

On the Quest for an NLP-Driven Framework for Value-Based Decision-Making in Automatic Agent Architecture

Alicia Pina-Zapata^a and Sara García-Rodríguez^b

CETINIA, Rey Juan Carlos University, Madrid, Spain

{alicia.pina, sara.garcia}@urjc.es

Keywords: Automatic Agents, Value-Aware Engineering, Natural Language Processing.

Abstract: As automatic agents begin to operate in high-stakes areas like finance and healthcare, the alignment of AI goals with human values becomes increasingly critical, addressing the so-called “alignment problem”. To tackle this challenge, the paper proposes the architecture of a Value-Based autonomous Agent capable of interpreting its environment through the lens of human values and guiding its decision-making processes in accordance with its own values. The agent utilizes a natural language processing (NLP) technique to detect and assess the values associated with various actions, selecting those most aligned with its moral guidelines. The integration of NLP into the agent’s architecture is crucial for enhancing its ability to make autonomous value-aligned decisions, offering a framework for incorporating ethical considerations into AI development.

1 INTRODUCTION

Artificial Intelligence (AI) systems are already well-integrated in our society. The scope of this technology spans from automated agents that manage home energy use or recommendation systems that help you make the perfect choice for your next movie night to autonomous vehicles, financial algorithms or medical diagnosis assistance.

This highly innovative field has grown rapidly in recent years and represent numerous advancements and benefits that are part of a major technological breakthrough. However, as autonomous agents begin to make decisions in high-risk domains, such as finance or healthcare, it becomes crucial to exercise caution.


One of the primary risks of AI, according to Stuart Russell, is that autonomous agents can inadvertently cause harm by pursuing goals that conflict with human values. This issue is often referred to as the “alignment problem” (Russell, 2022). For instance, an intelligent agent tasked with reducing pollution might shut down entire industries without considering the social and economic consequences. His view focuses on ensuring that AI systems are beneficial, controllable and aligned with human well-being. This perspective is closely related to the concept of value-


aware engineering, where the goal is to incorporate ethical and social considerations into the design and deployment of technology.

Within this framework arises the challenge of developing systems or intelligent agents capable of interpreting the environment in terms of human values—referred to as value-aware agents (Osman and d’Inverno, 2023). Once an agent can reason about values, it is crucial that it also acts according to its own moral or value guidelines, ensuring it can make value-aligned decisions.

This paper presents the architecture of an autonomous agent designed to guide its behavior based on human values. The agent can infer which values are promoted or demoted by a set of possible actions and select the one most aligned with its own values. To detect the values promoted or demoted by a particular option, an NLP technique is employed to extract human values from the text description of the option. This value-detection model consists of a pre-trained text analyser and a neural network to determine which values, ranked by their importance, are implicit in the descriptions of the actions. Additionally, an aggregation function is used to integrate the agent’s values into the decision-making process, determining the degree of alignment between the agent and each option.

The article’s content is structured as follows: Section 2 presents related work concerning value concepts and their computational extensions, along with a brief review of existing value detection techniques.

^a  <https://orcid.org/0009-0005-0412-4128>

^b  <https://orcid.org/0009-0001-4880-605X>

In Section 3, the proposed Value-Based Agent is explained, with a particular focus on the integration of the NLP model into the agent’s architecture. In Section 4, a real-world domain is introduced where the simulations will be conducted to observe the behavior of different agents. Finally, conclusions and future work are presented in Section 5.

2 STATE OF THE ART

A wide range of research across psychology, philosophy and social sciences agree that values guide human behaviour, playing a crucial role in human decision-making processes. Related works regarding the integration of human values in automatic agents decision-making schemes include research on value-based formal reasoning (VFR) frameworks (Wyner and Zurek, 2024), value-based argumentation frameworks (VAFR) (van der Weide et al., 2010) or the use of LLM to generate responses that align with human values (Abbo et al., 2024).

In state-of-the-art proposals, values are engineered into the decision-making architecture of autonomous agents as automatic behaviour, learned behaviour or through value-based reasoning (Noriega and Plaza, 2024). When considering the implementation of a value-based computational framework, it is essential to formally establish an explicit representation of human values and their relations. This involves defining sets of values and creating taxonomies. Among other theoretical frameworks that structure value concepts, the Moral Foundations Theory (MFT) (Haidt, 2013) proposes six fundamental moral values as universal across cultures: care, fairness, loyalty, authority, sanctity, and liberty. Another well established framework is the Basic Human Values (BHV) (Schwartz, 1992), also known as Schwartz’s Value Theory. This widely recognized theory, which explores values and the relationships between them, has been integrated into agents’ architectures in various ways (Heidari, 2022) (Karanik et al., 2024). In this article, Schwartz’s Value Theory serves as the foundation for the proposed value-based reasoning architecture, which will facilitate the implementation of a value-driven agent.

According to Schwartz, values are beliefs that relate to desirable end states or modes of conduct, which go beyond specific situations and guide the selection or evaluation of behaviour, people, and events. In BHV, Schwartz proposes ten fundamental values, based on the motivational goal they express: *self-direction*, *stimulation*, *hedonism*, *achievement*, *power*, *security*, *conformity*, *tradition*, *benevo-*

lence and *universalism*. In addition to identifying ten basic values, the theory explicates the structure of dynamic relations among them. One basis of this value structure is the idea that pursuing the promotion of a specific value will typically be congruent with fostering some values but will create conflict with others.

It is following this idea that he defines two bipolar dimensions in which the 10 basic values are classified. One refers to the emphasis of values on personal interests, or on the well-being of others: *social focus vs personal focus*. The second dimension captures the conflict between values that emphasize personal growth and exploration and those that focus on maintaining stability and preventing potential risks: *anxiety-free vs anxiety-based*. This classification leads to four main groups of values: *openness to change*, *conservation*, *self-enhancement* and *self-transcendence*.

The circular arrangement of the 10 basic values following the previous dimensions leads to a motivational continuum, as in Figure 1. Values located closer together on the circle are motivationally related, while those farther apart tend to be motivationally opposed.

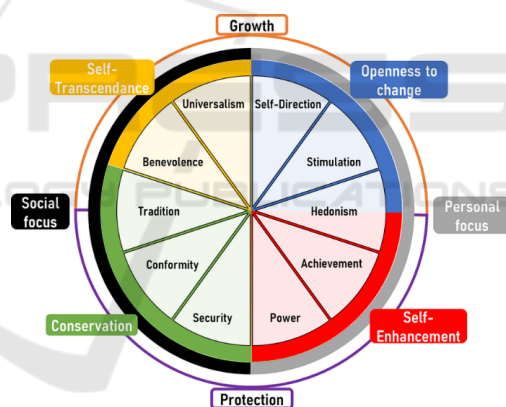


Figure 1: Basic Continuum (Schwartz, 1992).

In an extension of his theory (Schwartz et al., 2012), Schwartz refines the original 10 values into 19 to create a more comprehensive framework for understanding human motivations. This expansion allows for greater specificity in capturing diverse human experiences and highlights the complexity of value interactions across different contexts. The resulting extended continuum can be seen in Figure 2.

When making a value-aligned decision, individuals evaluate and select behavior that maximizes harmony with their values. The first step is to identify the degree to which an available action promotes or demotes each of the values (ex: SVT 19 values). The second step is to determine the degree of alignment

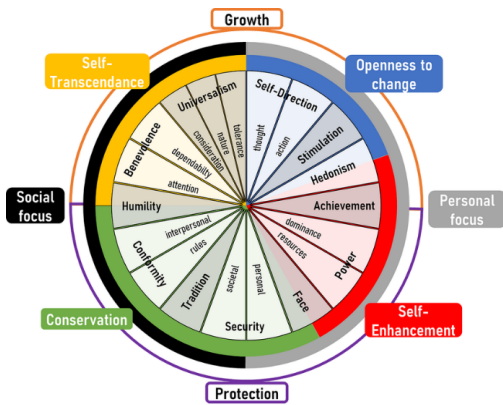


Figure 2: Extended Continuum (Schwartz et al., 2012).

of the individual with every possible choice. It is then that conflicts may arise, when trying to seek the promotion of all values. For example, one action might promote self-direction, while another action might prioritize security. In such cases, the individual must decide between the promotion of two different values. According to van der Weide (van der Weide et al., 2010), each person can determine the importance they give to each of their values, creating a personal Value System that guides their decision-making. This can help resolve conflict between values, by establishing an order of preference between them.

Given this two concepts, the promotion magnitudes that relate every possible action with the values and an individual’s own value system, the value alignment can be determined as a conjunction of both measures. Based on this concept of calculating alignment by combining or aggregating the two magnitudes, an automatic agent can be built so that it replicates the value-aligned decision-making process of an specific profile of individual (Karanik et al., 2024).

In this state-of-the-art Value-Based architecture, the agent receives a set of actions along with a pre-computed list indicating the values that each action promotes or demotes. This implies that the agent needs someone to interpret the set of actions in terms of values for them, which can be a major drawback when the goal is to construct an autonomous Value-Based Agent, as the agent depends on this value interpreter. The automatization of the value-elicitation process could enhance the agent’s autonomy while also eliminating the bias that a human annotator might introduce. If we consider the characterization of each action through a textual description that reflects the underlying motivations and objectives driving it, the value-extraction problem can be reframed as a value-detection in text problem.

This problem is still on ongoing challenge, with approaches that go from simple word-count-based

methods (Fulgoni et al., 2016) to feature-based methods utilizing word embeddings and sequences (Kennedy et al., 2021). The used methods can be classified in unsupervised or supervised. Some supervised methods are based on The Frame Axis technique (Hopp et al., 2020), that projects words onto micro-dimensions defined by two opposing sets of words to analyze their semantic orientation without labeled data, and others use the extended Moral Foundation Dictionary (MFD), which includes words associated with virtues, vices, and moral aspects related to the five dyads of Moral Foundations Theory (MFT) (Mokherian et al., 2020). Supervised techniques include performing multi-label classification, in which each label corresponds to a specific human value. The degree of association of the text with a category (human value) reflects the extent to which the text promotes that value. In this group it can be highlighted the use of NLP transformer-based models, such as XLNET or BERT (Bulla et al., 2024).

The following section explains an extension of the state-of-the-art Value-Based agent (Karanik et al., 2024), focusing on the integration of a NLP model into the agent’s architecture.

3 PROPOSED MODEL

The proposed model in this paper builds on the discussed Value-Based architecture concept. Along with the value-detection model incorporation, further contributions of this proposal include the use of the refined SVT with 19 values, instead of the basic 10-value theory used earlier or the inclusion of the concept of negative promotion (i.e., demotion) of values, which was absent in the base model, where only positive promotion was considered. This is an important upgrade, as it better reflects real-world scenarios where situations not only promote values positively but can also demote them. It also allows for the expression of disagreement with the possible actions, indicating a negative alignment with them. By considering both positive promotion and negative promotion (demotion) of values, the model more accurately captures the dynamic and sometimes conflicting nature of how values are influenced in real-life contexts. All in all, the result is a value-aware and value-aligned agent capable of perceiving its environment through the lens of values and acting accordingly.

The proposed architecture for an autonomous agent capable of making value-aligned decisions (see Figure 3) consists of two main components: the agent’s Value System, that represent its value preferences, and a Decision Module, which simulates a

value-based decision-making process. The central idea is that, given a set of options described in text, the agent’s Decision Module is able to make decisions guided by its Value System, and act in consequence.

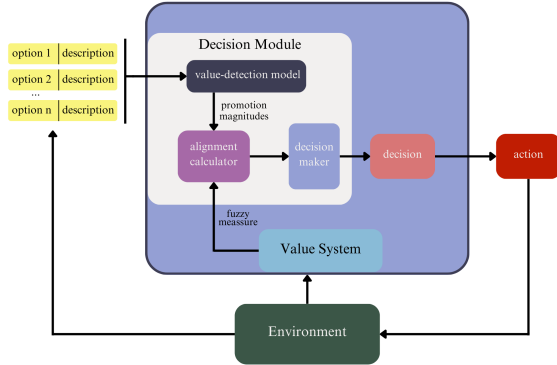


Figure 3: Value-aligned Agent model.

The Decision module includes an NLP-based value detection model that extracts the magnitudes of value promotion and demotion from the text descriptions of the actions. Based on these promotion and demotion magnitudes, along with the agent’s Value System, an alignment calculator computes the alignment magnitude for each option.

Once these magnitudes are calculated, the decision maker evaluates them to make a final decision. This could involve selecting the action with the highest alignment if only one action is possible, or determining whether to proceed with each potential action depending on the positivity or negativity of its alignment score. Finally, the agent will act according to the decision made.

Next, a more in-depth analysis of the various components will be presented.

3.1 Value System

As discussed in the previous section, the agent’s Value System is designed to represent its preferences concerning Schwartz’s 19 human values. These preferences can be captured by assigning importance magnitudes to each individual value. However, following the principles of Schwartz’s Value Theory (SVT), it is crucial to not only consider values in isolation but also the relationships and synergies between them. This approach leads to evaluating the importance of groups of values rather than solely focusing on individual ones.

In line with prior models (Karanik et al., 2024), a normalized fuzzy measure can be used to describe the importance weights assigned to different value groups. This allows for a more detailed representation, as it captures the interactions between values and

how they collectively influence the agent’s decision-making process.

In this extension of the model, it is crucial to impose a restriction on the values from the 19-refined-values set that are result from the disaggregation of one of the 10 Schwartz’s basic values. Specifically, the sum of these sub-values’ individual importances cannot exceed 1 (the maximum importance value allowed for their corresponding higher-level Schwartz value). This restriction is essential to preserve the monotonicity property of the constructed fuzzy measure, which could otherwise be violated.

The computation of the fuzzy measure (that is afterwards normalized) given a set of values is as follows:

$$\begin{aligned} \mathfrak{w}_i(\{v_s, \dots, v_t\}) &= \mathfrak{w}_i(\{v_s\}) + \dots + \mathfrak{w}_i(\{v_t\}) + \\ &+ \sum_{k=1}^{dp} ic(\{v_1, v_2\}_k) \times \mathfrak{w}_i(\{v_1\}) \times \mathfrak{w}_i(\{v_2\}), \end{aligned} \quad (1)$$

where dp is the number of distinct pairs within the set and ic is the interaction coefficient used to model the dynamic interaction of values. Following Schwartz, three main interaction between values are considered: (a) negative interaction between values in the same wedge. Likely, an agent who prefers one value will also prefer another of the same wedge, for example, *power* and *achievement*, and the weight of the importance of the group formed by both should be less than the sum of their single weights (subadditive measure); (b) positive interaction between values in opposite wedges. Due to it being unlikely that the agent would prefer both values of opposite wedges, such as *power* and *universalism*, the weight for this group should be greater than the sum of their single weights (superadditive measure) and (c) no interaction between values in adjacent wedges. Values belonging to multiple wedges, such as *hedonism*, *face* and *humility*, are considered to have a negative interaction with the values in the two wedges they are associated with and a positive interaction with the values in the other two wedges. The resulting expression of the interaction coefficient is

$$ic(\{v_1, v_2\}_k) = \begin{cases} +0.25 & v_1, v_2 \text{ in opposite wedges} \\ 0 & v_1, v_2 \text{ in adjacent wedges} \\ -0.25 & v_1, v_2 \text{ in the same wedge} \end{cases} \quad (2)$$

In this way, the fuzzy measure constructed based on the agent’s individual importance magnitudes over each value will effectively capture and represent the agent’s Value System.

3.2 Decision Module

The Decision Module enables the agent to make decisions aligned with its Value System. Firstly, the agent should detect the promotions and demotions of values for each possible action. Secondly, it needs to compute its alignment with each option, using the promotion magnitudes and the fuzzy measure that represents the importance the agent assigns to each subset of values. Lastly, given the collections of all the alignment magnitudes, the agent makes its final decision.

The value-detection model allows the agent to perceive how each action promotes or demotes distinct values. Given the text description of each available action, this model outputs the promotion magnitudes associated with each of them.

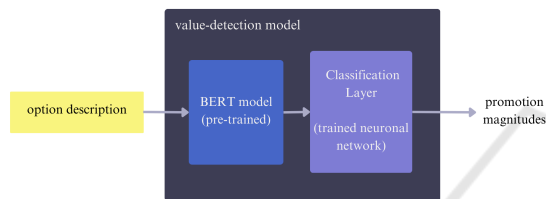


Figure 4: Value-detection model.

The proposed value-detection model, shown in Figure 4, is structured into two fundamental components: a pre-trained sub-model and a Classification Layer. The former refers to the well-known BERT (Bidirectional Encoder Representations from Transformers) model, while the latter consists of a neural network that serves as the Classification Layer. This layer allows to fine-tune the pre-trained model to our specific task and is trained using a labeled dataset (ValuesMLProject, 2024) consisting of text fragments and the corresponding promotion or demotion for each of the 19 Schwartz values.

The output of the model is a vector of promotion magnitudes for the input action, representing the promotion or demotion of each of the considered values.

Given the promotion magnitudes and the Value System, the alignment of the agent with each action is calculated using an aggregation function. The function selected in this paper is the signed Choquet's integral (Choquet, 1954). The positive Choquet's integral is computed considering the promoted values, while the negative Choquet's integral is calculated using only the demoted values. This two integrals represent the positive and negative alignment of the agent with the action, respectively. Then, the alignment is computed by subtracting the negative alignment from the positive alignment.

Once the alignment with each possible action is computed, the agent can make a decision, that ranges

from perceiving each action as desirable or undesirable to selecting the most aligned action.

4 CASE OF STUDY

4.1 Domain

Values not only guide human behaviour on an individual level, but they also shape human behaviour at a societal level. As several previous studies (Caprara et al., 2006) (Barnea and Schwartz, 2008) (Schwartz et al., 2010) have indicated, political voting is strongly related to and driven by personal values. This strong correlation between Schwartz's theory of values and political choices provides an ideal field of application for our current research.

It was demonstrated that voters' political choices in Western democracies depend more on personal preferences, especially values, than on other factors such as voters' social characteristics (Caprara et al., 2006). This study not only proved the primacy of values among the factors that drive voters' choices but also highlighted their lasting influence over time.

It has already been established that human decisions are driven by values, but the direct relationship between values and politics appears to be even more significant. Certain values are specifically related, both positively and negatively, to center-left and center-right political parties (Caprara et al., 2006), or more specifically to ideologies such as Classical Liberalism and Economic Egalitarianism (Barnea and Schwartz, 2008). Basic values are reflected in core political values (Schwartz et al., 2010), such as law and order, equality, or the acceptance of immigrants.

Taking into account these foundational studies that demonstrated the direct correlation between voters' values and their political choices, we propose to extend this research by considering not only Schwartz's 10 basic values but also his extended set of 19 values. One difficulty mentioned in previous studies on political psychology (Schwartz et al., 2010) was how to determine which values the political parties are promoting or demoting. The NLP model we proposed in Section 3 is the key element for overcoming this problem.

4.2 Simulations

The simulations consist of analyzing the political voting process of several Value-Based Agents with different value preferences. To do so, different profiles of agents are implemented following the proposed model, considering the relationship between

Schwartz’s values and their corresponding ideology as discussed in the prior subsection.

The voting process simulates the elections to the UK Parliament, in which the two major parties, the *Labour Party* and the *Conservative and Unionist Party*, present very strong ideologies (Social Democracy and Conservatism/Economic Liberalism).

In this voting scenario, agents are given two alternatives: the *Labour Party* or the *Conservative Party*. The agents will evaluate each party in terms of values, calculate their alignment with each of them and decide on the one that best aligns with their own values. To facilitate this process, two texts summarizing the ideology of each party (this is, the underlying motivation of each voting alternative) are considered. The promotion magnitudes extracted from these texts can be observed in Figure 5, and represent the values promoted and demoted by the two considered parties.

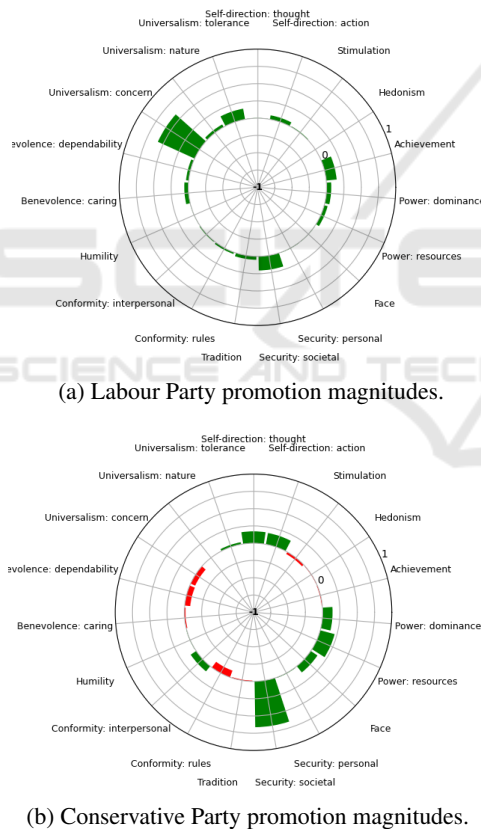
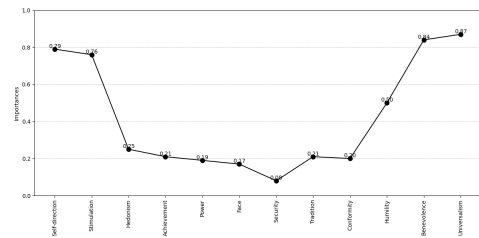


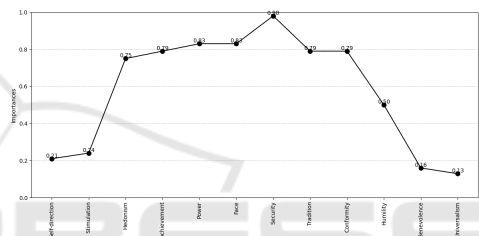
Figure 5: Promotion magnitudes detected.

As a starting point, the voting process is tested considering two agents with clearly defined ideological profiles: one representing a left-wing voter and the other representing a right-wing voter. The Value Systems of these agents are constructed based on Schwartz’s demonstrated correlations between human values and the political preferences of center-left vot-

ers. Assuming the opposite correlations for center-right voters, the fixed importances of each value for both agents are illustrated in Figure 6. Note that these importances are derived from the combination of sub-values that constitute each primal value, which are later broken down into their individual sub-values importances. The Value System of each agent is built according to this importances, computing the fuzzy measure described in Section 3.



(a) Left-wing voter importances.



(b) Right-wing voter importances.

Figure 6: Agent’s importances.

Each agent then computes its alignment with both parties, by aggregating the promotion magnitudes of each party (Figure 5) with the fuzzy measure that represents its Value System (Figure 6). The alignment magnitudes results can be seen in Table 1.

Table 1: Voter alignments.

	Labour Party	Conservative Party
Left-wing voter	0.0727	0.0265
Right-wing voter	0.0636	0.0924

Following this, two simulations are carried out, each considering a different number of agents and a distinct method for constructing their Value Systems.

For the first simulation, we generate a population of 200 agents, with 100 agents representing slight variations of the left-wing profile and 100 representing variations of the right-wing profile defined earlier. The Value Systems of these agents are constructed by making small modifications to the importance values of the ideological profiles. These modifications are introduced randomly, adjusting the importance $lw(\{v_i\})$ of each value i by up to $0.1 + 0.1 \cdot lw(\{v_i\})$.

The fuzzy measure is then constructed based on these modified ideological profiles. Having generated the population, the voting process is simulated.

To visualize the voting results (see Figure 7), each agent is positioned within Schwartz's two-dimensional value space according to its preference profile. As outlined in Section 2, this space is defined by two bipolar dimensions: the x-axis represents the continuum from *social focus* to *personal focus*, while the y-axis from *anxiety-based* (protection) to *anxiety-free* (growth). This two-dimensional division results in four quadrants, each of them corresponding to a wedge of Schwartz's continuum (Figure 2).

The positioning of each agent reflects its orientation along both dimensions, and therefore across each wedge, providing a clear graphical representation of their Value System orientation. Each agent is represented in red or blue, indicating their choice to vote for the *Labour Party* or the *Conservative Party*, respectively, based on their alignment with each party.

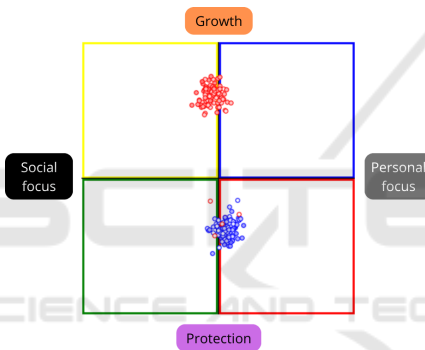


Figure 7: Random population voting results.

For the second simulation, the voting process is simulated for a population of 250 randomly generated agents.

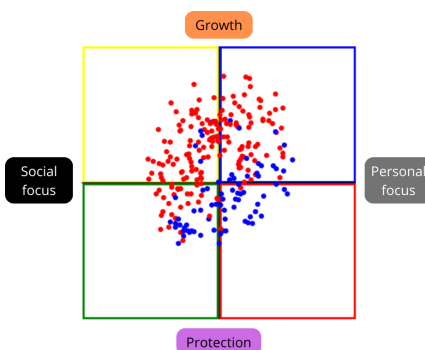


Figure 8: Random population voting results.

Instead of considering only agents with left-wing or right-wing profiles, each agent's preferences across the 19 values are randomly assigned, but always

adhering to the structure established by Schwartz. Specifically, the order of preferences follows the continuum of values (see Figure 2). The voting results are shown in Figure 7.

5 CONCLUSIONS AND FUTURE WORK

This paper proposes a model for a Value-Based Agent able to interpret a set of options in terms of values and make decisions based on its own values. By integrating an NLP model for value detection, the agent gains greater independence in its decision-making processes and can evaluate real-time situations based on relevant values, allowing it to respond dynamically to changing conditions and leading to more informed and effective decisions. Moreover, the inclusion of NLP within the agents architecture paves the way for future research, where each agent could develop its own value detection capabilities, creating different ways to perceive and interpret the environment.

The simulation results show the voting decisions made by agents with different preferences over values (i.e., different Value Systems). It can be observed (Table 1) that the two agents constructed with left and right-wing profiles are more aligned with the Labour Party (center-left) and the Conservative and Unionist Party (right-wing), respectively. Moreover, the agents whose profiles are generated as small variations of these profiles vote in 98 % of the cases in alignment with the profile from which they were generated. This behavior is consistent with Schwartz's value-based characterization of a left and right-wing voter. The graphical representation of the voting decisions of the population of agents with random value preferences (following Schwartz's restrictions) shows that agents with a tendency towards *anxiety-free* values, and more notably, those with values in the Self-Transcendence wedge, tend to vote for the *Labour Party*. In contrast, agents oriented towards protection or *anxiety-based* values, especially those with high preferences in the Self-Enhancement wedge, show a tendency to vote for the *Conservative Party*. These voting tendencies are appropriate given the values associated with each party's ideology. Moreover, the voting patterns are consistent with the agent's Value Systems, as agents with similar values tend to cast the same vote.

For future work, it is essential to explore improvements to the NLP model or even investigate alternative models, as the precision of the model is crucial for the agent's interpretation of the environment based on values. For instance, in the simulations conducted, it was observed that while the demotion of some values

was detected for the *Conservative party*, no negative promotion magnitudes were detected for the *Labour Party*, which results into a predisposition for agents with values not strongly associated with a specific ideology to vote for the *Labour Party* (see Figure 8). Refining the NLP model could lead to more reliable and accurate results, ultimately influencing the agents' decisions in a meaningful way.

Additionally, a future line of research could involve the integration of this Value-Based Agent architecture into an agent with practical applications, such as an intelligent traffic light or a chatbot.

ACKNOWLEDGEMENTS

This work has been supported by grant VAE: TED2021-131295B-C33 funded by MCIN/AEI/10.13039/501100011033 and by the "European Union NextGeneration EU/PRTR", by grant COSASS: PID2021-123673OB-C32 funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe", and by the AGROBOTS Project of Universidad Rey Juan Carlos funded by the Community of Madrid, Spain.

REFERENCES

- Abbo, G. A., Marchesi, S., Wykowska, A., and Belpaeme, T. (2024). Social value alignment in large language models. In Osman, N. and Steels, L., editors, *Value Engineering in Artificial Intelligence*, pages 83–97, Cham. Springer Nature Switzerland.
- Barnea, M. and Schwartz, S. (2008). Values and voting. *Political Psychology*, 19:17–40.
- Bulla, L., Gangemi, A., and Mongiovì, M. (2024). Do language models understand morality? towards a robust detection of moral content.
- Caprara, G., Schwartz, S., Capanna, C., Vecchione, M., and Barbaranelli, C. (2006). Personality and politics: Values, traits, and political choice. *Political Psychology*, 27:1–28.
- Choquet, G. (1954). Theory of capacities. *Annales de l'Institut Fourier*, 5:131–295.
- Fulgoni, D., Carpenter, J., Ungar, L., and Preoțiuc-Pietro, D. (2016). An empirical exploration of moral foundations theory in partisan news sources. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia. European Language Resources Association.
- Haidt, J. (2013). Moral psychology for the twenty-first century. *Journal of Moral Education*, 42.
- Heidari, S. (2022). *PhD Thesis: Agents with Social Norms and Values: A framework for agent based social simulations with social norms and personal values*.
- Hopp, F., Fisher, J., Cornell, D., Huskey, R., and Weber, R. (2020). The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text.
- Karanik, M., Billhardt, H., Fernández, A., and Ossowski, S. (2024). *Exploiting Value System Structure for Value-Aligned Decision-Making*, pages 180–196.
- Kennedy, B., Atari, M., Mostafazadeh Davani, A., Hoover, J., Omrani, A., Graham, J., and Dehghani, M. (2021). Moral concerns are differentially observable in language. *Cognition*, 212:104696.
- Mokhberian, N., Abeliuk, A., Cummings, P., and Lerman, K. (2020). *Moral Framing and Ideological Bias of News*, pages 206–219.
- Noriega, P. and Plaza, E. (2024). On autonomy, governance, and values: An agv approach to value engineering. In Osman, N. and Steels, L., editors, *Value Engineering in Artificial Intelligence*, pages 165–179, Cham. Springer Nature Switzerland.
- Osman, N. and d'Inverno, M. (2023). A computational framework of human values for ethical ai.
- Russell, S. (2022). *Artificial Intelligence and the Problem of Control*, pages 19–24.
- Schwartz, S., Caprara, G., and Vecchione, M. (2010). Basic personal values, core political values, and voting: A longitudinal analysis. *Political Psychology*, 31:421–452.
- Schwartz, S., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., Ramos, A., Verkasalo, M., Lönnqvist, J.-E., Demirutku, K., dirilen gumus, O., and Konty, M. (2012). Refining the theory of basic individual values. *Journal of Personality and Social Psychology*, 103:663–88.
- Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. volume 25 of *Advances in Experimental Social Psychology*, pages 1–65. Academic Press.
- ValuesMLProject (2024). Valuesml dataset. Data provided as tab-separated values files with one header line. In addition to the original files in nine languages, a machine-translated version in English is available.
- van der Weide, T. L., Dignum, F., Meyer, J. J. C., Prakken, H., and Vreeswijk, G. A. W. (2010). Practical reasoning using values. In McBurney, P., Rahwan, I., Parsons, S., and Maudet, N., editors, *Argumentation in Multi-Agent Systems*, pages 79–93, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Wyner, A. and Zurek, T. (2024). Towards a formalisation of motivated reasoning and the roots of conflict. In Osman, N. and Steels, L., editors, *Value Engineering in Artificial Intelligence*, pages 28–45, Cham. Springer Nature Switzerland.