A Computational Model of Affective Moral Decision Making that predicts Human Criminal Choices[1]

Matthijs A. Pontier[1], Jean-Louis Van Gelder[2], Reinout E. de Vries[3]

[1]Center for Advanced Media Research Amsterdam (CAMeRA@VU)
Buitenveldertselaan 3, Amsterdam, 1081HV, the Netherlands
matthijspon@gmail.com
[2]Netherlands Institute for the Study of Crime and Law Enforcement (NSCR)
JvanGelder@nscr.nl
[3]VU University Amsterdam, Faculty of Psychology and Education
re.de.vries@vu.nl

Abstract. In the present paper we show that a computational model of affective moral decision making can fit human behavior data obtained from an empirical study on criminal decision making. By applying parameter tuning techniques on data from an initial sample, optimal fits of the affective moral decision making model were found supporting the influences of honesty/humility, perceived risk and negative state affect on criminal choice. Using the parameter settings from the initial sample, we were able to predict criminal choices of participants in the holdout sample. The prediction errors of the full model were fairly low. Moreover, they compared favorably to the prediction errors produced by constrained variants of the model where either the moral, rational or affective influences or a combination of these had been removed.

Keywords: Moral Reasoning, Mathematical Modeling, Cognitive Modeling, Criminal Decision Making, Affective Decision Making, Machine Ethics, Empirical Data

1. Introduction

1.1. Ratio and Affect in criminal decision making

Although there is substantial evidence that emotions are fundamental inputs in the criminal decision making process [1], references to the role of emotions have largely remained confined to narrative or interpretative approaches and rarely made it into choice models of offending. These approaches are limited in terms of gaining insight into the decision making process, as they do not specify the psychological mechanisms according to which they operate [18] or how emotions influence the criminal calculus and alter concerns regarding risk.

The possible interplay between cognition and affect has been prominent in dual-process theories of information processing [5]. Van Gelder [18] argues that criminal decision making processes, perceived as a particular kind of risk taking, may also be insightfully portrayed as invoking these two types of processing. According to the hot/cool perspective of criminal decision making, the cognitive, 'cool', processing mode is sensitive to risk considerations and is therefore likely to respond to notions of sanction severity and certainty, as suggested by deterrence theorists [18]. The cognitive mode is also responsible for balancing costs against benefits and making projections about the long-term consequences of decisions and, consequently, functions much in accordance with the logic assumed by rational choice theory. The affective mode, on the other hand, remains largely unresponsive to probabilities [21]. The dual-process approach applied to criminal decision making can illuminate why notions such as severity of punishment in general have little or no effect on crime rates, why the effect of punishment certainty is only modest, and why recidivism rates are as high as they are.

The present study focuses on the relationship between personality, ratio, affect and criminal behavior. Our point of departure is the HEXACO model. Recently, reanalyses of the same lexical data that have yielded the Big Five model have suggested that instead of five, there are six main dimensions of personality. In the HEXACO model, a sixth cross-culturally corresponding personality dimension named *Honesty–Humility* is added [6]. This trait refers to individual differences in the tendency to be interpersonally genuine, to be unwilling to take advantage of others, to

[1] The full version of this paper will appear in: 16th International Conference on Principles of Practice in Multi-Agent Systems, PRIMA'13: Lecture Notes in Artificial Intelligence, Vol. 8291, pp. 502-509, Springer Verlag, In Press

avoid fraud and corruption, to be uninterested in status and wealth, and to be modest and unassuming. Recent research by Van Gelder and De Vries [20] suggests that the HEXACO model and its Honesty-Humility dimension in particular, is also a strong predictor of criminal behavior.

1.2. Predicting criminal behavior using a computational model

To be able to predict human criminal behavior, we created a computational model of affective moral decision making. To our knowledge, no agent models that also include affect and personality to predict crime have so far been proposed. As a first step in this direction, we integrated a moral reasoning system that matched the decision of medical ethical experts [12] and an empirically validated model of affective decision making [7]. We extended the affective moral decision making module so that the agent can take into account anticipatory emotions during the decision making process. The section below will explain the model in more detail.

We obtained empirical data to test whether the model can predict human criminal behavior. In simulation experiments, we optimized the weights for the moral, rational and affective influences in the decision making process and the morality of the criminal choice, based on the first half of the sample, using parameter tuning, similar to [2]. With the obtained weights, we tested the predictions for the holdout sample (i.e., the remaining half of the participants) using seven different versions of the model: the full model and constrained versions of the model, in which one or two of the three influences in the decision making process (i.e., personality, ratio and affect) were removed. We hypothesized that the full model would fit the data the best. Because of the generic form of the model, we expect that if the model successfully predicts human affective moral decision making (i.e., criminal behavior), it can also be used to simulate human affective ethical decision making.

2. The computational model of affective moral decision making

In the rational moral reasoning system [12], the agent tries to estimate the morality of actions by holding each action against the moral principles inserted in the system and picking actions that serve these moral goals best. The moral goals inserted into the system are (1) autonomy, (2) beneficence, (3) non-maleficence and (4) justice. The agent calculates the estimated level of Morality of an action by taking the sum of the ambition levels of the moral goals multiplied with the beliefs that the particular actions facilitate the corresponding moral goals:

$$\text{Morality(Action)} = \Sigma_{\text{Goal}}(\text{ Belief(Action facilitates Goal))} * \text{Ambition(Goal))} \quad (1)$$

This can be represented as a weighted association network, where moral goals are associated with the possible actions via the belief strengths that these actions facilitate the four moral goals.

However, only focusing on balancing principles through rational argumentation may lead to the underexposure of the role of social processes of interpretation and communication [10]. To be able to capture these human moral decision making processes, we integrated the moral reasoning system of Pontier and Hoorn [12] with Silicon Coppélia [8], a computational model of emotional intelligence that is capable of affective decision making. During the process, the agent retrieves beliefs about actions that facilitate or inhibit the desired or undesired goal-states. This is to calculate an *ExpectedUtility* [0, 1] of each action. Actions that facilitate desired goals or inhibit undesired goals will have a high ExpectedUtility [8]. In an affective decision-making module, affective and rational influences are combined in the decision-making process. By combining moral reasoning and affective decision making into Moral Coppélia, human moral decision making processes could be simulated that could not be simulated using the moral reasoning system alone [15].

In the previous affective decision making module in Moral Coppélia, emotions were only implicitly regulated, by picking actions that lead to desired goals. To be able to account for *Negative State Affect* in Moral Coppélia, we added *ExpectedEmotionalStateAffect* (EESA) [0, 1] to the affective moral decision making module. Here, a high EESA indicates that an action is expected to improve the emotional state of the agent, whereas a low EESA indicates that an action is expected to worsen the emotional state. Hereby, we more explicitly add the emotion regulation strategy *situation selection* of Gross' model of emotion regulation [4] to the system.

For calculating the EESA, we added *ActionEmotionBeliefs* (AEB) [0, 1] to the system. An AEB(action, emotion) represents the belief that an action will lead to a certain level of emotion. For example, an AEB(shoplifting, excitement) of 0.6 represents the belief that shoplifting will lead to a level of excitement of 0.6. The ExpectedEmotion (EE) [0, 1] is calculated using formula 2:

$$\text{EE(action, emotion)} = (1-\beta) * \text{AEB(action, emotion)} + \beta * \text{current\_emotion} \quad (2)$$

In this formula, the persistency factor β is the proportion of emotion that is taken into account to determine the EE. The new contribution to the emotion response level is determined by taking the appropriate AEB.
To determine the EESA of an action, a weighed sum of the discrepancy between desired emotions and expected emotions after performing the action is subtracted from 1. For simplification, the weights *w(i)* were set to the same level for all emotions added to the system:

$$EESA(action) = 1 - (\sum_{n}^{0} w(i) * (Desired(emotion(i)) - EE(action, i))) \quad (3)$$

To determine the *ExpectedSatisfaction* [0, 1] of a criminal choice, a weighed sum is taken of the Morality, the rational ExpectedUtility and the emotional EESA of the action:

$$\begin{aligned} &ExpectedSatisfaction(action) = \\ &w_{mor} * \ Morality(action) + \\ &w_{rat} * \ ExpectedUtility(action) + \\ &w_{emo} * \ ExpectedEmotionalStateAffect(action) \end{aligned} \quad (4)$$

3. Matching the data to the model

153 undergraduate psychology and educational science students from a university in the Netherlands were approached by email to participate in a short scientific study about dilemmas. Two scenarios were used to measure the mediating and outcome variables. Both scenarios described illegal behavior that can be classified as common, minor crime, i.e., illegal downloading and insurance fraud. Both scenarios were followed by a set of items measuring anticipated sanction probability and severity, negative affect, and criminal choice. For more information about the scenarios and the procedure, see [20].
For matching the data to the model, we transformed all obtained data to the domain [0, 1]. Subsequently, we populated a virtual environment with agents that estimated the probability of making a criminal choice. Each agent was coupled to a participant. The goals inserted into the system were 'profit from a criminal choice' and 'not getting caught'. The emotions inserted into the system were 'hope', 'fear', 'joy' and 'sadness'. For each agent, the rational beliefs about actions relating to goals were set to a level so that the ExpectedUtility of an action matched the Perceived Risk of the participant. Additionally, the beliefs about actions relating to emotions were set to a level that the EESA of the criminal choice matched the Negative State Affect. The weight of the morality in the decision-making process was set proportional to the level of the trait 'Honesty-Humility' in the participant. To divide the remaining weight for calculating the expected satisfaction of a criminal choice, the rational and emotional influence were each assigned a part of the remaining weight, where we made sure that $part_{rat} + part_{emo} = 1$. In formula 5 and 6, w_{rat_opt} and w_{emo_opt} represent the optimal weights found with parameter tuning for the rational and affective influences in the decision making process.

$$w_{rat} = (1-w_{mor}) + part_{rat} * w_{rat\_opt} \quad (5) \qquad w_{emo} = (1-w_{mor}) + part_{emo} * w_{emo\_opt} \quad (6)$$

With the found weights, we tested the predictions for the holdout sample (i.e., the remaining half of the participants) using seven different versions of the model: the full model and constrained versions of the model, in which one or two of the three influences in the decision making process (i.e., personality, ratio and affect) were removed.
The quality of fit was determined by investigating the discrepancy between the expected satisfaction of the agents (i.e., their prediction of the behavior of their human counterparts) and the likelihood of criminal choice as reported by the participants. The coefficient of determination R^2 [17] was calculated to determine the quality of the fit (the closer to 1 the better). The match was called satisfactory when the quality of fit did not increase anymore for several time steps. If the matching process seemed to be stuck into a local optimum, the parameters were adjusted by intuition to check whether the match could be improved.

4. Results

Table 1 shows the results of the simulation experiments. In experiment 1, we tried to predict the criminal choice of the participants by agents using only the rational expected utility in the decision making process. This resulted in an R^2 of 0.719 for the holdout sample. In experiment 2, only making use of the Expected Emotional State Affect (EESA) of a criminal choice resulted in an R^2 of 0.906 for the holdout sample. In experiment 3, optimally tuning a combination of ratio and affect resulted in a $part_{rat}$ of 0.34 and a $part_{emo}$ of 0.66, leading to an R^2 of 0.9323 for the holdout sample. In experiment 4, using only moral reasoning resulted in an R^2 of 0.9281 for the holdout sample. In experiment 5, an optimally tuned combination of moral reasoning and ratio resulted in an R^2 of 0.9803 for the holdout sample. In experiment 6, an optimally tuned combination of moral reasoning and affect resulted in an R^2 of

0.9778 for the holdout sample. Experiments 5, 6 and 7 found similar values for the morality of the criminal choice (mor_{cc}).

Table 1. Simulation results

	Exp 1	**Exp 2**	**Exp 3**	**Exp 4**	**Exp 5**	**Exp 6**	**Exp 7**
mor_{cc}	0	0	0	0.68	0.42	0.453	0.435
w_{mor}	0	0	0	1.00*hh	0.96*hh	0.97*hh	0.87*hh
$part_{rat}$	1	0	0.34	0	1	0	0.64
$part_{emo}$	0	1	0.66	0	0	1	0.36
R^2 initial	0.7553	0.8792	0.9222	0.9336	0.9871	0.9798	0.9881
R^2 holdout	**0.7192**	**0.9060**	**0.9323**	**0.9281**	**0.9803**	**0.9778**	**0.9821**

The optimal fit was found in experiment 7. Here, we tested the full model, including moral reasoning, ratio and affect in the decision-making process. Parameter tuning led to a $part_{rat}$ of 0.64 and a $part_{emo}$ of 0.36, resulting in an R^2 of for the simulation of the 0.9881. The R^2 for the predictions of the holdout sample was 0.9821.

5. Discussion

We asked the participants to estimate the probability of making a criminal choice in two scenarios, and assessed their perceived risk and the negative state affect with respect to different criminal choices using a scenario design. Additionally, we measured the personality dimension Honesty-Humility of the participants. We extended a model of affective moral decision making, Moral Coppélia [15], and matched the participants to agents equipped with the model. We applied parameter tuning techniques and found optimal parameter settings to fit the initial sample. Using the obtained parameter settings, we predicted the criminal choice of the participants in the holdout sample. The prediction errors that were found turned out to be fairly low. Thereby we have shown that an extended version of Moral Coppélia can fit empirical data. This can be seen as a form of ecological validation.

Moreover, we compared the prediction errors with those produced by constrained variants of the model where either the moral, rational or affective influences or a combination of these had been removed. The best predictions were produced by the full model, which confirms our hypothesis.

This is an important indication that making a criminal choice is dependent on the participants' personality, rational choice considerations, as well as emotions. This corresponds with recent informal models of criminal decision making [18, 19, 20]. Thereby the current findings strengthen these informal models. We show that the models can be used to reproduce and predict human criminal decision making.

There are many applications in which a combination of moral reasoning, rational choice considerations as well as emotions is useful. In the first place, the model can be used to predict criminal behavior in humans. Additionally, Moral Coppélia can be used to develop intelligent agents for a wide variety of applications, such as (serious) digital games, tutor and advice systems, or coach and therapist systems. Another possible use is in software and/or hardware that interacts with a human and tries to understand this human's cognitive and emotional states and processes and responds in an intelligent manner. The system can combine sensor data as input to project Moral Coppélia in the user to maintain their emotional state. This can enable the system to adapt the type of interaction to the user's needs.

Additionally, there are many applications in which agents should not behave ethically 'perfect' in a rationalist sense. They should be able to distinguish between right and wrong. In a training simulation or serious game, police officers may not always be effective when they 'play it nicely.' Sometimes they have to break the moral rules (e.g., lie or cheat) to achieve a higher goal (e.g., prevent a murder). The need to be context-sensitive and not rigidly follow rational principles is crucial in all human interaction.

Furthermore, Moral Coppélia can be used to develop agents for interactive storytelling. A trend in developing virtual stories is the movement from stories with a fixed, pre-scripted storyline toward emergent narratives; i.e., stories in which only a number of characters and their personalities are fixed, rather than the precise script of the story. In emergent narratives, ideally, all the designer (or writer) has to do is to determine which (types of) characters will occur in the play, although usually it is still needed to roughly prescribe a course of events. To accomplish complex personalities with human-like properties such as emotions and theories of mind, researchers have started to incorporate cognitive models within agents (e.g., [3]). Moral Coppélia can be seen as a next step into this direction. The agent can combine moral reasoning with rationality and emotions to make decisions on its own. The agent can simulate emotions, and regulate them upwards as well as downwards using various emotion regulation strategies.

Agents telling stories are not only useful to make the elderly feel less lonely. Autonomous agents that can affectively make moral decision are also applicable in an entertainment context (e.g., computer games, see [13]). Additionally,

the use of autonomous agents also proved to be useful for clinical experts in the treatment of behavior problems, family counseling, and training [11], education [16], or in persuasive contexts (e.g., science and health communication), or clinical therapy [9].
In particular, agents can play a useful role in the interaction between human and computer in a Web context. One of the application areas foreseen is in self-help therapy, in which humans with psychological disorders are supported through applications available on the Internet and virtual communities of persons with similar problems. An agent equipped with Moral Coppélia can respond empathically toward the user. Together with expert knowledge, the agent can use the model to behave emotionally intelligent and give 'the right response at the right moment'.
As is, the moral reasoner with rational and affective components only allows choosing from given decision options in scenarios. In future research, we additionally want to explore what happens if the Caredroid makes use of computational creativity to propose alternatives that include more information than the offered decision options. Finally, we would like to extend autonomy in the moral reasoning system to be able to distinguish positive and negative autonomy [14]

Acknowledgements. This study is part of the SELEMCA project within CRISP (grant number: NWO 646.000.003).

6. References

1. Athens, L.: Violent encounters, violent engagements, and tiffs. Journal of Contemporary Ethnography 34, 631–78 (2005)
2. Bosse, T, Brenninckmeyer, J., Kalisch, R., Paret, C., Pontier, M.A.: Matching Skin Conductance Data to a Cognitive Model of Reappraisal. In: Proceedings of of the 33th International Annual Conference of the Cognitive Science Society, CogSci'11, pp. 1888-1893 (2011)
3. Bosse, T., Pontier, M.A., Siddiqui, G. F., Treur, J.: Incorporating emotion regulation into virtual stories. In: IVA'07. Lecture notes in artificial intelligence, vol. 4722, pp. 339-347. Springer Verlag (2007)
4. Bosse, T., Pontier, M.A., Treur, J.: A Computational Model based on Gross' Emotion Regulation Theory. Cognitive Systems Research Journal, 11, 211-230 (2010)
5. Chaiken, S., Yaacov, T.: Dual-Process Theories in Social Psychology. New York: Guilford Press. (1999)
6. De Vries, R.E., Ashton, M.C., Lee, K.: De zes belangrijkste persoonlijkheidsdimensies en de HEXACO Persoonlijkheidsvragenlijst. Gedrag en Organisatie 22:232–274 (2009)
7. Hoorn, J.F., Pontier, M.A., Siddiqui, G.F.: When the user is instrument to robot goals. In: Proceedings of the Seventh IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT'08, pp. 296-301 (2008)
8. Hoorn, J.F., Pontier, M.A., Siddiqui, G.F.: Coppélius' Concoction: Similarity and Complementarity Among Three Affect-related Agent Models. Cognitive Systems Research Journal, 33-49 (2012)
9. Lee, E., Leets, L.: Persuasive storytelling by hate groups online: examining its effects on adolescents. American Behavioral Scientist, 45, 927-957 (2002)
10. Ohnsorge, K., Widdershoven, G.A.M.: Monological versus dialogical consciousness - two epistemological views on the use of theory in clinical ethical practice. Bioethics. 25(7), 361-369 (2011)
11. Painter, L.T., Cook, J.W., Silverman, P. S.: The effects of therapeutic storytelling and behavioral parent training on noncompliant behavior in young boys. Child and Family Behavior Therapy, 21(2), 47-66 (1999)
12. Pontier, M.A., Hoorn, J.F.: Toward machines that behave ethically better than humans do In: Miyake, N., Peebles, B., Cooper, R.P. (eds.) Proceedings of of the 34th International Annual Conference of the Cognitive Science Society, CogSci'12, pp. 2198-2203 (2012)
13. Pontier, M.A., Siddiqui, G.F.: An Affective Agent Playing Tic-Tac-Toe as Part of a Healing Environment In: J.J. Yang et al. (eds.), PRIMA'09, Lecture Notes in Artificial Intelligence, Vol. 5925, Springer Verlag, pp. 33-47 (2009)
14. Pontier, M.A., Widdershoven, G.A.M.: Robots that stimulate our autonomy. In: IFIP Advances in Information and Communication Technology, AIAI'13, in press.
15. Pontier, M.A., Widdershoven, G.A.M., Hoorn, J.F.: Moral Coppélia - Combining Ratio with Affect in Ethical Reasoning. In: Advances in Artificial Intelligence – IBERAMIA 2012, Lecture Notes in Computer Science, Vol. 7637, pp. 442-451 (2012)
16. Schlosser, R.W., Lloyd, L.L.: Effects of initial element teaching in a story-telling context on Blissymbol acquisition and generalization. J of Speech and Hearing Research, 36, 979-995 (1993)
17. Steel, R.G.D. Torrie, J.H.: Principles and Procedures of Statistics, New York: McGraw-Hill, pp. 187-287 (1960)
18. Van Gelder, J.L.: Beyond rational choice: The hot/cool perspective of criminal decision making. Psychology, Crime & Law. 19, 745-763 (2013)
19. Van Gelder, J.L., De Vries, R.E.: Traits and states: Integrating personality and affect into a model of criminal decision making. Criminology, 50, 637-671 (2012)
20. Van Gelder, J.L., De Vries, R.E.: Rational misbehavior? Evaluating an integrated dual-process model of criminal decision making. Journal of Quantitative Criminology (2012)
21. Van Gelder, J.L., De Vries, R.E., Van der Pligt, J.: Evaluating a dual-process model of risk: affect and cognition as determinants of risky choice, J Behavioral Decision Making, 22, 45–61 (2009)