Value-sensitive knowledge evolution

Adriana Luntraru

Master of Science in Informatics at Grenoble
Master Informatique, Université Grenoble Alpes
Specialization Data Science and Artificial Intelligence

Value-Sensitive Knowledge Evolution

Adriana LUNTRARU

July 7, 2023

Research project performed at LIG

Under the supervision of:
Jérôme EUZENAT

Defended before a jury composed of:
Massih-Reza AMINI, President
Sylvain BOUVERET, External Expert
Franck IUTZELER, Examiner

June 2023

Abstract

Cultural values are cognitive representations of general objectives, such as independence or mastery, that people use to distinguish whether something is "good" or "bad". More specifically, people may use their values to evaluate alternatives and pick the most compatible one. Cultural values have been previously used in artificial societies of agents with the purpose of replicating and predicting human behavior. However, to the best of our knowledge, they have never been used in the context of cultural knowledge evolution. We consider cooperating agents which adapt their individually learned ontologies by interacting with each other to agree. When two agents disagree during an interaction, one of them needs to adapt its ontology. We use the cultural values of independence, novelty, authority and mastery to influence the choice of which agent adapts in a population of agents sharing the same values. We investigate the effects the choice of cultural values has on the knowledge obtained. Our results show that agents do not improve the accuracy of their knowledge without using the mastery value. Under certain conditions, independence causes the agents to converge to successful interactions faster, and novelty increases knowledge diversity, but both effects come with a large reduction in accuracy. We however did not find any significant effects of authority.

Résumé

Les valeurs culturelles sont des représentations cognitives d'objectifs généraux, tels que l'indépendance ou le maîtrise, que les individus utilisent pour distinguer si quelque chose est "bon" ou "mauvais". Plus précisément, ils peuvent utiliser leurs valeurs pour évaluer des alternatives et choisir celle qui est la plus compatible avec leurs valeurs. Les valeurs culturelles ont déjá été utilisées dans des sociétés d'agents artificiels dans le but de reproduire et de prédire le comportement humain. Cependant, à notre connaissance, elles n'ont jamais été utilisées dans le contexte de l'évolution culturelle de la connaissance. Nous considérons des agents coopérants qui adaptent leurs ontologies individuellement apprises en interagissant les uns avec les autres dans le but d'atteindre un accord. Lorsqu'il y a désaccord entre deux agents pendant une interaction, l'un d'eux doit adapter son ontologie. Nous utilisons les valeurs culturelles d'indépendance, de nouveauté, d'autorité et de maîtrise pour influencer le choix de celui qui s'adapte dans une population d'agents partageant les mêmes valeurs. Nous étudions les effets du choix des valeurs culturelles sur les connaissances obtenues. Nos résultats montrent que les agents n'améliorent pas l'exactitude de leurs connaissances sans utiliser la valeur de maÃ®trise. Dans certaines conditions, l'indépendance permet aux agents de converger plus rapidement vers des interactions réussies, et la nouveauté augmente la diversité des connaissances, mais ces deux effets s'accompagnent d'une réduction importante de l'exactitude. En revanche, nous n'avons pas trouvé d'effets significatifs de l'autorité.

Acknowledgement

I would like to express my gratitude to my supervisor Jérôme Euzenat for his endless patience and invaluable advice throughout the duration of this research project. I would also like to thank INRIA for providing a wonderful work environment, and my colleagues from the mOeX team for supporting me and always offering help. This work was made possible by funding from the MIAI institute.

Notes

This report has been edited since the defense day on June 27, 2023. The edits are minor, fixing small mistakes and adding clarifications, with little impact on the overall content. The date on the first page denotes the last time this report was edited.

Contents

— 1 —
Introduction

1.1 Background: Cultural Evolution

Cultural evolution refers to the use of evolutionary theory in the context of culture [15]. It aims to understand the ways cultural traits are transmitted, selected and varied in human societies.

One component of culture that influences human behavior is cultural values [19, 7, 10]. Cultural values are mental representations of general objectives (e.g. independence, authority, security) which help people decide whether something is "good" or "bad". More specifically, people weigh different alternatives according to the values they hold and choose the one that is most compatible with their value system. For instance, if two people, belonging to a culture in which authority is a very important value, disagree on which decision to take, they will likely choose the decision proposed by the one which is in a position of higher authority (e.g. a student would follow their professor's decision, an employee their team leader's).

Cultural evolution can be explored using multi-agent simulations [22]. Human-inspired cultural values have been previously used in artificial societies of agents for different purposes [13, 14, 23]. However, they have not been used in the context of cultural knowledge evolution. The present work therefore aims to (a) explore mechanisms of using cultural values to guide agent behavior in the context of cultural knowledge evolution, and (b) analyze the consequences of the choice of cultural values.

1.2 Problem Statement: Designing Value-Sensitive Agents

Experiments of cultural knowledge evolution consist in a population of agents equipped with knowledge which play random interaction games [1, 4, 6, 3]. As a result of these interactions, agents may adapt their knowledge. By designing rules for the interaction games and observing the results, hypotheses may be tested.

The interaction games are designed to make agents use their knowledge to communicate to each other about their environment. Whenever the communication fails, the agents adapt their knowledge using adaptation operators. It has been shown that, under certain circumstances, knowledge adaptation leads to a faster convergence of the successful communication rate and improves the correctness of the knowledge [6, 3].

Agent knowledge is represented using ontologies. If agents obtain their initial ontologies independently through training on a subset of labeled objects from their environment, their ontologies might not be complete, correct or consistent with each other. When agents interact with each other, they might therefore not agree with each other.

When agents disagree, they need to adapt their ontologies. We believe the choice of how they adapt can be influenced using a cultural value system. We therefore consider value-sensitive agents as agents which are able to evaluate the compatibility of alternative adaptation actions with their cultural value system, and elect the most compatible one.

Our goal is to design value-sensitive agents in order to study if and how values can influence cultural knowledge evolution.

1.3 Scientific Approach, Methodology and Results

Our approach consists in first comparing different social theories of cultural values to choose one that best fits our problem. The Hofstede [7] and Schwartz [19] theories were both considered. We chose the Schwartz theory because of its robustness and high level of detail.

We defined the cultural value system of a population of artificial agents as relying on a set of cultural value indices, along with their associated importance in the population. Each value index is used to determine the compatibility of an adaptation action with respect to the cultural value. These compatibility values are then aggregated, along with their associated importance in the population, into a final score denoting the compatibility of an adaptation action with the cultural value system of the population.

For this purpose, we chose four value indices, inspired by Schwartz: independence, novelty, authority and mastery. Independence favors knowledge held by agents which have a higher level of disagreement with other agents. Novelty supports adaptation actions that result in the highest amount of change in knowledge. Authority promotes knowledge held by agents which adapt the least often. Finally, mastery favors the knowledge of the agents which are the best at correctly classifying objects in their environment.

We are interested in the effects of the cultural value system on the following three measures: the quality and diversity of knowledge and the success rate.

We therefore formulated one hypotheses for each value index:

(H1) The mastery index is needed for improving knowledge correctness. A negative mastery index would cause the accuracy to converge to 0.

(H2) Using a negative independence weight results in a higher value and faster convergence of the success rate.

(H3) Using a negative novelty weight increases the knowledge diversity.

(H4) Using a positive authority weight alongside a positive mastery weight increases knowledge correctness.

To test these hypotheses, we considered homogeneous populations of agents, all sharing the same cultural value system. We then ran a full factorial experiment, varying the weights associated with each value index, and recording the three measures at each iteration.

To assess the statistical significance of each value index on each measure, we performed one-way ANOVA and Tukey HSD tests. Our results show the first hypothesis to be supported.

The second and third hypotheses are supported under certain circumstances, but they lead to low knowledge quality. Lastly, the fourth hypothesis is not supported.

1.4 Contents

Chapter 2 presents the state-of-the-art of cultural values in sociology, as well as previous work on cultural knowledge evolution in multi-agent systems. Chapter 3 then reconciles the two by proposing the design of value-sensitive agents. Chapter 4 proposes an experiment plan for simulating the behavior of the aforementioned agents. It also defines the measures and methodology, and formulates the hypotheses. The hypotheses are tested by analyzing the results of the experiments in Chapter 5. Chapter 6 then draws the final conclusions and offers perspectives on future work.

— 2 —
State-of-the-Art

2.1 Cultural Values

Culture is composed of unwritten rules which affect people's behavior and help differentiate members of different groups (e.g. citizens of different countries) [9]. Based on surveys conducted on IBM workers in 53 countries and regions, the four Hofstede cultural dimensions were empirically identified: power distance, collectivism vs individualism, femininity vs masculinity and uncertainty avoidance [8]. Two more dimensions were later added: pragmatism and indulgence [7].

Cultural values are "broad goals that serve as guiding principles in people's lives" [19, 21]. Schwartz values are cognitive representation of goals needed to cope with the three universal requirements of human survival: a) biological needs of individuals, b) needs for coordinated social interaction (preserving the social fabric), and c) survival and welfare of groups. People use values to decide whether something is "good" or "bad". More specifically, they use the respective importance of each of the values they hold to compare alternative choices and pick the "best" one.

Schwartz values are based on social theory [16, 12], and empirically tested through surveys in close to 100 countries. According to the Schwartz theory of values [19], there are a total of 10 basic value types which are present across cultures: self-direction, stimulation, hedonism, achievement, power, security, conformity, tradition, benevolence and universalism. They can be organized through three principles: a) compatibility between the values (congruence or conflict), b) the interests served by holding a value (personal or social), and c) the relation of values to anxiety (self-protection or growth).

Figure 2.1 shows the 10 basic value types, the 4 higher order values and the 3 organizing principles used in the Schwartz Value Survey [19, 21, 18]. The first organizing principle, value compatibility, is represented by the value position on the wheel. The closest two values are on the wheel, the more compatible they are. Conversely, the furthest apart two values are, the more conflicting they are.

The other two organizing principles are represented through the two outermost circles. Concerning the interests served, the values on the left half of the wheel have a social focus, while the values on the right half have a personal focus. Regarding the relation to anxiety, the values on the top half of the circle promote growth, while the values on the bottom half promote self-protection.

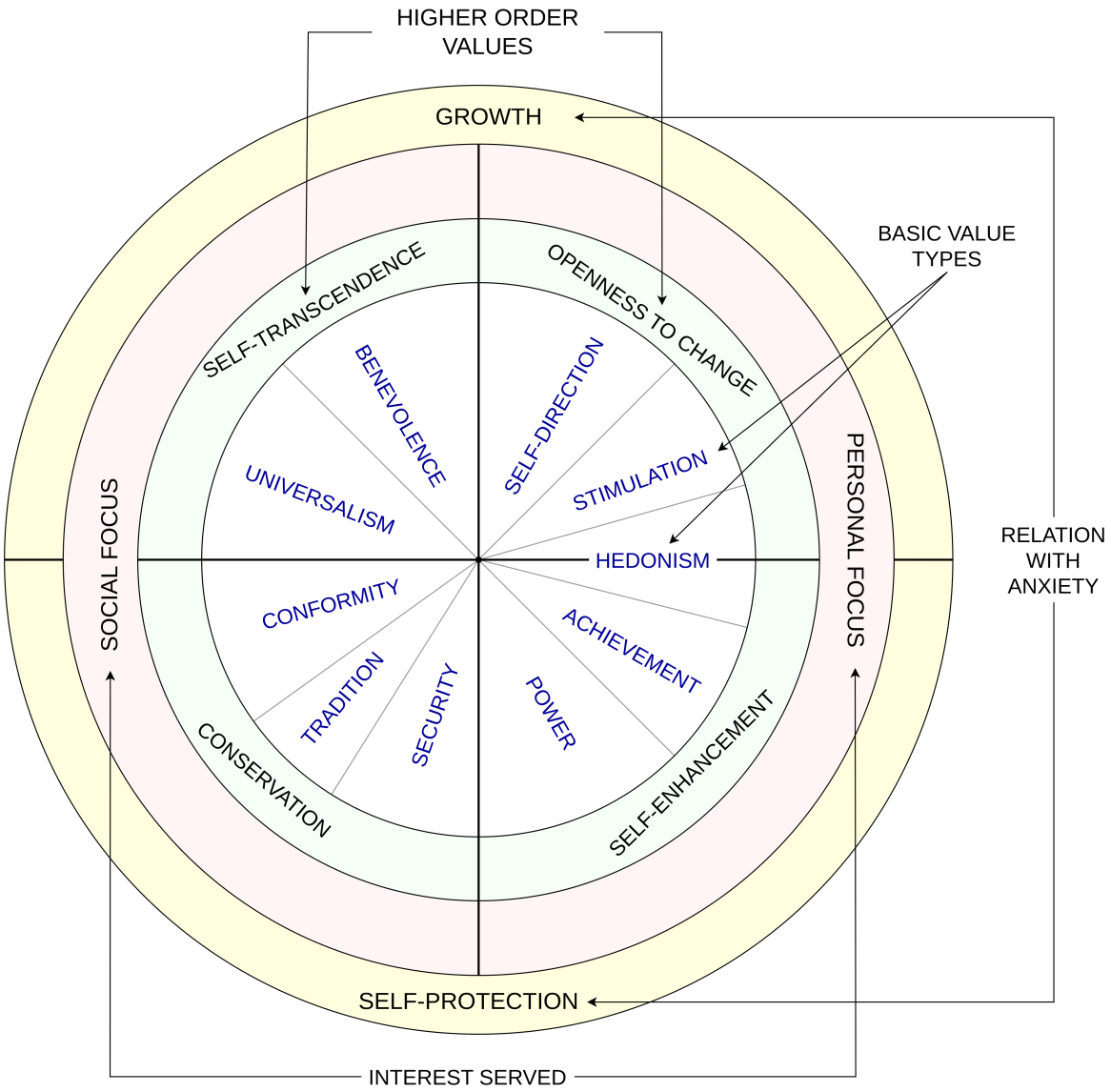

Figure 2.1: Schwartz Values Wheel illustrating the values and their organizing principles. Figure adapted from [21] and [18].

While Hofstede cultural dimensions were identified empirically and Schwartz cultural values were based on social theory, there is a significant overlap between them [20], namely (a) Hofstede's individualism/collectivism and Schwartz's autonomy/embeddedness, (b) Hofstede's power distance and Schwartz's egalitarianism/hierarchy, and (c) Hofstede's masculinity and Schwartz's mastery.

Value systems based on either Schwartz or Hofstede definitions have been previously used in multi-agent systems for different purposes. A Hofstede-inspired value system has been designed and used for flexible agent coordination [23]. Agents holding the values of "wealth" and "fairness", inspired by the Schwartz theory, were used to successfully replicate human behavior in the ultimatum game [14]. The Hofstede dimensions of individualism/collectivism and indulgence have been used to improve the prediction of people's privacy decisions across

cultures [13]. The results were successfully replicated using the equivalent Schwartz values of embeddedness and egalitarianism.

Our goal is to exploit cultural values in the context of artificial cultural knowledge evolution.

2.2 Cultural Evolution and Multi-Agent Systems

Cultural evolution is concerned with understanding the transmission, selection and variation of cultural traits within human societies [15]. Building upon the foundations of biological evolution, it examines the complex dynamics of cultural change and adaptation.

Cultural evolution can be explored using multi-agent simulations, focusing on the dynamics of knowledge acquisition, transformation and transmission of knowledge within a population of interacting agents. One of its applications is the exploration of cultural knowledge evolution in the context of cooperation and coordination among agents.

Ontologies can be used in multi-agent systems to represent agent knowledge. When simulating agent interaction games, agents might be in a position in which their knowledge is different, but they need to come to an agreement. For achieving agent agreement, one approach is to have agents define correspondences between their concepts and the concepts of other agents. This can be represented using a network of alignments between agents' ontologies [5, 1, 6, 2]. When a pair of agents disagrees, they may change the alignment between their ontologies, using adaptation operators.

A different approach is to allow agents to adapt their own ontology when faced with an interaction failure [3, 24, 11]. Informally, it can be described as agents changing their mind (adapting their knowledge), rather than trying to understand other agents' knowledge by making correlations to their own. The next section goes deeper into explaining the second approach, on which the current work is based.

2.3 Multi-Agent Systems with Interaction-Based Adaptation

The present work builds on the foundations laid by [3] in designing multi-agent systems in which agents evolve their knowledge through interaction. In [3], the environment contains a set of objects I, each being characterized by binary properties $p \in P$ (Fig. 2.2).

Let D be the set of decisions and $d^* : I \to D$ the oracle mapping each object to one correct decision. Since the decisions are disjoint, they can be represented as an ontology O^* (Fig. 2.3).

A set of agents A is placed within this environment. Each agent has access to the public ontology O^* (the set of decisions it can take when presented with an object) and to the properties of each object. However, the agents do not know which decision is the correct one for a given object.

Each agent $a \in A$ has its own private ontology O^a for representing its knowledge, which is learned independently. The learning phase occurs only once, in the beginning of the experiment, before the interaction games are played. During this phase, each agent is presented with a (possibly different) random subset of objects I and their corresponding correct decision. Each agent $a \in A$ uses the ID3 algorithm [17] to learn a decision tree classifier, which is then converted into its private ontology O^a and correspondences between its leaf nodes and decision classes $d \in D$ from the public ontology O^* (Fig. 2.4).

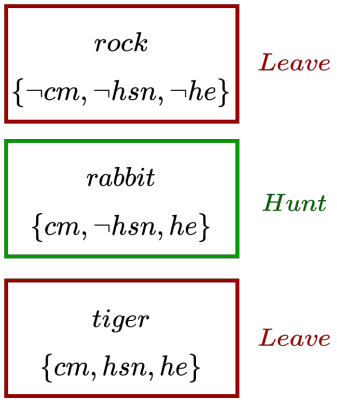

Figure 2.2: Objects (*rock, rabbit, tiger*) with binary properties (*cm = can move, hsn = has sharp nails, he = has eyes*) and their corresponding correct decisions (*Leave, Hunt*). Figure adapted from [3].

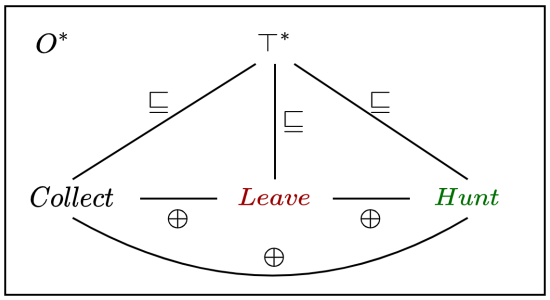

Figure 2.3: Decision ontology O^* with 3 disjoint decision classes (*Collect, Leave, Hunt*). Figure adapted from [3].

The agents then play n interaction games, each game being described as follows:

(i) Two random agents a, b and a random object o are picked from the environment.

(ii) The two agents disclose their decision classes d_a and d_b about object o.

(iii) If $d_a = d_b$, the interaction ends in success.

(iv) Otherwise, the interaction is a failure and one of the agents adapts its ontology.

In the event of an interaction failure between agent a and agent b about object o, if agent b is the one chosen to adapt, the adaption can be described as follows:

(i) Agents a and b disclose the most specific (leaf) classes C_a, C_b of their respective ontologies O^a, O^b to which the object o belongs, along with the corresponding decisions D_a, D_b.

(ii) If there are no object classified as C_b but not classified as C_a ($C_b \sqsubseteq C_a$), replace the correspondance $\langle C_b, \sqsubseteq, D_b \rangle$ with $\langle C_b, \sqsubseteq, D_a \rangle$.

(iii) Otherwise, add $\langle (C_b \sqcap \neg C_a), \sqsubseteq, D_b \rangle$ and replace $\langle C_b, \sqsubseteq, D_b \rangle$ by $\langle (C_b \sqcap C_a), \sqsubseteq, D_a \rangle$ (Fig. 2.5).

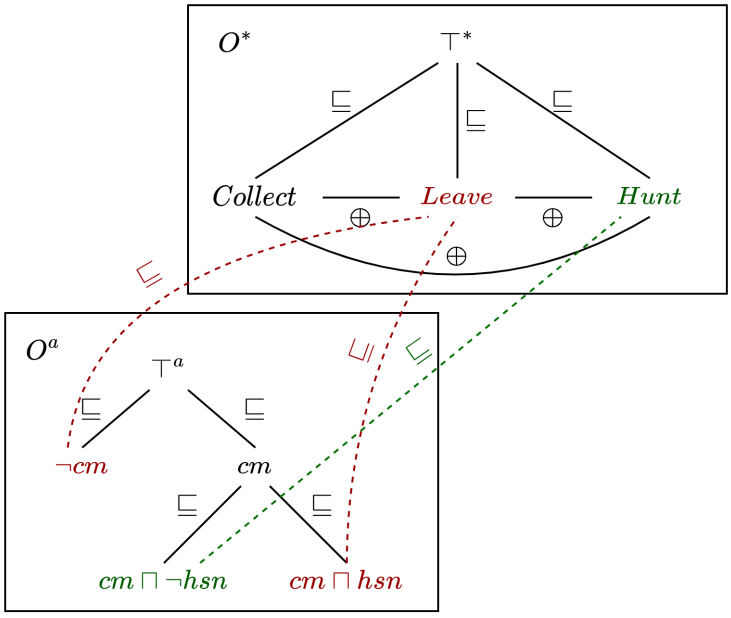

Figure 2.4: Agent a's ontology O^a with correspondences between its most specific classes and decision classes in O^*. Figure adapted from [3].

This adaptation mechanism allows the agents to only change the part of their knowledge that causes the current disagreement, while keeping the rest of their knowledge intact.

The choice of which agent will adapt is explored in [3], and later in [24]. In [3], this choice is solely based on feedback from the environment: at each iteration, before interacting, the agents perform an individual task in which they classify a subset of objects from the environment $S \subset I$ according to their decisions. As a result, the agent gets a payoff denoting the proportion of correctly classified objects. When the interaction ends in failure, the agent which has the lower payoff is the one which will adapt.

In addition to the environment feedback, [24] also introduces a "conformity" bias to help determine which agent adapts. This bias takes into account the frequency of agreement in each agent's previous interactions. [24] has shown that a) the communication success rate does not converge without using any transmission bias, and b) the accuracy of the knowledge does not improve without using feedback from the environment.

2.4 Conclusion

Our goal is to explore the impact cultural values can have on multi-agent cultural knowledge evolution. We will therefore propose ways to integrate cultural values within the previously described experimental framework.

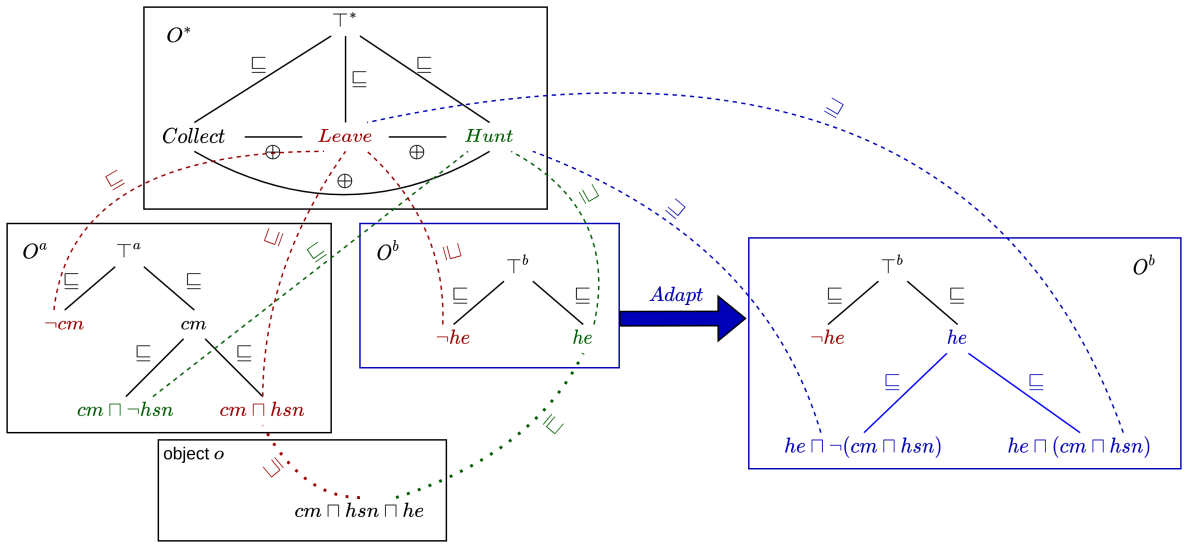

Figure 2.5: Interaction failure between agent a and agent b about object o, resulting in agent b adapting its ontology O^b. Changes caused by the adaptation are colored in blue.

— 3 —
Design of Value-Sensitive Agents

Our goal is to design mechanisms of influencing agents' behavior using cultural values, and then analyze their impact on cultural knowledge evolution. To achieve this, we will propose a design which will be integrated within the *Lazy lavender* [25] experimental framework, described in Section 2.3.

We define value-sensitive agents as agents which, in the event of an interaction failure, can use their cultural values to make the choice of which one of them adapts their ontology. To reach value-sensitive agents, we design a Cultural Value System and define means of using it to guide the agents in making the aforementioned choice.

To design our Cultural Value System, we choose a social value theory to base it on. Then, we decide which cultural values to use, define ways to model them to fit our system, and define the relationships between them.

3.1 Cultural Value System

There are two principal value theories that have both been extensively researched and previously used in multi-agent systems: the Schwartz value theory [19, 21] and the Hofstede cultural dimensions [7, 9]. We chose to base our Cultural Value System on Schwartz values for several reasons: a) the robustness of his theory, b) its high level of detail, allowing us to choose values as finely grained as we wish, and c) the relationships of compatibility/conflict between the values.

However, it must be noted that our goal is not predicting human behavior, and our cultural value system is merely *inspired by* Schwartz values. Some liberties were therefore taken in the design.

For the purpose of this work, we use the 4 higher order values defined in the Schwartz Value Theory [21]: Openness to Change, Self-enhancement, Conservation and Self-transcendence. In order to keep the model simple, we make the following design choices:

(i) Openness to Change and Conservation are perfectly negatively correlated;

(ii) Self-enhancement and Self-transcendence are perfectly negatively correlated;

(iii) There is no correlation between neighboring values.

The first two choices allow us to keep the design simple, while the last one allows us to keep the values independent in order to accurately identify their impact. We can therefore consider

two cultural dimensions: a) Openness to change - Conservation and b) Self-transcendence - Self-enhancement. For each cultural dimension, we choose two indices: the *independence* and *novelty* indices for the Openness to change - Conservation dimension and the *authority* and *mastery* indices for the Self-transcendence - Self-enhancement dimension (Fig. 3.1). These indices will be used to influence agent behavior, more specifically to allow agents to evaluate and compare alternative actions. They are formally defined in Section 3.2.

The reasons behind choosing two values per cultural dimension is a better coverage of the basic value types (8 out of 10). Hedonism and tradition were left out because we did not find a way to make our agents benefit from them.

Tradition was not used because of its overlap with conformity [21]. In our design, we chose to exploit the subordination (obedience) component of conformity, which is also part of the tradition value. We therefore decided not to separate the two.

We did not exploit hedonism because it refers to self-indulgence and seeking pleasures for oneself, which goes beyond our goals.

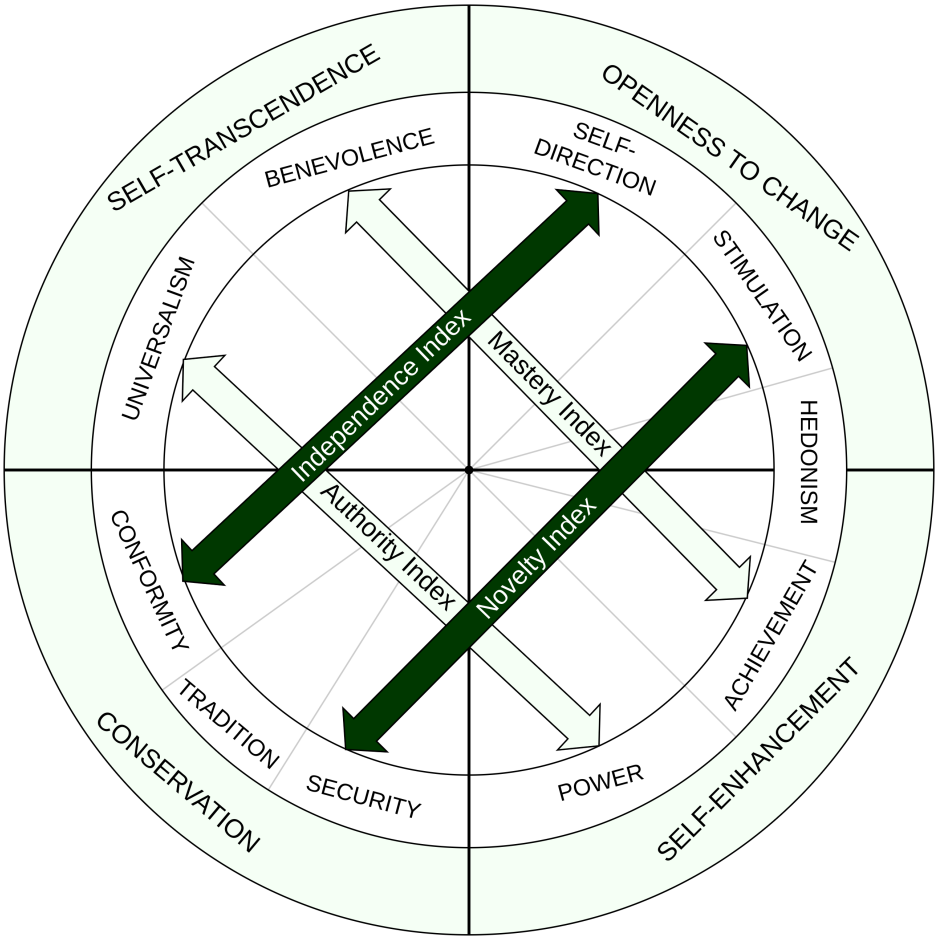

Figure 3.1: The 4 value indices used for the Cultural Value System: independence, novelty, authority and mastery.

We define an adaptation action $\langle a, b, n \rangle$ as agent b adapting its ontology to agree with agent a at iteration n. We use cultural values to measure the degree of compatibility between an adaptation action and the cultural value system adopted by the agents. In our design, the cultural

values are global and constant at the population level, meaning that all agents which belong to the same population share the same cultural values and the cultural values do not change over time. In our experiments, we only have one population of agents, which means that they all share the same cultural values.

3.2 Action Compatibility with Values

In the case of an interaction failure between agent a and agent b at iteration n, there are two alternative adaptation actions to choose from:

- $\langle a,b,n\rangle$ - agent b adapts its ontology to agree with agent a, or
- $\langle b,a,n\rangle$ - agent a adapts its ontology to agree with agent b.

We want this choice to be compatible with the cultural values of the agents' populations. Hence, these two alternatives will be evaluated with respect to the indices enabling us to assess their compatibility with the population's values.

We will first define the four individual cultural value indices (independence, novelty, authority and mastery) and ways of measuring an action's compatibility with each of them. We will then propose a way to aggregate these measures into a final score representing an action's compatibility with a population's cultural values.

The naming of the cultural value indices is meant to indicate the cultural values they are based on from the Schwartz theory.

3.2.1 Independence Index

The independence index relates to the opposing values of self-direction and conformity. People who value self-direction treasure freedom, choosing their own goals, and taking actions independently, while people who value conformity appreciate obedience and restraint from actions which might violate the social norms [19, 21].

We consider agents to be more independent when they disagree during their interactions and less independent (or more obedient) when they agree. The compatibility of an action $\langle a,b,n\rangle$ with the independence value will therefore be computed as the disagreement rate over the previous interactions of the agent which does not adapt:

$$c^{ind}_{a,b,n} = \frac{\sum_{i=1}^{n} \gamma_{ind}^{n-i} \times dagr_a^i}{\sum_{i=1}^{n} \gamma_{ind}^{n-i}}$$

where $\gamma_{ind} \in (0,1]$ is a discount factor, $dagr_a^i$ is the disagreement of agent a at iteration i:

$$dagr_a^i = \begin{cases} 0, & \text{if the interaction was a success (ending in agreement)} \\ 1, & \text{otherwise} \end{cases}$$

and i only iterates over the interactions that agent a participated in.

A population of agents which puts a high importance on independence will prefer the action in which the more obedient agent adapts to agree with the more independent one. Informally, this can be seen as the population supporting independence by letting the independent agents take the decisions.

3.2.2 Novelty Index

The novelty index corresponds to the conflict between stimulation and security. People who value stimulation are daring, seeking new and exciting things. On the other hand, people who value security are moderate and want to preserve social order [19, 21].

We consider an adaptation action to be more novel if it results in a big change in the ontology of the agent which adapts. Conversely, we consider an action to be less novel (or more secure) if it results in a small change in the adapting agent's ontology.

When agents a and b play the n^{th} interaction game about an object o and disagree, they disclose the leaf classes C_a^n, C_b^n from their respective ontologies to which object o belongs (Section 2.3). The compatibility of an action $\langle a, b, n \rangle$ with the novelty value is defined as the ratio of objects that would have their decision changed in agent b's ontology:

$$c_{a,b,n}^{nov} = \frac{|C_a^n \sqcap C_b^n|}{|C_b^n|}$$

In the example provided in Fig. 3.2, agents a and b interact about object o_2. Agent a classifies object o_2 in the class $C_a \equiv cm \sqcap hsn$, while agent b classifies it as $C_b \equiv he$. The corresponding decisions, $d_a(o_2) = Leave$ and $d_b(o_2) = Hunt$ are different, leading to an interaction failure (disagreement).

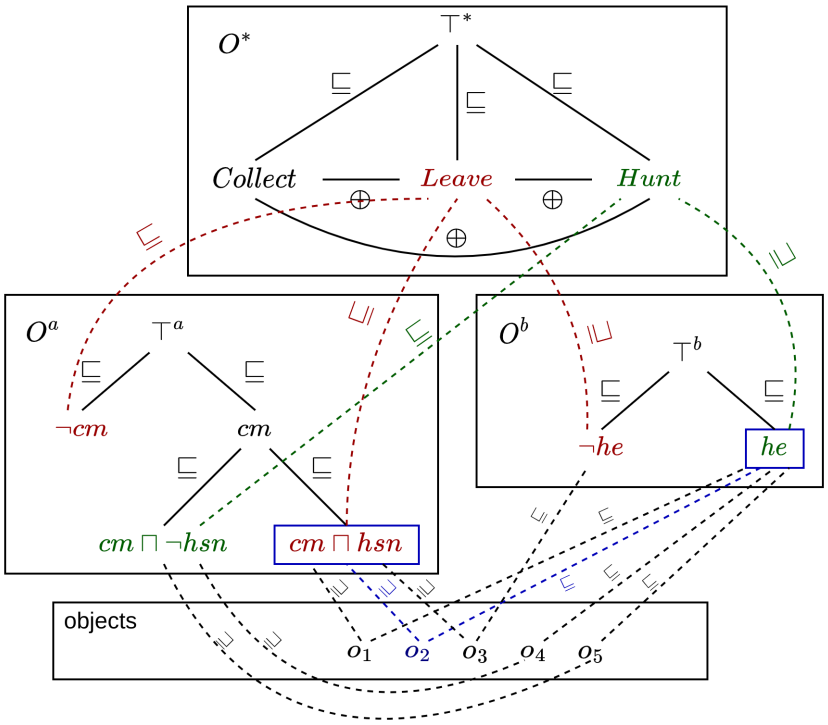

Figure 3.2: Interaction between agents a and b about object o_2, illustrating the concept of novelty.

In the environment, there are 3 objects (o_1, o_2, o_3) classified by agent a with the class $C_a \equiv cm \sqcap hsn$, and 4 objects (o_1, o_2, o_4, o_5) classified by agent b as $C_b \equiv he$. There are a total of 2 objects (o_1, o_2) which can be classified as $C_a \sqcap C_b \equiv (cm \sqcap hsn) \sqcap he$.

If action $\langle b,a,n\rangle$ was elected, it would result in reclassifying $c^{nov}_{b,a,n} = \frac{2}{3}$ objects. Otherwise, if action $\langle a,b,n\rangle$ was elected, it would result in reclassifying $c^{nov}_{a,b,n} = \frac{2}{4}$ objects. Therefore, the action $\langle b,a,n\rangle$ is more compatible with the value of novelty.

3.2.3 Authority Index

The authority index is based on the opposing values of power and universalism. People who value power want to be the ones who make the decisions and want others to follow them. Universalism, on the other hand, promotes equality, harmony, conflict avoidance and listening to people who hold different opinions [19, 21].

We consider agents to be more authoritative when they do not adapt during disagreements, and less authoritative (or more harmonious) when they do. The compatibility of an action $\langle a,b,n\rangle$ with the authority value will therefore be defined as follows:

$$c^{auth}_{a,b,n} = \frac{\sum_{i=1}^{n} \gamma_{auth}^{n-i} \times win_a^i}{\sum_{i=1}^{n} \gamma_{auth}^{n-i}}$$

where $\gamma_{auth} \in (0,1]$ is a discount factor, win_a^i is the winning value of agent a at iteration i:

$$win_a^i = \begin{cases} 1, & \text{if interaction } i \text{ was a failure (disagreement) and agent } a \text{ was not the one to adapt} \\ 0.5, & \text{if interaction } i \text{ was a success (agreement)} \\ 0, & \text{if interaction } i \text{ was a failure (disagreement) and agent } a \text{ was the one to adapt} \end{cases}$$

and i only iterates over the interactions that agent a participated in.

3.2.4 Mastery Index

The mastery index corresponds to the opposing values of achievement and benevolence. Achievement promotes intelligence, success and mastery, while benevolence promotes forgiveness and harmony [19, 21].

We define masterful agents as agents which are good at correctly classifying objects in their environment. In order to measure their accuracy, before each interaction game, each agent plays an individual task from which it receives a reward. During the individual task, each agent is presented with a random subset of objects from the environment $S_i \subset I$ and asked to provide its classification for each of them. The reward received by agent a at iteration i is:

$$r_a^i = \frac{|\{o \in S_i \mid d_a^i(o) = d^*(o)\}|}{|S_i|}$$

where $d_a^i : I \to D$ is agent a's function for mapping objects to decisions based on its knowledge at iteration i, and $d^* : I \to D$ is the oracle mapping objects to the correct decisions. The task ratio t is a parameter of the experiment, such that $|S_i| = t \times |I|$.

The compatibility of an action $\langle a,b,n\rangle$ with the mastery index is then computed as:

$$c^{mast}_{a,b,n} = \frac{\sum_{i=1}^{n} \gamma_{mast}^{n-i} \times r_a^i}{\sum_{i=1}^{n} \gamma_{mast}^{n-i}}$$

where $\gamma_{mast} \in (0,1]$ is a discount factor.

3.2.5 Aggregated Value Compatibility Score

In order to compare the two alternative actions, the indices above will be aggregated into a single score. The compatibility score of the action $\langle a,b,n\rangle$, defined as agent b adapting to agree with agent a at iteration n, is the sum of the four indices (independence, novelty, authority, mastery) weighted by their importance in the adapting agent's population:

$$c_{a,b,n} = \sum_{v \in V} w_b^v \times c_{a,b,n}^v$$

where $V = \{ind, nov, auth, mast\}$, $c_{a,b,n}^v \in [0,1]$ and $w_b^v \in [-1,1]$ is the weight representing the importance of value v in agent b's population.

The weights may be positive or negative, in order to represent the opposing values on the Schwartz value wheel (Fig. 3.1). For example, a population with $w^{ind} = 1$ will prefer adaptation actions that support independence, while a population with $w^{ind} = -1$ will prefer actions that support obedience.

The compatibility scores for the two alternative actions $\langle a,b,n\rangle$ and $\langle b,a,n\rangle$ are computed and compared:

- If $c_{a,b,n} > c_{b,a,n}$, agent b will adapt its ontology to agree with agent a.
- If $c_{a,b,n} < c_{b,a,n}$, agent a will adapt its ontology to agree with agent b.
- Otherwise, the adapting agent will be chosen randomly.

3.3 Design Conclusions

Due to the overlap between Schwartz and Hofstede values [20], it may be argued that our cultural values conform to certain Hofstede dimensions as well as with Schwartz values. The independence and novelty indices relate to Hofstede's individualism/collectivism dimension, the authority index to Hofstede's power distance, and the mastery index to Hofstede's masculinity. We reiterate the fact that we do not aim to precisely replicate and/or predict human behavior. Therefore, designing our cultural value system using Hofstede dimensions would have also been a valid choice.

We have now defined a cultural value system and designed agents which behave according to it. In the following two chapters, we will propose simulation experiments using these agents. Our final goal will be to analyze the impact of the choice of cultural values.

— 4 —
Experiment

Having defined the cultural value system and the mechanisms which allow agents to benefit from it, the following step is experimentation. We will start by describing the experimental process. Then, we will define measures and an experimental plan. Finally, we will discuss our methodology and formulate hypotheses.

4.1 Experiment Description

In the beginning of an experiment run, each agent is presented with a (possibly different) subset of objects from the environment $S \subset I$, along with their properties and their corresponding correct decisions. Each agent learns their initial ontologies by training on this labeled data using the ID3 algorithm [17]. We set the training ratio $t = \frac{|S|}{|I|}$ to 0.2.

Then, the following steps are performed for n iterations (Section 2.3):

(i) Two random agents $a, b \in A$ are selected.

(ii) Agents a and b each play one individual classification task and receive the rewards r_a, r_b (Section 3.2.4).

(iii) Agents a and b play an interaction game together about a random object $o \in I$ from the environment.

(iv) If the interaction ends in failure, one of the agents adapts.

For deciding which agent adapts, the value compatibility score is computed for the two alternative actions $\langle a, b, n \rangle$ and $\langle b, a, n \rangle$ (Section 3.2.5). Then, the choice is made as follows:

- If $c_{a,b,n} > c_{b,a,n}$, agent b adapts to agree with agent a.
- If $c_{b,a,n} > c_{a,b,n}$, agent a adapts to agree with agent b.
- Otherwise, either agent a or agent b is randomly chosen to adapt.

At the end of each iteration, 3 measures are computed and saved, as described in the following section.

4.2 Measures

The measures we take have already been defined and used within the *Lazy lavender* framework [3, 24], namely: success rate, accuracy and distance.

The success rate of the experiment e at iteration n is defined as the ratio of previous successful interactions (interactions in which the agents agreed):

$$srate(e_n) = \frac{\sum_{i=0}^{n} s(e_i)}{n}$$

where

$$s(e_i) = \begin{cases} 0, & \text{if interaction } i \text{ was a failure (disagreement)} \\ 1, & \text{if interaction } i \text{ was a success (agreement)} \end{cases}$$

The accuracy of the agents' ontologies at iteration n of the experiment e is an average of the accuracy of each individual agent's ontology:

$$accuracy(e_n) = \frac{\sum_{a \in A} acc(O_n^a)}{|A|}$$

where A is the set of all agents in the experiment and O_n^a is agent a's ontology at iteration n.

The accuracy of agent a's ontology at iteration n is the ratio of objects correctly classified by agent a:

$$acc(O_n^a) = \frac{|\{o \in I \mid d_n^a(o) = d^*(o)\}|}{|I|}$$

where I is the set of all objects in the environment, and d_n^a, d^* the decision functions of agent a and the oracle, respectively.

The ontology distance at iteration n of experiment e is defined as the average of the distance between each distinct pair of agents:

$$distance(e_n) = \frac{\sum_{a \in A} \sum_{b \in A \setminus \{a\}} dist(O_n^a, O_n^b)}{|A| \times (|A| - 1)}$$

The distance between two ontologies O^a and O^b is computed as follows:

$$dist(O^a, O^b) = 1 - \frac{equiv(O^a, O^b)}{max(|O^a|, |O^b|)}$$

where $equiv(O^a, O^b)$ is the number of equivalent classes in the ontologies O^a, O^b, meaning the pairs of classes which classify the same set of objects:

$$equiv(O^a, O^b) = |\{(C^a, C^b) \in O^a \times O^b \mid O^a, O^b \models C^a \equiv C^b\}|$$

This last measure is used as an indication of the knowledge diversity.

4.3 Experiment Plan

Experiments were run in *Lazy lavender* [25], using a full factorial plan, using only one population of agents and varying the parameters relating to the population's cultural values. All other parameters were used with fixed values. Table 4.1 shows the values of the fixed parameters, while table 4.2 shows the values of the varied parameters. We do not vary all parameters for practical reasons, namely the computational time. We are only interested in assessing the impact of the parameters used to influence one population of agents through cultural values. We have therefore made informed choices on the values of all the other parameters.

4.3.1 Fixed Parameters

The impact of the number of agents, number of properties, number of decisions, training ratio, task ratio, training algorithm, and adaptation operator has been previously assessed [24].

Meaning	**Variable**	**Values**
Number of agents	$\|A\|$	20
Number of properties	$\|P\|$	4
Number of decisions	$\|D\|$	3
Training ratio	r	0.2
Task ratio	t	0.2
Training algorithm	*trainer*	*ID3*
Adaptation frequency	*adFreq*	0.05
Independence discount	γ_{ind}	0.9
Authority discount	γ_{auth}	0.9
Mastery discount	γ_{mast}	0.9
Adaptation operator	*op*	*allCom*
Number of iterations	n	100000
Number of runs	*nbRuns*	5

Table 4.1: Fixed experiment parameters and their values.

The adaptation operators defined in the *Lazy lavender* framework control how many properties agents disclose from the class in their ontology with which they classified the object they disagree about [24]. The choice of adaptation operator is not statistically significant for accuracy and distance, but it is for success rate. For our experiments, we choose the adaptation operator *allCom*, which forces agents to disclose all the properties of the class in which they classify the object. This choice was made with respect to the novelty index (Section 3.2.2), which computes a reclassification ratio and therefore profits from the full description of classes, but also because *allCom* yields a higher success rate compared to the other operators.

All the discount factors γ_{ind}, γ_{auth}, γ_{mast} were set to 0.9 in order to put more emphasis on the most recent interactions of the agents. The novelty index is only concerned with the current interaction, therefore it does not need a discount factor.

When two agents interact and disagree, one of them adapts. Therefore, at the end of the interaction the two agents are in agreement with each other. We chose to make the agents adapt in only 5% of their failed interactions. This allows the independence index to better reflect

the current agreement rate of each agent with the rest of the agents [24]. Since this causes the agents to converge more slowly, we set the number of iterations to 100000. Each experiment is run 5 times.

The algorithm used in the initial training proved to not have a statistically significant effect on any of the three measures [24], therefore we chose the ID3 algorithm [17] simply because it is the fastest one available in the *Lazy lavender* framework [25].

Since the values of independence and mastery have already been used as transmission biases in a previous experiment [24], we chose the same values for the number of agents, properties and decisions to facilitate comparisons.

4.3.2 Cultural Values Parameters

The parameters relating to the population's cultural values are the weights corresponding to the four value indices: independence, novelty, authority and mastery (Table 4.2). In our experiment, all agents belong to the same population.

Meaning	**Variable**	**Values**
Independence index weight	w_{ind}	$\{-1, 0, 1\}$
Novelty index weight	w_{nov}	$\{-1, 0, 1\}$
Authority index weight	w_{auth}	$\{-1, 0, 1\}$
Mastery index weight	w_{mast}	$\{-1, 0, 1\}$

Table 4.2: Varied experiment parameters and their values.

We used both positive and negative weights for each index to indicate the importance of two opposing cultural values using one index (Section 3.2):

- w_{ind} corresponds to self-direction vs. conformity
- w_{nov} corresponds to stimulation vs. security
- w_{auth} corresponds to power vs. universalism
- w_{mast} corresponds to achievement vs. benevolence

For instance, $w_{ind} = 1$ denotes a population which supports actions compatible with self-direction (independence), while $w_{ind} = -1$ denotes a population which supports actions compatible with conformity (obedience).

We ran a full factorial experiment with the parameters in Table 4.1 and Table 4.2. The $3^4 = 81$ combinations of parameters were each run 5 times over $100,000$ iterations, resulting in $81 \times 5 = 405$ total runs, and $40,500,000$ total iterations.

4.4 Methodology and Hypotheses

In order to assess the effects of the cultural value indices on our measures, we perform one-way ANOVA (analysis of variance) tests using our value indices as independent variables and the measures at the final iteration as dependent variables. We exclusively study the effects of

each independent variable alone. For a given pair of dependent and independent variables, ANOVA computes the probability (*p-value*) of the independent variable having no effect on the dependent one. We consider a *p-value* lower than 0.01 sufficient for rejecting the null hypothesis that the independent variable has no effect on the dependent one.

ANOVA only informs on whether an independent variable has a statistically significant effect on a dependent one, but we are also interested in knowing how it is affected. Therefore, in addition to ANOVA, we also run post-hoc Tukey HSD (honestly significant difference) tests. While the following chapter only covers the more interesting results, the complete analysis is publicly available in the experiment logbook [27]. The result dataset can be found at [26].

Before formulating our hypotheses, we reiterate the results found in previous experiments [24], which we expect to be verified by our experiments:

(R1) If the adaptation action is chosen randomly, or if the independence value is used alone, the accuracy does not improve.

(R2) The success rate converges faster and to a higher value for an independence weight of -1.

Based on (R1), we do not expect to be able to improve the quality of the agents' knowledge without any feedback from the environment, i.e. the rewards the agents receive during the individual tasks (Section 3.2.4).

Based on (R2), we expect an "obedient" population of agents to converge to a high interaction success rate faster because it favors the decisions taken by the agents which agree more often.

A population of agents which discourages novelty will prefer adaptation actions that result in the smallest changes in the agents' ontologies (Section 3.2.2). We believe this will help preserve more of the initial knowledge, leading to a higher final knowledge diversity.

An agent population which supports authority will favor the knowledge of the agents which adapted less often in their previous interactions (Section 3.2.3). We think that authority, when used together with mastery, can enhance its positive effect on knowledge accuracy.

Considering all of the above, we formulate our hypotheses as follows:

(H1) The accuracy does not improve without using a positive mastery index weight and using a negative mastery index weight results in the accuracy converging to 0.

(H2) A negative independence weight will cause the success rate to converge faster to a higher value.

(H3) A negative novelty weight will lead to a high final ontology distance.

(H4) A positive authority weight will lead to an increase in final accuracy when used together with a positive mastery weight.

Next, we will analyze the results of the formerly described experiments, test our hypotheses, and eventually propose ideas for further exploitation of our cultural value system.

— 5 —
Result Analysis

Our goal is to find the effects of our cultural value system (Section 3.2) on the previously defined measures (Section 4.2). More precisely, we want to see how different weights for the four value indices (independence, novelty, authority and mastery) affect the accuracy and diversity of the agents' ontologies and the success rate of the interactions. For this purpose, we defined an experiment plan (Section 4.3) and formulated four hypotheses (Section 4.4).

The objective of the experiment is to analyze the results, which allows the testing of the hypotheses. To assess the impact of each individual index, we will first start by interpreting the results obtained when the value indices are used on their own.

5.1 Using One Index on Its Own

To analyze the effects of using one cultural value index on its own, we are considering the results in which only one weight has a non-zero value. Table 5.1 shows the three measures for each index weight. As a point of reference, the last column displays the data from the runs in which none of the indices are used, i.e. $w_v = 0$, $\forall v \in V$. As expected, choosing the adaptation actions randomly leads to a very low final accuracy and success rate.

Additionally, Table 5.2 shows the ANOVA results for each index. On each column, ANOVA is applied on a different subset of the results, namely the results obtained when the weight of one index has a given value and all other indices have the value 0.

When looking at the measures for mastery, along with their plots (Figure 5.1), we observe that the positive mastery weight seems to make the agents converge to a very high accuracy (0.97) within $60,000$ iterations. None of the other value indices, when used on its own, leads to a high accuracy. In most cases, the accuracy even decreases over time and sometimes it does not converge after $100,000$ iterations (i.e. negative novelty weight, positive authority weight).

A mastery weight value of -1 causes the accuracy to converge to 0. Together with the findings mentioned above, this confirms our first hypothesis (H1).

The negative novelty weight encourages diversity (i.e. it leads to the highest final distance between ontologies), confirming our third hypothesis (H3). The resulting final accuracy is however very low.

Fig. 5.2 seems to support the second hypothesis (H2): the independence weight of -1 causes the success rate to converge faster and to a higher value.

Since the accuracy does not improve without a positive mastery weight, we will next look into using a mastery weight of 1 along with one other index.

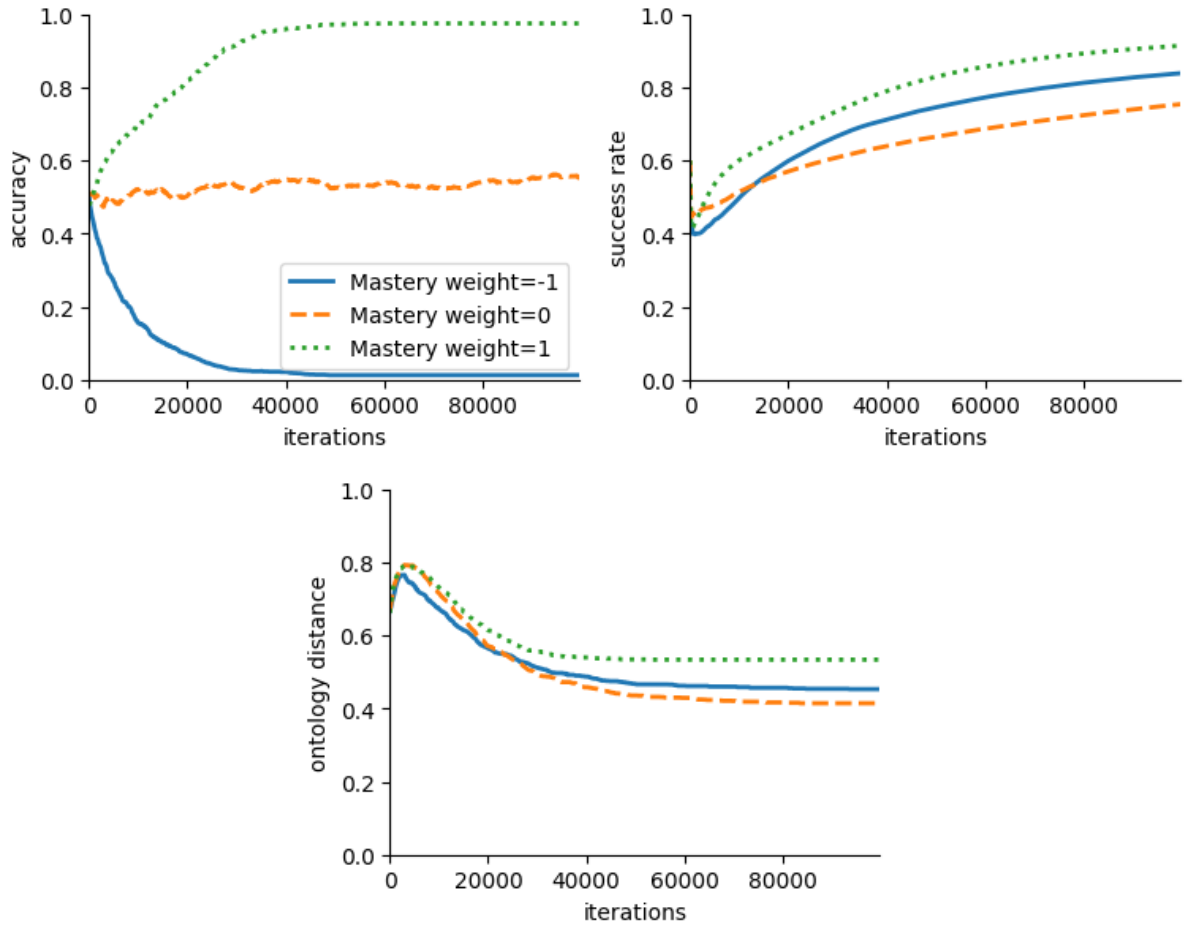

Figure 5.1: Ontology accuracy (top left), success rate (top right) and ontology distance (bottom) when the mastery index is used on its own. The plotted lines for mastery weight $= 0$ (orange dashed) correspond to the runs in which no index is used (all weights are zero).

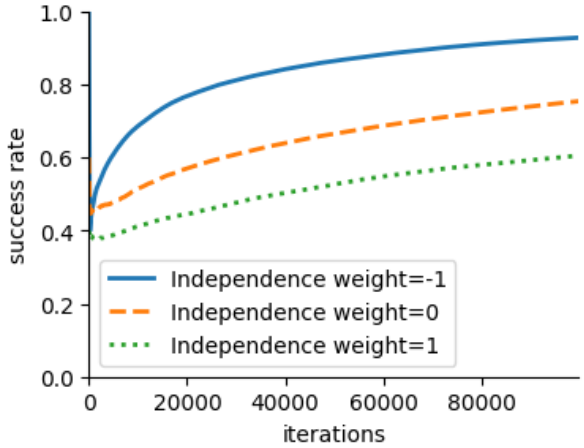

Figure 5.2: Success rate while the independence index is used on its own. The plotted line for independence weight $= 0$ (orange dashed) corresponds to the runs in which no index is used (all weights are zero).

Independence weight		-1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Novelty weight		0.0	0.0	-1.0	1.0	0.0	0.0	0.0	0.0	0.0
Authority weight		0.0	0.0	0.0	0.0	-1.0	1.0	0.0	0.0	0.0
Mastery weight		0.0	0.0	0.0	0.0	0.0	0.0	-1.0	1.0	0.0
accuracy	1	0.46	0.46	0.46	0.45	0.49	0.46	0.50	0.48	0.51
	20000	0.53	0.35	0.39	0.43	0.50	0.45	0.07	0.82	0.50
	40000	0.52	0.38	0.39	0.43	0.48	0.45	0.02	0.96	0.55
	60000	0.53	0.40	0.39	0.42	0.49	0.44	0.01	**0.97**	0.54
	80000	0.54	0.39	0.39	0.43	0.47	0.44	0.01	**0.97**	0.54
	100000	0.54	0.39	0.39	0.42	0.46	0.44	0.01	**0.97**	0.55
ssrate	1	1.00	0.40	0.00	0.20	0.20	0.40	0.40	0.60	0.60
	20000	0.77	0.44	0.82	0.55	0.54	0.57	0.60	0.67	0.57
	40000	0.84	0.50	0.91	0.63	0.61	0.65	0.71	0.79	0.64
	60000	0.88	0.55	0.94	0.69	0.67	0.70	0.77	0.86	0.69
	80000	0.91	0.58	**0.96**	0.74	0.72	0.73	0.81	0.89	0.72
	100000	0.93	0.61	**0.96**	0.77	0.75	0.76	0.84	0.91	0.75
distance	1	0.62	0.66	0.66	0.66	0.66	0.67	0.67	0.66	0.67
	20000	0.62	0.45	**0.75**	0.37	0.55	0.55	0.56	0.61	0.57
	40000	0.58	0.35	**0.75**	0.33	0.44	0.44	0.49	0.54	0.46
	60000	0.58	0.32	**0.75**	0.32	0.42	0.42	0.46	0.53	0.43
	80000	0.58	0.32	**0.75**	0.32	0.42	0.42	0.46	0.53	0.42
	100000	0.58	0.32	**0.75**	0.32	0.41	0.42	0.45	0.53	0.41

Table 5.1: Accuracy, success rate and distance while using only one cultural value index. Last column shows the measures while not using any index. Highest values for each measure in bold.

	Independence weight		Novelty weight		Authority weight		Mastery weight	
	p-value	Infl	p-value	Infl	p-value	Infl	p-value	Infl
accuracy	0.113955	False	0.067150	False	0.411979	False	0.000000	**True**
ssrate	0.000002	**True**	0.000004	**True**	0.899969	False	0.000043	**True**
distance	0.000511	**True**	0.000000	**True**	0.997836	False	0.004629	**True**

Table 5.2: ANOVA results for using only one cultural value index, showing whether the index weight has an influence on each measure.

5.2 Using the Mastery Index with One Other Index

Table 5.3 shows the measures taken at different iterations for the experiments runs in which one non-zero index weight is used alongside a mastery weight of 1. For comparison, the last column shows the data for the runs in which a mastery weight of 1 is used on its own. Additionally, Table 5.4 displays the ANOVA results for the effects of each index weight on each measure.

Independence weight		-1.0	1.0	0.0	0.0	0.0	0.0	0.0
Novelty weight		0.0	0.0	-1.0	1.0	0.0	0.0	0.0
Authority weight		0.0	0.0	0.0	0.0	-1.0	1.0	0.0
Mastery weight		1.0	1.0	1.0	1.0	1.0	1.0	1.0
accuracy	1	0.42	0.46	0.46	0.42	0.43	0.49	0.48
	20000	0.62	0.72	0.47	0.63	0.68	0.79	0.82
	40000	0.72	0.83	0.47	0.83	0.82	0.91	0.96
	60000	0.75	0.91	0.47	0.88	0.87	0.94	**0.97**
	80000	0.75	0.94	0.47	0.88	0.89	0.94	**0.97**
	100000	0.76	0.95	0.47	0.88	0.89	0.94	**0.97**
ssrate	1	0.00	0.20	0.60	0.20	0.00	0.40	0.60
	20000	0.66	0.56	0.88	0.61	0.60	0.71	0.67
	40000	0.77	0.66	0.94	0.72	0.71	0.80	0.79
	60000	0.83	0.73	0.96	0.80	0.79	0.86	0.86
	80000	0.87	0.79	0.97	0.85	0.84	0.89	0.89
	100000	0.90	0.82	**0.98**	0.88	0.87	0.92	0.91
distance	1	0.68	0.64	0.64	0.66	0.65	0.66	0.66
	20000	0.61	0.53	**0.73**	0.39	0.53	0.63	0.61
	40000	0.55	0.45	**0.73**	0.36	0.44	0.54	0.54
	60000	0.54	0.43	**0.73**	0.36	0.44	0.54	0.53
	80000	0.54	0.43	**0.73**	0.36	0.44	0.54	0.53
	100000	0.54	0.43	**0.73**	0.36	0.44	0.54	0.53

Table 5.3: Accuracy, success rate and distance while using a mastery weight value of 1 together with one other index. The last column shows the measures when the mastery index is used on its own. Highest values for each measure in bold.

	Independence weight		Novelty weight		Authority weight	
	p-value	Infl	p-value	Infl	p-value	Infl
accuracy	0.000459	**True**	0.000001	**True**	0.090169	False
ssrate	0.008515	**True**	0.000075	**True**	0.046613	False
distance	0.034418	False	0.000000	**True**	0.005096	**True**

Table 5.4: ANOVA results for using one value index together with a mastery weight of 1.

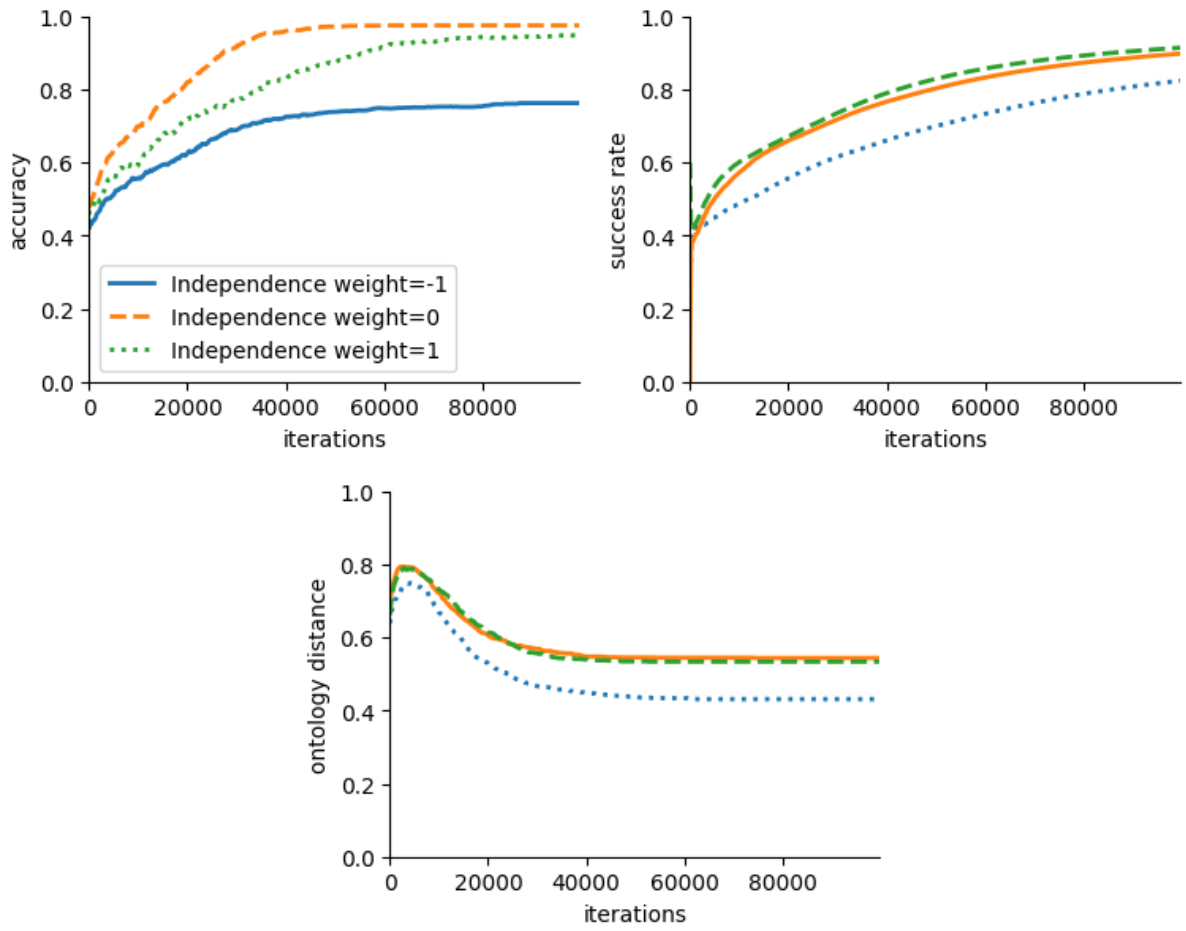

Figure 5.3: Ontology accuracy (top left), success rate (top right) and ontology distance (bottom) when the independence index is used along with a mastery weight value of 1. The plotted lines for the independence weight of 0 (orange dashed) correspond to the runs in which a mastery weight value of 1 is used on its own.

5.2.1 Independence

Taking a look at the plotted measures for different independence weights (Fig. 5.3), it seems to yield a similar final accuracy to using the mastery index on its own, only taking longer to converge. Although the ANOVA results (Table 5.4) show that independence influences accuracy, the Tukey HSD test shows that the independence weight of 1 does not have a statistically significant influence on accuracy ($p > 0.01$). The same applies to the success rate: according to the Tukey HSD results, an independence weight of -1 is not significant for the success rate, which does not support our second hypothesis (H2).

5.2.2 Novelty

The ANOVA results (Table 5.4) show that novelty has an influence on accuracy, but the Tukey HSD test shows this to be true only for a novelty weight of -1, which has a negative impact (Fig. 5.4). The positive influence of a novelty weight of -1 on the ontology distance does hold

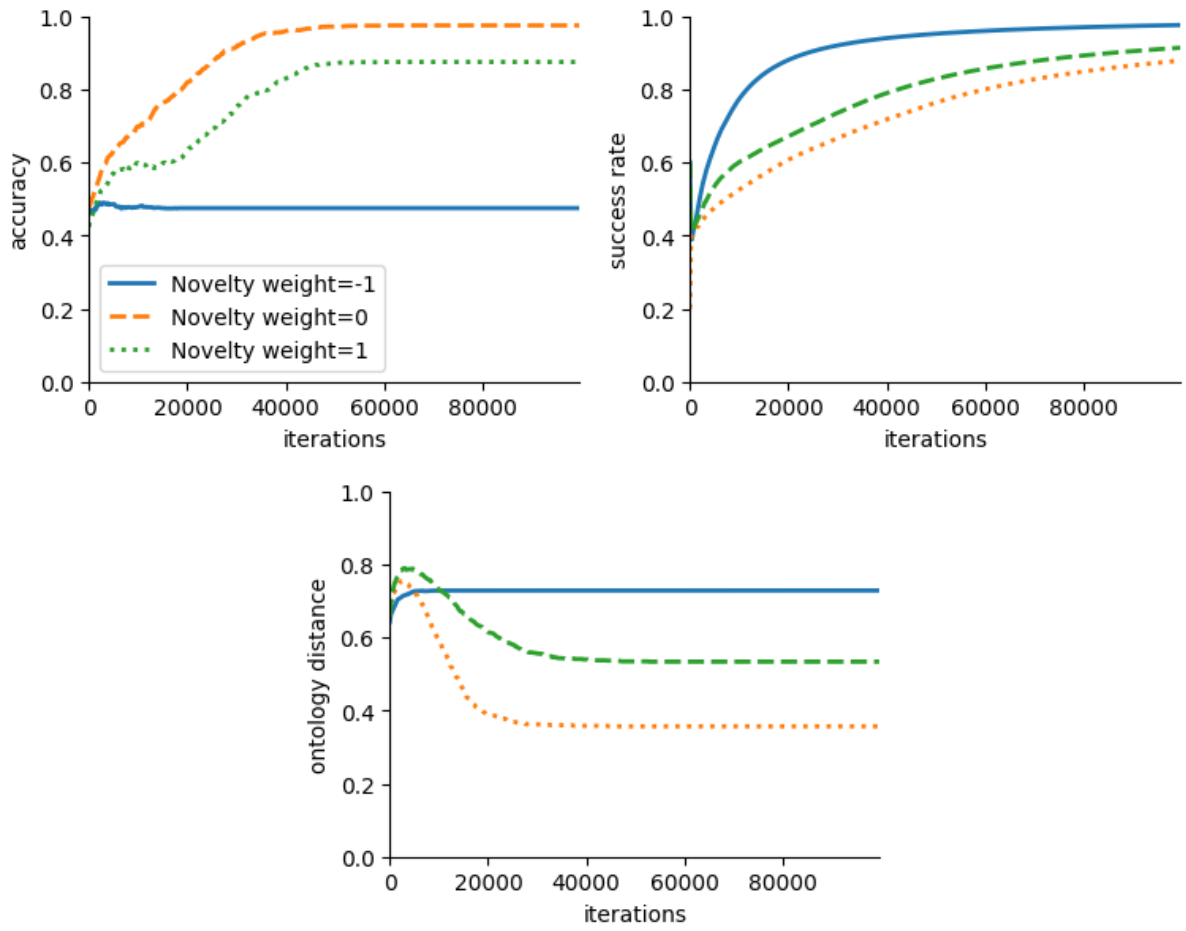

Figure 5.4: Ontology accuracy (top left), success rate (top right) and ontology distance (bottom) when the novelty index is used along with a mastery weight value of 1. The plotted lines for the novelty weight of 0 (orange dashed) correspond to the runs in which a mastery weight value of 1 is used on its own.

true according to both tests, reinforcing our third hypothesis (H3). The negative impact this has on accuracy is however very large.

5.2.3 Authority

Surprisingly, when used jointly with mastery, the authority index seems to decrease accuracy, either with weights -1 or 1 (Table 5.3). However, these results are not statistically significant (Table 5.4). Hence, hypothesis (H4) is not supported.

The Tukey HSD test shows the influence on distance to be significant only for the weight of -1, which produces the smallest ontology distance (Fig. 5.5), denoting a low knowledge diversity.

5.3 Result Conclusions

After testing all our hypotheses, we arrived to the following conclusions:

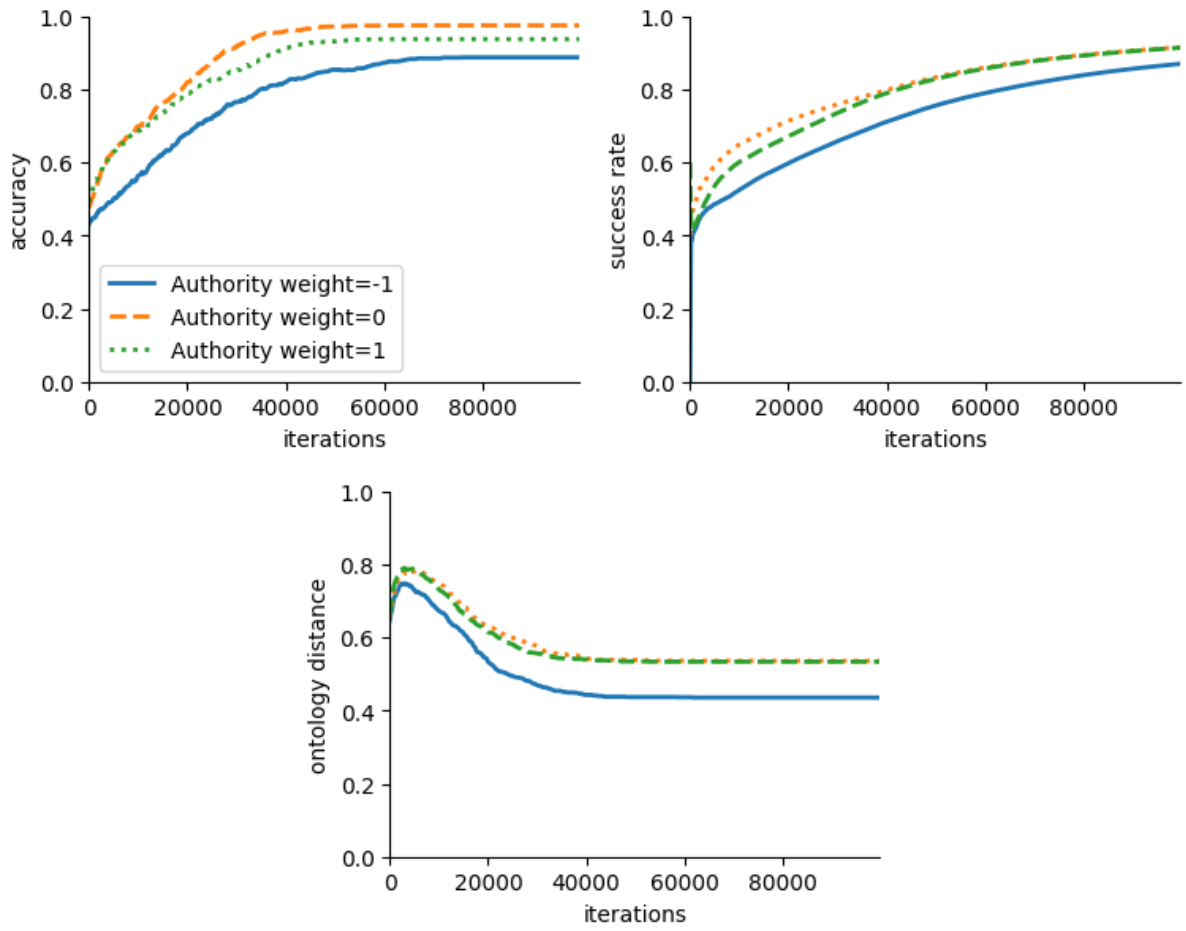

Figure 5.5: Ontology accuracy (top left), success rate (top right) and ontology distance (bottom) when the authority index is used along with a mastery weight value of 1. The plotted lines for the authority weight of 0 (orange dashed) correspond to the runs in which a mastery weight value of 1 is used on its own.

- (H1) is supported: the accuracy only improves when using a mastery weight of 1. Furthermore, using a mastery weight of -1 results in the accuracy converging to 0.
- (H2) is partially supported: an independence index weight of -1 does increase the final success rate, but only when it is used on its own. Using it without the mastery index does not however improve accuracy.
- (H3) is partially supported: a novelty weight of -1 increases the final ontology distance, both when using it on its own, and in combination with the mastery index. However, even when used alongside the mastery index, it leads to a very low accuracy.
- (H4) is not supported: the authority index has no significant effect on the accuracy and success rate. Furthermore, its effect on ontology distance when used alongside the mastery index is negative.

Having analyzed our results and tested our hypotheses, the next and final chapter will draw the final conclusions and propose ideas for further work.

— 6 —
Conclusions and Further Work

6.1 Summary of the Problem, Approach and Methodology

Our initial goal was to find ways of using cultural values to influence the behavior of artificial agents in the context of knowledge evolution. To achieve this, we first studied social theories about cultural values, as well as how cultural values have been previously used in multi-agent systems. We chose the Schwartz social theory of values [19] to base our design on, and proceeded to define value-sensitive agents.

In our experimental framework [25], agents represent their knowledge using private ontologies. They play interaction games with each other, with the goal of agreeing. When agents disagree, causing their interactions to fail, they may take adaptation actions to agree with each other. We want the decision of which agent adapts to be influenced by cultural values.

We therefore defined 4 individual value indices: independence, novelty, authority and mastery, which we use to assess the compatibility of an adaptation action with a cultural value.

The cultural value system of a population of agents is defined by the value indices and their associated weights, which denote the value's importance in the population. The compatibility of an adaptation action with a population's cultural values is an aggregation of the compatibility with the 4 indices and their importance in the population. When their interactions fail, agents will choose the adaptation action most compatible with their population's cultural value system.

To evaluate the effects of the four value indices, we ran a full factorial experiment varying each index weight. We then formulated four hypotheses, one corresponding to each index:

(H1) A positive mastery weight is needed to achieve a significant increase in accuracy. A negative mastery weight causes the accuracy to converge to 0.

(H2) A negative independence weight causes the interactions' agreement to converge faster to a higher value.

(H3) A negative novelty weight has a positive effect on knowledge diversity.

(H4) Using a positive authority weight alongside a positive mastery weight leads to an increase in accuracy.

We analyzed the effects on three measures: (a) the accuracy of agent ontologies, (b) the success rate of agent interactions, and (c) the distance between agent ontologies. ANOVA and Tukey HSD tests were performed to assess the statistical significance of the observed effects.

The result dataset and the complete analysis are available at [26] and [27], respectively.

6.2 Results

The results showed that a mastery index weight of 1 was needed to improve the accuracy. Additionally, a mastery weight of -1 causes the accuracy to converge to 0, confirming our first hypothesis (H1). This was expected, as mastery is the only index which provides agents with feedback from the environment.

The independence weight of -1 only increases the success rate when used without the mastery index. It however does not improve accuracy when used alone. Our second hypothesis (H2) is therefore only partially supported.

The novelty weight of -1 increases the ontology distance, but it leads to a very low accuracy, even when used alongside the mastery index. Therefore, the third hypothesis (H3) is also only partially supported.

The authority index does not have a significant effect on accuracy, even when used alongside the mastery index, rejecting our fourth hypothesis (H4).

6.3 Comments

As suggested by our results, our indices did not fully achieve the desired effects. We believe that their design can be improved.

For instance, the authority index proved to have no significant effect on neither the accuracy nor the success rate. In our design, a population which favors authority will prefer the adaptation actions that preserve the knowledge of the agent which has adapted the least in its past interactions. This, however, is merely an amplifier of the other components of the cultural value system: when an agent adapts, it does so because the action of adapting is the most compatible alternative with the cultural value system of its population. In our experiment, all agents belong to the same population, rendering this index worthless. It would perhaps be interesting to use it in simulations of multiple populations with distinct cultural value systems.

In the case of independence and novelty, we did observe desired effects on the interaction success rate and ontology distance, respectively. They however came at the cost of low accuracy. We believe that the weight values of -1 may be too extreme. A "weaker" disincentive for novelty, for instance, might still increase the ontology distance without such a dramatic drop in accuracy.

6.4 Perspectives

This work may call for more exploration of human-inspired cultural values in artificial knowledge evolution. We showed that it is possible to use human-inspired cultural values to influence the behavior of artificial agents in the context of knowledge evolution.

As previously stated, we believe that it is worth further exploiting these indices in simulations of multiple populations of agents, using distinct cultural value systems. For instance, it would be interesting to simulate the interactions between two populations of agents, only one of which using the mastery index. Could a population that does not employ a mastery index

learn correct knowledge from a population that does and preserve it over time? Could a population that uses a positive mastery index diversify its knowledge by interacting with a different population, without a large reduction of knowledge correctness?

In our experiment, we only used the discrete values $\{-1, 0, 1\}$ for the weights. It might also be worth using intermediate values, especially for independence and novelty (i.e. $-0.5, -0.25, -0.1$) and verify whether their positive effects on success rate and diversity hold, with a smaller or no decrease in accuracy.

Going even further, one could re-think the decisions that are influenced by cultural values. In our work, cultural values influence the choice of the agent which adapts. They might however also be used to influence *how* agents adapt. For instance, one might design different adaptation operators and have the agents choose which one to use based on their cultural values.

Bibliography

References

[1] Michael Anslow and Michael Rovatsos. “Aligning experientially grounded ontologies using language games”. In: *Graph Structures for Knowledge Representation and Reasoning: 4th International Workshop, GKR 2015, Buenos Aires, Argentina, July 25, 2015, Revised Selected Papers 4*. Springer. 2015, pp. 15–31.

[2] Line van den Berg, Manuel Atencia, and Jérôme Euzenat. “Agent ontology alignment repair through dynamic epistemic logic”. In: *AAMAS 2020-19th ACM international conference on Autonomous Agents and Multi-Agent Systems*. ACM. 2020, pp. 1422–1430.

[3] Yasser Bourahla, Manuel Atencia, and Jérôme Euzenat. “Knowledge improvement and diversity under interaction-driven adaptation of learned ontologies”. In: *AAMAS 2021-20th ACM international conference on Autonomous Agents and Multi-Agent Systems*. 2021, pp. 242–250.

[4] Paula Chocron and Marco Schorlemmer. “Attuning ontology alignments to semantically heterogeneous multi-agent interactions”. In: *Proceedings of the Twenty-second European Conference on Artificial Intelligence*. 2016, pp. 871–879.

[5] Jérôme Euzenat. “First experiments in cultural alignment repair (extended version)”. In: *The Semantic Web: ESWC 2014 Satellite Events: ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers 11*. Springer. 2014, pp. 115–130.

[6] Jérôme Euzenat. “Interaction-based ontology alignment repair with expansion and relaxation”. In: *IJCAI 2017-26th International Joint Conference on Artificial Intelligence*. AAAI Press. 2017, pp. 185–191.

[7] Geert Hofstede. *Culture’s consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage, 2001.

[8] Geert Hofstede and Michael H Bond. “Hofstede’s culture dimensions: An independent validation using Rokeach’s value survey”. In: *Journal of cross-cultural psychology* 15.4 (1984), pp. 417–433.

[9] Geert Hofstede, Gert Jan Hofstede, and Michael Minkov. *Cultures and organizations: Software of the mind*. Vol. 2. Mcgraw-hill New York, 2005.

[10] Ronald Inglehart. “Values, objective needs, and subjective satisfaction among western publics”. In: *Comparative Political Studies* 9.4 (1977), pp. 429–458.

[11] Andreas Kalaitzakis and Jérôme Euzenat. “À quoi sert la spécialisation en évolution culturelle de la connaissance?” fr. In: *Actes 31^{e} journées francophones sur Systèmes multi-agent (JFSMA), Strasbourg (FR)*. Ed. by Maxime Morge. 2023. URL: https://moex.inria.fr/files/papers/kalaitzakis2023a.pdf.

[12] Florence R Kluckhohn and Fred L Strodtbeck. *Variations in value orientations.* Row, Peterson, 1961.

[13] Yao Li et al. “Cross-Cultural Privacy Prediction.” In: *Proc. Priv. Enhancing Technol.* 2017.2 (2017), pp. 113–132.

[14] Rijk Mercuur, Virginia Dignum, and Catholijn Jonker. “The value of values and norms in social simulation”. In: *Journal of Artificial Societies and Social Simulation* 22.1 (2019).

[15] Alex Mesoudi, Andrew Whiten, and Kevin N Laland. “Towards a unified science of cultural evolution”. In: *Behavioral and brain sciences* 29.4 (2006), pp. 329–347.

[16] Talcott Parsons. *The social system*. Free Press, 1951.

[17] J. Ross Quinlan. “Induction of decision trees”. In: *Machine learning* 1 (1986), pp. 81–106.

[18] Lilach Sagiv and Shalom H Schwartz. “Personal values across cultures”. In: *Annual review of psychology* 73 (2022), pp. 517–546.

[19] Shalom Schwartz. “A theory of cultural value orientations: Explication and applications”. In: *Comparative sociology* 5.2-3 (2006), pp. 137–182.

[20] Shalom H Schwartz. “National culture as value orientations: Consequences of value differences and cultural distance”. In: *Handbook of the Economics of Art and Culture*. Vol. 2. Elsevier, 2014, pp. 547–586.

[21] Shalom H Schwartz et al. “An overview of the Schwartz theory of basic values”. In: *Online readings in Psychology and Culture* 2.1 (2012), pp. 2307–0919.

[22] Luc Steels. “Experiments in cultural language evolution”. In: *Experiments in Cultural Language Evolution* (2012), pp. 1–318.

[23] LCBC Vanhée. “Using Cultures and Values to Support Flexible Coordination”. PhD thesis. Utrecht University, 2015.

Unpublished Sources

[24] Yasser Bourahla, Manuel Atencia, and Jérôme Euzenat. “Accuracy and diversity of agent ontologies adapted by interaction”. Submitted for publication. 2022.

Internet Sources

[25] *Lazy lavender*. https://gitlab.inria.fr/moex/lazylav. 2023.

[26] Adriana Luntraru. *20230523-VBCE: Agents adapt ontologies to agree on decision taking. Introducing cultural values*. July 2023. DOI: 10.5281/zenodo.8124000. URL: https://doi.org/10.5281/zenodo.8124000.

[27] Adriana Luntraru. *20230523-VBCE Experiment description*. https://sake.re/20230523-VBCE/. 2023.