T

Federico Bina

Models of moral decision-making: Recent advances and normative relevance

1. *Dual-process models and the normative challenge*

Decades of experimental research have been regarded by many as supporting dual-process theories of human cognition, according to which two types of processes – one automatic (type 1), the other controlled (type 2) – are involved in the psychology of judgment and choice (Kahneman 2011; Evans & Stanovich 2013). Dual-process frameworks, however, are controversial, both in descriptive terms and for their potential normative implications. Specifically, disagreement persists about the interactions between type 1 and type 2 processes and their relative reliability. I will refer to the problem of drawing normative conclusions from a better understanding of decision processes as the *normative challenge*[1].

According to dual-process views, type 1 processes provide quick and efficient solutions to ordinary problems. However, these responses are often statistically inaccurate, biased, and unreliable in front of new and complex problems and decisions, due to their inflexible dependence on limited information and insensitivity to new and/or relevant ones (Kahneman 2011)[2]. On the contrary, type 2 operations are more flexible and sensitive to new and relevant information and changes in the decisional environment; they are also responsible for hypothetical thinking, simulation of alternatives,

[1] Evans calls unjustified inferences from description to normative conclusions about reasoning "normative fallacies" (Evans 2019).

[2] At least at the time of decision. As discussed below, type 1 processes are not completely inflexible, since they can significantly learn over time; the point is that they cannot be updated in real time.

 DOI: 10.4454/teoria.v42i2.166

and cost-benefit analyses (CBA). This, of course, requires higher computational costs.

The idea that these differences render type 2 more reliable than type 1 processes has been widely criticized. In particular, critics have emphasized a greater interaction between processes, suggesting that the dual-process image is not accurate (Kruglanski 2013) and that type 1 processes can be subject to sophisticated learning mechanisms, made sensitive to relevant information, and attuned to considered normative standards. Controlled processes can in fact be translated into automatic ones both implicitly and through exercise, as it happens for skill-acquisition and expertise in several domains (Hogarth 2001; Kahneman & Klein 2009). In light of their flexibility, penetrability, and ability to learn, it has been argued that type 1 processes should be considered very reliable in guiding decisions (Gigerenzer 2007).

In what follows, I will explain why these reasons are not sufficient to consider type 1 processes reliable, especially to address new and complex problems, and specifically in the moral domain. This claim is based on a vindicatory etiological and procedural reply to the normative challenge: the reliability of decision strategies is assessed in light of new (non-normative) understanding of the basic processes underlying their functioning, combined with relevant features – e.g. novelty, uncertainty, stakes – of the problems at hand.

2. *Dual-process moral cognition (beyond the reason/emotion divide)*

Dual-process models have been very influential also in recent (neuro) psychological research and empirically-informed ethical debates on moral judgment and decision-making. In the past two decades, empirical studies on (in)famous moral dilemmas have found correlations between characteristically deontological (D) responses and type 1 processes, while characteristically consequentialist (C) judgments correlate with type 2 reasoning (Conway & Gawronski 2013; Greene 2014; Patil et al. 2020).

A few scholars have concluded that these data support consequentialism as a normative theory (Greene 2014; Singer 2005). In sections 4 and 5, I suggest that this conclusion is problematic. Nonetheless, I will argue that empirical research and updated dual-process frameworks can still support significant conclusions for moral theory, though the nature of these conclusions is *procedural* rather than *substantive*.

A big part of the recent scientific and philosophical debate has questioned both Greene's dual-process account and the normative implications that he drew from it. Many critics have stressed that type 1 and 2 processes interact much more than Greene acknowledges; that empirical evidence does not show strong correlations between D judgments–type 1 processes and C judgments-type 2 reasoning; and that type 1 processes can learn and be reason-sensitive, attuned, educated, or trained. For these reasons, critics conclude, type 1 processes are more reliable than Greene maintains (Cecchini 2021; Sauer 2017; Railton 2014, 2017).

Although these claims are true from a descriptive point of view, inferring from them that type 1 processes are reliable in moral decision-making is problematic. As I formulated it, the normative challenge consists in understanding whether we are justified to infer normative conclusions from an increased understanding of the processes underlying moral judgments and decisions[3]. A more detailed description of these processes, therefore, might be of help.

In the past decades, dual-process frameworks have been characterized in several ways: fast vs. slow, automatic vs. controlled, unconscious vs. conscious, habitual vs. goal-oriented, affective vs. rational. I will focus here on a dual-process framework for morality which I believe to be more promising than others for several reasons (see section 3). First of all, this framework denies the problematic – though extremely common and influential – emotion/reason divide. Although this distinction has (historically) been a favorite way of philosophers to understand moral psychology, both critics and advocates of dual-process models have recognized that positing a clear distinction between emotions and reason (or affective and "cognitive" processes) is incorrect, since both type 1 and 2 processes always involve integrative information-processing as well as affective and motivational components (Saunders 2016)[4].

[3] Note that the same strategy is adopted by those who defend the higher reliability of type 1 processes: since they can learn and be sensitive to reasons – they argue – type 1 processes can be reliable.

[4] For instance, processes leading to C judgements do not just elaborate the factual information "5 is more than 1", but also affective elements leading to endorse, or choose, that "saving 5 lives is *better* than saving 1". Moreover, both D and C judgements involve factual information processing: D judgements and emotional reactions are always driven by a clear representation of structural features of the situation, such as personal interaction, the exercise of bodily force (Greene et al. 2009), or direct vs. indirect harm (Royzman & Baron 2002; Cushman et al. 2006).

Denying the emotion/reason distinction, however, does not mean leaving *any* dual-process accounts of moral cognition behind. Experimental research shows that two types of processes can be distinguished in moral as well as in non-moral decision-making, although framed in different ways, and portrayed as deeply interacting and cooperating.

3. *Action-outcome and computational frameworks*

A promising strand of dual-process models (Crockett 2013; Cushman 2013), relatively under-considered in the philosophical literature, frames moral cognition by stressing the distinction between:

1) Attributing value *directly to actions* by associating positive or negative value to them on the basis of a history of feedback (e.g. rewards or losses);
2) Attributing value to expected *outcomes* on the basis of a causal model (a "cognitive map") representing options, values, and transition functions.

These frameworks have two immediate advantages. First, they account for the presence of affective and cognitive information-processing in both types of processes; second, their reliance on learning models account for the diachronic dimension of moral cognition significantly more than first-wave dual-process models did.

These models are also consistent with several studies in moral psychology reporting a preference for indirect over direct harm (Rozyman & Baron 2002), strong aversion to typically harmful actions even when fake or victimless (Cushman et al. 2012; Haidt et al. 1993), and the systematic presence of moral norms across history and societies prescribing the wrongness of specific action-types independently of outcomes (e.g. rituals, food and sexual taboos) (see Graybiel 2008). In these cases, characteristically deontological responses are elicited by the value directly associated with actions, regardless of other relevant information, such as expected outcomes or empathic concern for the subjects involved.

In addition to this evidence, action-outcome frameworks are supported by recent research in computer science and computational neuroscience, reflecting the difference between two basic kinds of reinforcement learning: *model-free* and *model-based* algorithms (Dolan & Dayan 2013).

3.1. Model-free learning and decision-making

Model-free (MF) algorithms work by associating positive or negative value to specific and immediately available actions after a history of rewards, independently of a causal representation of the environment. Imagine an agent A who, when turning right in a state *r* (*round*), gets a reward. If this association occurs a significant number of times, A will associate a positive value to the option "turn right" when in *r* states. Now imagine that A reaches state *r* after turning left in a state *s* (*squared*). Since A associates positive value to state *r*, A will also associate positive value to the option "turn left" when in *s*; and so on, creating adaptive chains of actions.

This mechanism brings A to associate value to the available actions in each particular state on the track leading to a reward, treating each of them as if it was itself a reward. The main advantage of this algorithm is that it is computationally cheap: at each step, it decides on the basis of the value associated with the immediately available action, avoiding costly simulations of future or hypothetical states and comparisons between them. However, and precisely for this reason, MF algorithms are not farsighted. They cannot be goal-oriented – nor prospective in general – because they lack a causal representation of the relation between possible actions and outcomes. This precludes them from any chance to make plans at all: MF algorithms are fundamentally retrospective.

Moreover, although very efficient, MF algorithms are inflexible. They cannot use information to adjust values associated with states, actions, and outcomes (and, consequently, preferences and behavior) because they lack a global representation of them. Value representations can be updated, but this requires time, trial-and-error learning, or interference of strong opposing values (Dickinson et al. 1995).

3.2. Model-based learning and decision-making

By contrast, model-based (MB) algorithms choose by considering available courses of action on the basis of a causal representation – a model, or a cognitive map – of the environment. The model includes causal relations between events (actions, outcomes, rewards, and transition functions) to which A attributes different values; the expected values of the available options are compared, and choices are taken by exploring the decision tree and via CBAs (Dolan & Dayan 2013).

The main downside of this algorithm are its computational costs. Nonetheless, MB strategies can be very flexible, because the model can be updated at any moment by integrating new information and changes in the environment. Imagine that agent A has identified the optimal strategy to reach a reward. Knowing that an obstacle is obstructing the optimal policy (e.g. the fastest route) can make A choose the preferred alternative option in the most efficient way (e.g. without having to face the obstacle on the fastest route before finding an alternative). MB algorithms can be very farsighted, because they can identify clear and complex policies made of long chains of actions, simulating and evaluating consequences of consequences, and modulating value representation accordingly.

In human (moral) cognition, these two types of algorithms interact deeply (Cushman & Morris 2015; Kool et al. 2018). MF mechanisms do not only regulate motor habits or personal harm-aversion, but also the application of rules, principles, and concepts (Dayan 2012); they also facilitate MB decision-making by proposing limited sets of possibilities, thus avoiding the consideration of potentially infinite options in deliberative planning (Phillips & Cushman 2017). But to what extent can the differences between these algorithms – and/or their interaction – be normatively significant?

4. *Addressing the normative challenge*

Greene (2017) argued that the MF-MB distinction provides further support for consequentialism[5]. Like fast-and-frugal heuristics, MF decision-making is generally reliable in front of ordinary contexts and problems, but «it would be a cognitive miracle if we had reliably good moral instincts about unfamiliar moral problems» (Greene 2014, 715). New, complex, and controversial moral problems require MB reasoning. Since empirical research shows strong correlations and similarities between MB thinking and consequentialism, Greene concludes that the latter is the best normative theory to address those kinds of problems.

[5] Greene (2014) illustrates this idea through the analogy with a camera's automatic vs. manual settings. As he noticed later, however, this analogy can be misleading because the automatic settings of standard cameras do not change after they leave the factory, whereas «people's "automatic settings" are constantly evolving through learning [...] The key point, however, is that at the time of decision one is stuck with the automatic settings that one has, regardless of how circumstances might have changed» (Greene 2017, 5).

Note that according to Greene – as for many other advocates of consequentialism – this does not mean that agents should engage in CBA *all the time* (Hare 1981; Brink 1989). MF decision-making can work well in many circumstances, but MB reasoning is more reliable when we have to decide about complex cases, as well as about moral principles, rules, procedures, decision strategies, and whether or not to trust our intuitions. Advocates of deontological and virtue theories, Greene argues, deny this, favoring forms of MF thinking such as reliance on norms or the moral perception of virtuous agents.

These conclusions are partly convincing, but also partly problematic. On the one hand, Greene addresses the normative challenge in a promising way. Consider the following characterization that Railton (2017) recently gave of moral inquiry. Unlike other domains (but similarly to science) the moral discourse aspires to overcome subjective, tribal, elitist, or esoteric points of view and interests by following procedures, and looking for understanding and justification that are impartial, general, consistent, authority-independent, shareable, thinking- and action-guiding, and non-instrumentally concerned with interests and reasons of those actually or potentially affected (Railton 2017, p. 173).

If this characterization is plausible, then the only decision strategy able to accomplish these tasks cannot but be MB reasoning. Consistency, for instance, would be impossible without a model representing the value associated with principles, actions, and outcomes. MB reasoning is also the only strategy allowing us to consider the interests and reasons of others beyond our natural and cultural inclinations, and to evaluate them critically in light of relevant information and alternative possibilities. Moreover, consistent and intersubjectively acceptable moral justifications (Songhorian et al., this volume) cannot but be MB. Referring to a model – models are non-perspectival by definition – is the only way to make one's reasons intelligible to others. Finally, MB reasoning is necessary to link immediately available actions with distant goals, and to consider alternative courses of action (Railton 2017).

On the other hand, however, the idea that the higher reliability of MB reasoning supports consequentialism is problematic. The empirical literature is partly inconsistent on this matter; there are, nonetheless, at least four reasons to doubt such a bold normative conclusion.

1) Studies on confidence and decision-time in moral decision-making suggest that non-C judgments might be the result of MB reasoning *also at the time of decision* (Koop 2013; Gürcay & Baron 2017; Bialek & De Neys 2017);

2) MB reasoning should not be identified uniquely with CBA in act-utilitarian terms, but rather as a broader reflective operation that considers i) information, potential courses of actions and outcomes, ii) intuitions, feelings, rules and principles, and iii) reasons, testing their reciprocal consistency and discarding recalcitrant options (Brink 1989; Campbell & Kumar 2012; Bazerman & Greene 2010).
3) D/non-C judgments can be justifiable even when they are the proximate output of MF processes. First of all, they can be the (distal) output of previous MB reasoning or rationalization. In some cases, justificatory reasons can even track some processes that led to the new "educated" intuition, even if these processes did not intervene at the time of decision (Sauer 2017; Kumar 2017).
4) Finally, also C judgments can be the result of MF processes (Bago & De Neys 2019). For instance, Trémolière and Bonnefon (2014) have shown that the higher the number of lives involved in sacrificial dilemmas, the more intuitive C judgments are. This suggests that C responses can be model-free too, requiring MB reasoning when they are more counterintuitive (Kahane 2012).

To sum up, empirical research and the MF-MB framework support important normative conclusions, though mostly in "procedural" terms, i.e. suggesting how we should think in front of complex or new decisions, and how to justify them. This, however, has no clear direct implications for normative ethical theory in a more substantive way.

5. *Facilitators, conflict detectors, and metacognition*

Some readers might still be unconvinced about the procedural normative conclusion that MB moral reasoning is more reliable than MF mechanisms to address new and complex moral problems. I will briefly consider two possible reasons in favor of this skepticism:

i) In a recent paper, Cecchini argued that default-interventionist models of moral cognition – according to which type 2 (MB) processes intervene to control, endorse, or reject type 1 (MF) outputs – are inaccurate because (MB) moral reflection *fundamentally depends* on (MF) intuitions (Cecchini 2021, 301). In fact, recent research suggests that:

> i.i) MF mechanisms often *facilitate* MB reasoning, providing by default limited sets of options within potentially infinite ones (Phillips & Cushman 2017);

i.ii) MF mechanisms *detect conflicts* between intuitions, reasons, and non-moral information, signaling the need for further reflection (De Neys 2014).

Although these claims are descriptively true, by no means they constitute an objection to the normative conclusion defended here. Operations such as cognitive filtering and conflict detection are not intrinsically reliable: they might be based on, and lead to, either reliable learning histories and actions, or biased and unjustifiable ones[6].

Consider these two cases. First (i.i), agent A might not even consider being fair or kind to a member of a discriminated group, or engaging in sustainable behaviors, because these options might not be included in the default set provided by MF processes as a result of her learning history. Her habits are different and pretty inflexible; she can contemplate different possibilities, but she does not consider *those* actions since the value associated to them is significantly lower than alternatives available at the time of decision. Second (i.ii), intuitive conflict detection and resolution might result in discarding reasonable options (e.g. the less harmful, or the more supported by evidence) because too costly to hold; the conscious reasoning process called upon by intuitive conflict detection might be merely confirmatory of pre-reflective intuitions (Kunda 1990; Haidt 2001).

There is hence no reason to hold MF mechanisms trustworthy in the moral domain just because of their causal role: decisions are often driven by intuitive (MF) processes, but in no way this justifies them. On the contrary, the aforementioned limits of MF algorithms cast doubt on their outputs if no specific convergent support is provided by MB reasoning. In both the aforementioned cases, only MB strategies can critically evaluate whether to endorse the input provided by MF default options or to consider alternative ones. Moreover, only MB reasoning can test whether intuitions are reciprocally consistent and supported by reasons, independently of pre-reflective confidence about their rightness. Deciding uniquely based on the strength of "feelings" or "seemings" is not a defensible strategy (Brink 1989, ch. 5; Harris 2012, 294).

[6] In order to respect Railton's criteria for non-perspectival moral inquiry mentioned above – i.e. for being intersubjectively communicable, understandable and justifiable –, the normative standards needed to assess the reliability of cognitive processes and behavioral outputs cannot but be model-based.

ii) Finally, MF-type 1 mechanisms have been indicated as responsible for the meta-cognitive task of deciding whether MF or MB strategies should be implemented to address specific problems (Cecchini 2021; Thompson et al. 2011)[7]. However, recent studies suggest that when facing a problem, people often engage in CBA weighing the expected outcomes of each strategy (including, in the calculation, the computational costs of MB reasoning), rather than relying on heuristics. Specifically, data show that engagement in MB reasoning – both as metacognitive arbitrator and as the ultimate decision strategy – is proportional to the stakes and levels of uncertainty involved (Kool et al. 2017, 2018). These results are consistent with previous research suggesting that at each time point agents estimate the expected costs and rewards from engaging in a full MB estimation of action-outcome values (Keramati et al. 2011). Although MF processes do play a role in this arbitration, there is no reason for holding them reliable detectors of the right decision mode for specific and complex problems (Bazerman & Greene 2010).

6. *Conclusions*

In this paper I argued that dual-process models of moral cognition are plausible, though they should not be framed in terms of the problematic emotion/reason dichotomy. I also suggested that the distinction between model-free and model-based learning and decision-making algorithms can lead us to draw important normative conclusions. Specifically, in light of a) how they function, and b) the problems we have to face, this framework supports the higher reliability of model-based moral decision-making in front of new, uncertain, and/or complex scenarios. Reliability can be conceived of in terms of justifiability: people would more likely provide – and freely accept – good moral justifications based on non-perspectival model-based reasons, rather than on the subjective "feeling" or "smell" of what is right (although this latter strategy can give rise to effective *post-hoc* rationalizations; see Songhorian et al., this volume).

These conclusions, however, are procedural rather than substantive. Indeed, model-based moral reasoning should not be seen as merely eval-

[7] Evans (2019) hypothesizes a 'type 3' process for this task, presenting aspects of similarity with both type 1 and type 2 processes.

uating outcomes (Cushman 2013), nor as a kind of purely consequentialist form of thinking (Greene 2017), since it can be open to the consideration of several non-consequentialist reasons, norms, intuitions and evaluations (Białek & De Neys 2017). The coherentist mechanism needed to balance all these considerations is a form of model-based reasoning, though it looks closer to a reflective equilibrium than to a pure cost-benefit analysis.

References

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.

Bazerman M.H., Greene J.D. 2010. In favor of clear thinking: Incorporating moral rules into a wise cost-benefit analysis, *Perspectives on Psychological Science*, 5(2), 209-212.

Białek M., De Neys W. 2017, Dual processes and moral conflict: Evidence for deontological reasoners' intuitive utilitarian sensitivity, *Judgment and Decision Making*, 12(2), 148.

Brink D.O. 1989, *Moral realism and the foundations of ethics*, Cambridge University Press.

Campbell R., Kumar V. 2012, Moral reasoning on the ground, *Ethics*, 122(2), 273-312.

Cecchini D. 2021, Dual-process reflective equilibrium: rethinking the interplay between intuition and reflection in moral reasoning, *Philosophical Explorations*, 24(3), 295-311.

Conway P., Gawronski B. 2013, Deontological and utilitarian inclinations in moral decision making: a process dissociation approach, *Journal of Personality and Social Psychology*, 104(2), 216.

Crockett, M.J. 2013, Models of morality, *Trends in cognitive sciences*, 17(8), 363-366.

Cushman F. 2013, Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, 17(3), 273-292.

Cushman F., Gray K., Gaffey A., Mendes W.B. 2012, Simulating murder: the aversion to harmful action, *Emotion*, 12(1), 2.

Cushman F., Morris A. 2015, Habitual control of goal selection in humans, *Proceedings of the National Academy of Sciences*, 112(45), 13817-13822.

Cushman F., Young L., Hauser M. 2006, The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm, *Psychological Science*, 17(12), 1082-1089.

Dayan P. 2012, How to set the switches on this thing, *Current Opinion in Neurobiology*, 22(6), 1068-1074.

De Neys W. 2014, Conflict detection, dual processes, and logical intuitions: Some clarifications, *Thinking & Reasoning*, 20(2), 169-187.

Dickinson A., Balleine B., Watt A., Gonzalez F., Boakes R.A., 1995, Motivational control after extended instrumental training, *Animal Learning & Behavior*, 23(2), 197-206.

Dolan R.J., Dayan P. 2013, Goals and habits in the brain, *Neuron*, 80(2), 312-325.

Evans J.S.B., 2019, Reflections on reflection: the nature and function of type 2 processes in dual-process theories of reasoning, *Thinking & Reasoning*, 25(4), 383-415.

Evans J.S.B., Stanovich, K.E. 2013, Dual-process theories of higher cognition: Advancing the debate, *Perspectives on Psychological Science*, 8(3), 223-241.

Gigerenzer G. 2007, *Gut feelings: The intelligence of the unconscious*. Penguin.

Graybiel A.M. 2008, Habits, rituals, and the evaluative brain, *Annual Review of Neuroscience*, 31(1), 359-387.

Greene J.D. 2014, Beyond point-and-shoot morality: Why cognitive (neuro) science matters for ethics, *Ethics*, 124(4), 695-726.

Greene J.D. 2017, The rat-a-gorical imperative: Moral intuition and the limits of affective learning, *Cognition*, 167, 66-77.

Greene, J.D. Cushman F.A., Stewart L.E., Lowenberg K., Nystrom L.E., Cohen J.D. 2009, Pushing moral buttons: The interaction between personal force and intention in moral judgment, *Cognition*, 111(3), 364-371.

Gürçay B., Baron J. 2017, Challenges for the sequential two-system model of moral judgement, *Thinking & Reasoning*, 23(1), 49-80.

Haidt J. 2001, The emotional dog and its rational tail: a social intuitionist approach to moral judgment, *Psychological Review*, 108(4), 814-834.

Haidt J., Koller S.H., Dias M.G. 1993, Affect, culture, and morality, or is it wrong to eat your dog?, *Journal of Personality and Social Psychology*, 65(4), 613.

Hare R.M. 1981, *Moral thinking: Its levels, method, and point*. Oxford University Press.

Harris J. 2012, What it's like to be good, *Cambridge Quarterly of Healthcare Ethics*, 21(3), 293-305.

Hogarth R.M. 2001, *Educating Intuition*, University of Chicago Press.

Kahane G., Wiech K., Shackel N., Farias M., Savulescu J., Tracey I. 2012, The neural basis of intuitive and counterintuitive moral judgment, *Social Cognitive and Affective Neuroscience*, 7(4), 393-402.

Kahneman D. 2011, *Thinking, fast and slow*. Macmillan.

Kahneman D., Klein G. 2009, Conditions for intuitive expertise: a failure to disagree, *American Psychologist*, 64(6), 515.

Keramati M., Dezfouli A., Piray P. 2011, Speed/accuracy trade-off between the habitual and the goal-directed processes, *PLoS Computational Biology*, 7(5), e1002055.

Kool W., Gershman S.J., Cushman F.A. 2017, Cost-benefit arbitration between multiple reinforcement-learning systems, *Psychological Science*, 28(9), 1321-1333.

Kool W., Cushman F.A., Gershman S.J. 2018, Competition and cooperation between multiple reinforcement learning systems, in Morris, R.W., Bornstein, A., & Shenhav, A. (Eds.), *Goal-directed decision making: Computations and neural circuits*, Academic Press, 153-178.

Koop G.J. 2013, An assessment of the temporal dynamics of moral decisions, *Judgment and Decision Making*, 8(5), 527.

Kruglanski A.W. 2013, Only one? The default interventionist perspective as a unimodel, *Perspectives on Psychological Science*, 8(3), 242-247.

Kumar V. 2017, Moral vindications, *Cognition*, 167, 124-134.

Kunda Z.1990, The case for motivated reasoning, *Psychological Bulletin*, 108(3), 480-498.

Patil I., Zucchelli M.M., Kool W., Campbell S., Fornasier F., Calò M., Cikara M., Cushman, F. 2021, Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures, *Journal of Personality and Social Psychology*, 120(2), 443-460.

Phillips J., Cushman F. 2017, Morality constrains the default representation of what is possible, *Proceedings of the National Academy of Sciences*, 114(18), 4649-4654.

Railton P. 2014, The affective dog and its rational tale: Intuition and attunement. *Ethics*, 124(4), 813-859.

Railton P. 2017, Moral learning: Conceptual foundations and normative relevance, *Cognition*, 167, 172-190.

Royzman E.B., Baron J. 2002, The preference for indirect harm, *Social Justice Research*, 15(2), 165-184.

Sauer H. 2017, *Moral judgments as educated intuitions*. MIT Press.

Saunders L.F. 2016, Reason and emotion, not reason or emotion in moral judgment, *Philosophical Explorations*, 19(3), 252-267.

Singer P. 2005, Ethics and intuitions, *The Journal of Ethics*, 9(3), 331-352.

Songhorian S., Guma F., Bina F., Reichlin M. 2022, Moral progress: Just a matter of behavior?, *forthcoming*.

Thompson V.A., Turner J.A. P., Pennycook G. 2011, Intuition, reason, and metacognition, *Cognitive psychology*, 63(3), 107-140.

Trémolière B., Bonnefon J.F. 2014, Efficient kill–save ratios ease up the cognitive demands on counterintuitive moral utilitarianism, *Personality and Social Psychology Bulletin*, 40(7), 923-930.

Abstract

In the last decades, research in cognitive psychology and neuroscience fueled a rich debate about i) the main mechanisms underlying human (moral) decision-making and ii) their reliability. In this paper, I first make clear that the emotion/reason distinction should be set aside, although this does not imply casting doubt on dual-process models in general. To support this idea, I discuss a dual-process framework for moral decision-making informed by computational models of reinforcement learning. I finally consider some normative implications of this research, stressing their procedural, rather than substantive, nature.

Keywords: dual-process; moral cognition; reinforcement learning; intuition; consequentialism

Federico Bina
Vita-Salute San Raffaele University
f.bina@studenti.unisr.it