Carambola: Enforcing Relationships Between Values in Value-Sensitive Agent Design

Luis Garcia[1][0000−0002−8823−1967] and Chris Martens[1][0000−0002−7026−0348]

Northeastern University, Boston MA 02115, USA

Abstract. Carambola is a text-based strategy game that operationalizes the Theory of Basic Values (TBV) to model the motivations of its non-player characters (NPC) and the dilemmas it presents to players. The player takes on the role of the Emperor of a nation, making a series of executive decisions while noting the subsequent reactions of their NPC advisors. After a fixed number of rounds in which they choose actions, their NPC advisors vote on whether they should dethrone the player based on the affinity they have with the other subjects of the game. Advisor affinity is affected by the Emperor's actions, which each harm and promote a subset of their values. Our implementation of the TBV is a geometric interpretation that enforces restrictions on the attitudes that agents can have toward the values. We give a brief overview of the theory, and then describe our implementation and our plans for evaluating how this usage of the TBV affects the advisors' believability.

Keywords: social simulation, social psychology, Theory of Basic Values, values, value-sensitive narrative, dilemma generation, character generation, believability, affinity

1 Introduction

Role-playing games occasionally present ethical dilemmas to players wherein the values of at least two opposing parties are at stake. For example, in the game *Fallout: New Vegas*, players must choose to side with one of three factions who have conflicting ideological views toward governing a post-apocalyptic Mojave Desert. Researchers in the field of computational narrative have sought to utilize dilemmas in order to generate stories that maintain some level of narrative interest [6] or train people on ethical behavior [8][11]. Widely speaking, the operationalization of dilemmas can make our games not only more entertaining, but also more significant as vectors for ethical discourse [12].

Past projects that feature the theme of dilemma resolution rely on the arbitrary specification of values and mechanics (e.g., actions) extending those values that, by careful authoring, manifest an ethical system that is specific to the game-world in question [6][8][9]. In narrative generation tools, this reliance can pose issues for the believability of the game's characters. Authors must maintain coherence between the attitudes of their characters toward the values in their

system, the actions available to those characters (and the player), and the ethical and moral consequences of those actions. While specifying values and the mechanics extending them, authorial errors can bring about the generation of agents whose values are incoherent. For example, they may cherish values that are supposed to be diametrically opposed according to the game's world. In such a situation, players may see the agent's behavior as inconsistent, or "buggy" [3].

Toward reducing the possibility of incoherent character behavior, we introduce a method of value-sensitive agent generation in our game, Carambola. It is a turn-based game wherein the player, taking the role of an emperor making executive decisions, is presented a sequence of dilemmas that lead either to them retaining their throne or losing it. Each choice they make elicits reactions from their NPC advisors based on the advisors' attitudes toward the values that the actions affect. Their reactions nudge them toward voting for either one of the player's possible outcomes (see Figure 1). The advisors are randomly generated at each game start. Their attitudes and the values are given according to a geometric interpretation of the TBV that ensures that no character can cherish two diametrically opposed values at once [7]. We describe the underlying theory and our implementation. We hypothesize that our design improves the advisors' believability. A plan for evaluating our claim is described.

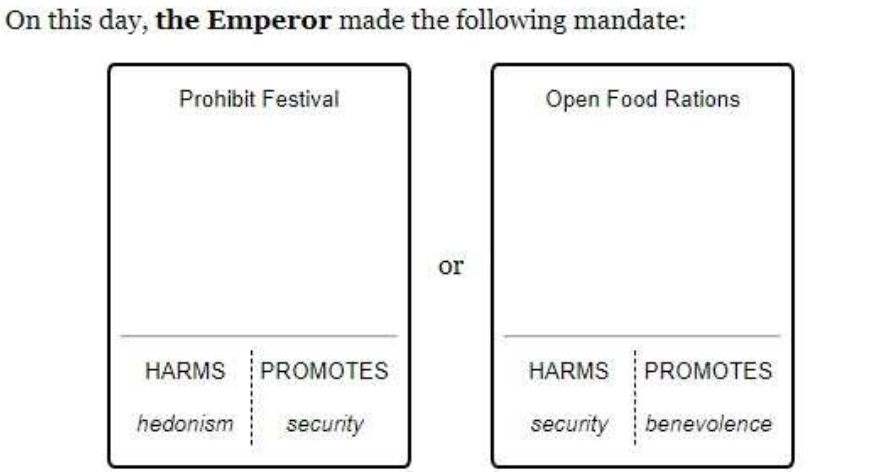
2 Related Work

Carambola's design is inspired by past interactive projects which feature value-sensitive agents [5][10]. Behind one of these projects ([10]) is one of the earliest examples of value-sensitive narrative generation, IDTension [9]. Here, an author defines values via abstract keywords (e.g., *non-violence*, *law*). The author also configures both agents' attitudes toward each of these values and actions that are symbolic of the values. Carambola takes a similar approach to agent and action design. However, in Carambola the values are not defined by us, the designers, but are rather fixed in accordance to the TBV. Furthermore, our implementation enforces relationships between the values, precluding authorial mistakes that can produce situations in which agents can simultaneously hold positive attitudes toward diametrically opposed values, such as *violence* and *non-violence*.

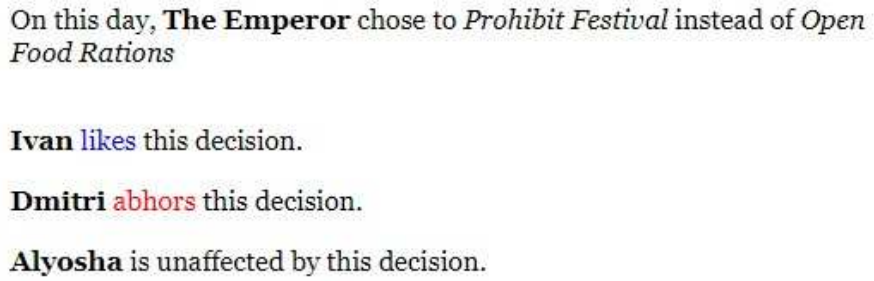
The main mechanic of our game, dilemma resolution, is largely inspired by proposed frameworks for dilemma generation [1][4]. We follow the EGAD framework, which introduces the use of the TBV to supplement author-specified values [4]. According to the framework, agents may cherish, despise, or be ambivalent to each of the values, while actions may promote, harm, or do nothing to each. We implement this framework for Carambola using only the values from the TBV. We extend EGAD by operationalizing the TBV's geometric property, which is left ignored by the framework.

Fig. 1: On each round, the player is presented with a dilemma (a). After a decision is made, the advisors react (b). At the end of the game, the advisors decide whether to retain the player as emperor by majority vote (c).

(a) A dilemma presented to the player. They must choose either one to progress the game.



(b) The player's choice and the reactions of each of their advisors.



(c) The "dethrone" ending of the game, one of two possible endings for the player.

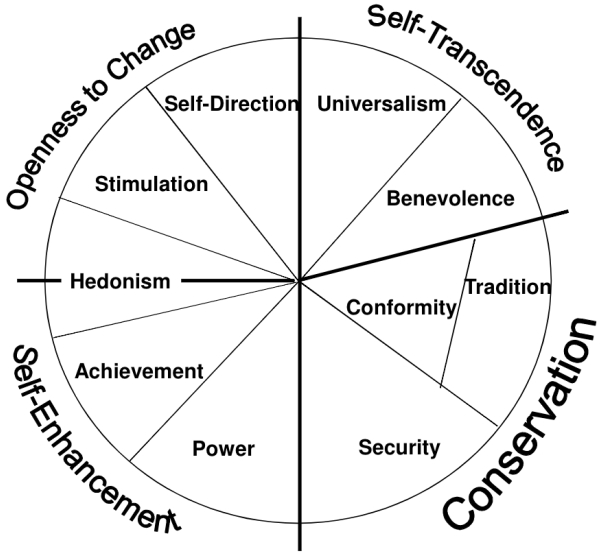


Fig. 2: The continuum of values introduced by Schwartz's Theory of Basic Values. The proximity of the values reflects how similar their underlying motivations are.

3 Theoretical Background

At the core of Carambola's design is an implementation of the TBV, which posits that there is a shared set of values across cultures worldwide [7]. The theory has two facets which are relevant to our implementation: the values themselves and a geometric model for how the values are related to each other. According to the TBV, these universal values can be placed in a circular continuum in which the proximity of the values represents the amount of similarity their underlying motivations have (see Figure 2).

In Carambola, we use this adjacency relationship from the TBV to enforce restrictions on the possible configurations of advisor attitudes toward the values. The advisors are each motivated to promote or maintain their empire's general wellness. However, that motivation manifests differently for each advisor through their personal attitudes toward the values that are promoted and harmed by the player's choices. Following the EGAD framework, the advisors can hold three attitudes: cherish, despise, or ambivalent about [4]. We extend the framework by ensuring that the values that advisors cherish (or despise) lie adjacent to each other on the continuum.

4 Game Design

To facilitate the player's ability to make reasoned decisions in the game, we sought to construct the advisors so that what their reactions and reasoning are

consistent and clear to the player. We achieve consistency by ensuring that the values that the advisors cherish (or despise) are adjacent to each other on the value continuum, thus being more similar to each other with regard to their underlying motivations [7]. Clarity is achieved by plainly stating the advisors' reactions, their affinities toward the player, and by labeling actions such that there is a simple thematic link between what they represent and the values they promote.

4.1 Action Design

On each round, Carambola presents the player with two alternative actions that each promote and harm a value (see Table 1). The action specifications were handwritten so that their effects fit in thematically with their labels. For example, *Authorize Military March* shows off the glory of Carambola's military (promoting their *achievements*) while reinforcing the force that the empire has over its citizens (harming *universalism*). Upon choosing an action, the player triggers its effects on the values, which elicit reactions from the advisors. Furthermore, we wrote the actions so that the values they promote are always diametrically opposed to the values that they harm.

Table 1: A list of all available actions Carambola, along with their effects on the values. Every action has an opposite version, where its effects are flipped from the original.

Label	**Value Promoted**	**Value Harmed**
Maintain Barracks	power	universalism
Authorize Military March	achievement	universalism
Authorize Festival	hedonism	security
Maintain Art Museum	stimulation	conformity/tradition
Pardon Criminal	self-direction	conformity/tradition
Maintain Hospital	universalism	power
Open Food Rations	benevolence	security
Enforce Mass	conformity/tradition	stimulation
Maintain Prison	security	self-direction

4.2 Advisor Design

To facilitate discussion about our advisor design, we will refer to Dmitri, an example advisor that can be generated in the game (see Figure 3).

Value Attitudes Consistent with dilemma generation systems in the past, we designed advisors so that they each have values that they cherish, despise or are ambivalent toward [1][2][4]. At game start, we generate their attitudes toward the values according to the following rules:

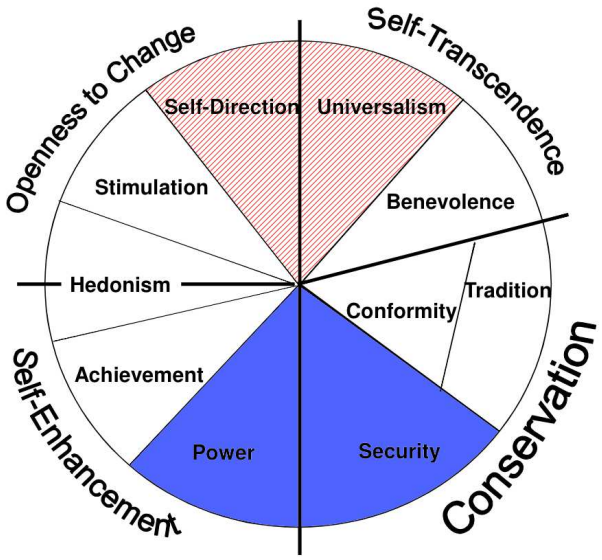


Fig. 3: Value attitude configuration for Dmitri, an example advisor. Dmitri cherishes *power* and *security*, and despises *self-direction* and *universalism*. He is ambivalent toward the rest of the values.

1. The two values that an advisor cherishes (despises) are adjacent on the value continuum.
2. For an advisor to despise a value, it must be on the opposite half of the value continuum from the ones that they cherish.

These two rules enforce a geometric interpretation of the TBV, where adjacent values have similar motivations, while values that are opposite from each other on the continuum can conflict [7]. In our example, Dmitri cherishes *power* and *security*, but not *power* and *universalism* because the latter pair are on opposite sides of the continuum. For the purpose of ensuring that Dmitri's values are clear to the player, this is ideal: while *power* and *security* together emphasize control and the overcoming of threats, *universalism* invites diversity and self-expression.

Reactions After the player chooses one of the alternatives presented to them, each of their advisors takes a turn to react. Computationally, an advisor's reaction is the sum of points that the player's choice yields, with points being given according to Table 2. Thus, an advisor's reaction can range from being very positive (yielding 2 points) to neutral (yielding 0 points) to very negative (yielding -2 points). This sum is added to the advisor's overall affinity to the player.

To illustrate, suppose the player chooses *Maintain Barracks*. Because Dmitri cherishes *power* and this choice promotes it, the player receives +1 point. Because Dmitri despises *universalism* and this choice harms it, the player receives

Table 2: An illustration of how the effects (promote or harm) of the player's choice and an advisor's attitude (cherish or despise) interact to produce points.

		Player Choice Effect	
		promote	*harm*
Advisor Attitude	*cherish*	+1	-1
	despise	-1	+1
	ambivalent	0	0

+1 additional point. Dmitri's overall reaction, then, is very positive, giving +2 points. His affinity toward the player moves in the positive direction.

5 Future Work

We plan on evaluating the effect that Carambola's implementation of the TBV has on the believability of its NPCs. Our hypothesis is that using this representation of the TBV to generate the advisors' reactions makes them more believable than if their reactions were generated without regard for the values that the player's choices affect. To assess this, we will extend the list of actions in Carambola so that there are pairs of actions that affect the same values in the same way. For example, we may introduce the following two pairs:

1. *Increase Weapons Manufacturing* and *Occupy a Neighboring City*, which each promote *power* and harm *universalism*.
2. *Enforce Attendance to Mass* and *Close All Business for Holiday*, which each promote *tradition* and harm *self-direction*.

Our evaluation will have two study cases that differ in how the advisors would react to player choices. Our test case will be of the implementation detailed thus far, where advisors react according to their values. For each action pair, they will react exactly the same way for either action. In the control case, an advisor will favor one action in a pair, but disfavor the other one. Intuitively, if our hypothesis is correct, the control case would introduce a level of inconsistency in the advisors' reactions that will break players' suspension of disbelief.

To test our hypothesis, we will use the methods proposed in Gomes et al. [2013] to quantify the difference in character believability between the control and test cases [3]. Believability is split into a number of dimensions that describe different aspects of agent behavior. For example, one dimension is *behavior coherence*, which is the degree to which a human observer may deem an agent's actions to be logical according to their mental model of the agent's state. During a user study, we will measure these dimensions, and then find if the test case outperforms the control case in terms of the criteria also described by Gomes et al. [2013].

5.1 Extending the Model

Currently, the advisors' attitudes toward the values are of a ternary set: cherish, despise, or ambivalent toward. Furthermore, as designed, the advisors cannot hold positive attitudes toward conflicting values. Lastly, the advisors can only care about four of the values at a time. While these limitations work toward our current goals for Carambola's player experience, Schwartz's theory alone does not enforce them. One can imagine a character in a different piece of media who holds more complex values, with their attitudes having been formed from a combination of practical necessity and lived experience. We speculate that switching from a discrete set of attitudes to a continuous one and allowing agents to have attitudes for all values on the continuum will be a step toward introducing such nuance.

References

1. Barber, H., Kudenko, D.: Generation of adaptive dilemma-based interactive narratives. IEEE Transactions on Computational Intelligence and AI in Games **1**, 309–326 (12 2009). https://doi.org/10.1109/TCIAIG.2009.2037925
2. Battaglino, C., Damiano, R., Lesmo, L.: Emotional range in value-sensitive deliberation (2013), http://hdl.handle.net/2318/147231
3. Gomes, P., Paiva, A., Martinho, C., Jhala, A.: Metrics for character believability in interactive narrative (2013)
4. Harmon, S.: An expressive dilemma generation model for players and artificial agents. The Twelfth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE-16) pp. 176–182 (2016)
5. Kybartas, B., Verbrugge, C., Lessard, J.: Subject and subjectivity: A conversational game using possible worlds. vol. 10690 LNCS, pp. 332–335. Springer Verlag (2017). https://doi.org/10.1007/978-3-319-71027-337
6. Mateas, M., Mawhorter, P., Wardrip-Fruin, N.: Intentionally generating choices in interactive narratives. pp. 292–299 (2015), http://inform7.com/
7. Schwartz, S.H.: An overview of the schwartz theory of basic values. Online Readings in Psychology and Culture **2** (12 2012). https://doi.org/10.9707/2307-0919.1116
8. Si, M., Marsella, S.C., Pynadath, D.V.: Thespian: An architecture for interactive pedagogical drama. pp. 595–602 (2005)
9. Szilas, N.: Idtension: a narrative engine for interactive drama (2004)
10. Szilas, N.: The mutiny: an interactive drama on idtension. pp. 539–540 (2008)
11. Upright, R.L.: To tell a tale: The use of moral dilemmas to increase empathy in the elementary school child (2002)
12. Zagal, J.P.: Ethically notable videogames: Moral dilemmas and gameplay (2009)