely, that somehow there has to exist a learnable mechanism for connecting the given sensory input data with the ‘innate’ moral intuitions, for example recognising that in this particular situation someone is being cheated, thus making the theory insufficient to explain integrated moral behaviour. Notice that on this account, the only place there can be an impact of the representation of a moral concept of one’s behaviour is via this unspecified mechanism, as the moral intuitions are by definition non malleable by acts of cognition such as being presented with a representational stimulus.

 Following [Railton 2017] we can present additional evidence to the influence of rational arguments and deliberation on moral judgments, namely an experiment (although not fully rigorous, see the cited work for details) showing statistically significant change of subject’s judgments regarding the desired behaviour of an autonomous vehicle faced with a choice to ride over five pedestrians or drive into a wall even if such an action risks rider’s life. Initially, the majority of the subjects (62%, n=37) answered that it should not drive into the wall. However, after being asked whom they have imagined themselves as when considering the best course of action in this situation, 66% answered: the rider (n=35). Influenced by their own realisation, when answering the same dilemma a week later, the majority decision flipped, as only (37%, n=33) responded with a ‘no’. It seems that an explanation of such a behaviour in the social intuitionist framework would be strained at best due to the character of stimulus, and would have to rest on the allusion to the unspecified mechanism for recognising which intuitions are operative in a given moral scenario.

There exists an extension of the social intuitionist approach, Greene’s Dual Process theory [Greene 2009], which proposes that the intuitional, emotions-based, fast moral judgments operate as ‘Type 1’ processes, whereas there also exist ‘Type 2’ processes, which consist of a rational analysis of costs and benefits. As this model again introduces the neurobiologically unlikely separation of emotions and reasoning, instead of considering this approach here, we will study its modification in the moral learning framework due to [Crockett 2013] in the next section.

### Moral learning

There is however another strand of thought which does not see the need for the strong nativist assumptions. One possibility is to propose a modification of Greene’s Dual Process theory [Crockett 2013], which construes moral deliberation as based on two intertwined yet different processes: one model-free, and the other model-based (the nomenclature being taken from reinforcement learning theory). The first one functions as an ‘intuitive’, fast, and inexpensive - yet relatively rigid and susceptible to bias - assignment of values to <state, action> pairs (c.f. value-function learning in reinforcement learning, see e.g. seminal work of [Mnih et al. 2013]). The other combines a relatively accurate, but computationally expensive causal / evaluative model of reality, with a search algorithm, which queries the model to simulate outcomes of various actions in order to search for the most desirable one (as [Crockett 2013] proposes, this latter process can be seen as a variation on the UMG-style explanation).

 How may we view the moral inconsistency problem through those lenses? Let’s consider two situations whose structural description (in the UMG sense) is identical, yet in one of them the agent is the judgement-making subject, whereas in the other she is an impartial spectator. Now, if she inconsistently judges only one of those scenarios as morally acceptable, then it cannot be the case that those judgments both used model-based approach (as then the arguments against explainability of the moral inconsistency in the UMG approach apply), nor would it be possible that she relied solely on model-free evaluation (as then arguments against explainability of the moral inconsistency in the intuitionist approach apply). Therefore, on this account, it seems that the only possibility is for the agent to rely on model-based reasoning in one judgement, and on a model-free evaluation in the other. However, as presented in [Crockett 2013], this theory lacks the mechanism for directing attention to one of the two processes based on circumstances, not allowing for the discrepancy described (although we do not rule out the possibility of extending this approach so that some such mechanism is incorporated in it).

 Another option in this avenue of inquiry is to assume that moral judgements need not necessarily rely on some task-specific architecture, but rather are ‘implemented’ via general cognitive mechanisms, integrated so that they allow for navigating social landscape [Railton 2017]. This approach proposes to provide understanding of the celebrated sacrificial dilemmas (e.g. the trolley problem introduced in [Foot 1967]) without appeal to intuition / model-free evaluation, but rather (inspired by studies of maze navigation of rats [Langston et al 2010]) through learning with reinforcement of perspectival and non-perspectival models of social interactions. An important aspect of this theory is to insist on the moral model’s spontaneous emergence via the learning process, being a first-order structure, rather than a modification of some pre-existing, native, ‘moral module’. A handful of features of this approach will be useful for us when defining the notion of moral dimensionality, let us emphasise here two of them. Firstly, as a model-based approach, it admits, unlike social intuitionist explanation, quick adaptation through learning and deliberation (see our arguments in the section on social intuitionist theory). Secondly, through positing perspectival and non-perspectival ‘maps’ of social landscape, it easily allows for a perspectival mistake (ie. locating oneself in a situation involving a moral choice) without simultaneously demanding that the subject’s non-perspectival ‘map’ is somehow internally incoherent.

### Moral dimensionality of a representation

The discussion of how one might view the moral inconsistency in various CSM frameworks, led us through nativist accounts, to dual-process learning, and finally to reinforcement learning model-based theory. It is this last one that we will assume as a CSM context for our explication of moral dimensionality.

We first note, that in [Demaree-Cotton and Kahane 2018] an account of understanding the moral situation in which one finds herself is given, on which affective signals from amygdala and rTPJ (right temporoparietal junction) encoding salient moral features of the situation are being fed into vmPFC (ventromedial prefrontal cortex), which in turn mediates integrating it with such forms of assessing the problem as imaginative projections and cost-benefit analysis. This way, rather than sundering emotion-based instincts and cold reasoning, an image of integrity and interconnectedness emerges. A similar conclusions are drawn for macaques in a study on their subjective valuations [Lak et al. 2014], where dopaminergic prediction-error signals are observed to integrate inputs from various reward planes (e.g. amount, risk, type), producing complete and stochastically transitive ordering of choices.

In our assumed CSM context, we will propose the following understanding of the process of perspectival model evaluation. Assume one has just two possible choices in a given morally-relevant situation: A and B, and that after some processing, value embeddings (represented here as real vectors, e.g. of activations in a population of neurons) of those choices have been computed: $v\_{A}, v\_{B}$ which then can be integrated over the vector $w\_{S}$ representing the computed importance in the given situation, of each of the dimensions along which (dis)values are predicted to occur in the wake of taking an action. Thus, we can compute logits of taking action A as

$l\_{A} = <v\_{A}, w\_{S}>$,

and the probability of taking action A as softmax over logits, given ‘temperature’ T, which functions as a parameter for whose large values, the actions will become more equiprobable, and for whose small values the probability will centre on the action with highest logits (see [Lak et al. 2014] for discussion of probable dopamine ‘temperature-like’ effects on softmax-based action sampling)

$P(A) = \frac{e^{Tl\_{A}}}{e^{Tl\_{A}} + e^{Tl\_{B}}}$.

 Notice that in order to compute $l\_{A}$, we have to obtain equidimensional vectors $v\_{A}$ and $w\_{S}$, meaning that for each value we notice in realising the option A (e.g. via the process of imaginative projection), its relevance to the given situation has to be noted, and for each relevant factor of the given situation, a corresponding value for each of the options has to be computed. Therefore, agent’s evaluation of the perspectival model, $w\_{S}$, is influenced by the construal of non-perspectival value predictions $v\_{A}, v\_{B}$. Now, obviously, this is a very simplified model, as it assumes that the final choice is a matter of static integration, rather than one involving time-based heuristics. For example, one might speculate that if the logits computed for one action are very high, some computation-saving mechanism might lead to the decision being made prior to fully evaluating other choices.

Let us also recall what we have said in the case of ‘ethics’ of autonomous vehicles above, namely that sometimes if an aspect of a choice has not been noted previously, the corresponding value that has to be computed for it ‘on the fly’ can be a very poor approximation to the one which would be obtained when all the relevant ‘planes’ of the imaginative projection would be considered, or more cost-benefit analysis allowed.

 Given this framework, we define moral dimensionality of a representation of a moral concept to be the minimal dimensionality of $v\_{R(C)}$ over all situations, and alternative choices available in them in the above decision process. In an informal way, we may describe it as the extent to which one can make such a representation essentially low-dimensional (e.g. summarise it in a statement like ‘abortion is morally acceptable’, or ‘abortion is morally unacceptable’). An example of a high-dimensional representation would be a touching story of one’s relative choosing abortion in some dire circumstances, or detailed ethical analysis of a pro-life stance. We propose that this notion of dimensionality might be correlated with the inconsistency problem: the higher the moral dimensionality, the lower the probability of inconsistency.

### Conclusions

We have reviewed major CSM approaches, and considered how the moral inconsistency problem may be analysed in each of them. We have also introduced a novel concept of moral dimensionality in the context of a very general decision-making model.

 We may speculate that the solution to the moral inconsistency problem will be, given our analysis, cultural-specific, and seems unlikely to be solved in general. For example, someone whose ethical outlook is deontological might favour representations which pertain to law or figures of authority, whereas for a person with a consequentialist convictions a detailed cost-benefit analysis might prove to be high-dimensional. Experimentally determining moral dimensionality as a function of both the concept’s representation, and one’s socio-cultural and ethical background will be an important step forward in understanding the optimal mechanisms for efficient communication of morally-relevant issues.

### Bibliography

**[Mikhail 2006]**, *Universal moral grammar: theory, evidence and the future.* In Trends Cogn Sci. Vol. 11, No. 4

**[Hale and Pillow 2015]** *Asymmetries in perceptions of self and others’ hypocrisy: Rethinking the meaning and perception of the construct* in Eur J Soc Psychol. 2015 February 1

**[Graham et al. 2013]** *Moral Foundations Theory: The Pragmatic Validity of Moral Pluralism* in Advances in Experimental Social Psychology Vol. 47, pp. 55-130.

**[Demaree-Cotton and Kahane 2018]** *The Neuroscience of Moral Judgment.* In Aaron Zimmerman, Karen Jones & Mark Timmons (eds.), The Routledge Handbook of Moral Epistemology. New York: Routledge. pp. 84–104.

**[Railton 2017]** *Moral Learning: Conceptual foundations and normative relevance* in Cognition,

Vol. 167, 2017, pp. 172-190.

**[Crockett 2013]** *Models of morality* in Trends Cogn Sci. Vol 17. No. 8

**[Mnih et al. 2013]** *Playing Atari with Deep Reinforcement Learning*, arxiv preprint

**[Foot 1967]** *The Problem of Abortion and the Doctrine of the Double Effect* in Virtues and Vices Oxford: Basil Blackwell, 1978 (originally appeared in the Oxford Review, Number 5, 1967.)

**[Lak et al. 2014]** *Dopamine prediction error responses integrate subjective value from different reward dimensions* in PNAS Vol. 11, No. 6, pp. 2343-2348.

**[Langston et al. 2010]** *Development of the Spatial Representation System in the Rat* in Science, Vol. 328, Issue 5985, pp. 1576-1580.

**[Rozuel 2011]** *The Moral Threat of Compartmentalization: Self, Roles and Responsibility* in Journal of Business Ethics Vol. 102, No. 4, pp. 685-697.

**[Greene 2009]***The Cognitive Neuroscience of Moral Judgment* in The Cognitive Neurosciences IV, pp. 987-1001.