UNIVERSITAT DE BARCELONA

Value-aligned norm selection

Marc Serramia Amoros

Value-aligned norm selection

by

Marc Serramia Amoros

A dissertation submitted in partial satisfaction
of the requirements for the award of the degree of
Doctor of Philosophy in Engineering and Applied Sciences.

Advisors:
Dr. Maite López-Sánchez
Dr. Juan A. Rodríguez-Aguilar

June 29, 2021

Abstract

Norms have been widely enacted in both human and agent societies to regulate the actions that individuals can perform. However, although legislators may have ethics in mind when establishing norms, moral values are seldom explicitly considered. This thesis advances the state of the art in normative multi-agent systems by providing quantitative and qualitative methods for a decision maker to select the norms to enact within a society that best align with the moral values of such society. We call the problem of selecting these norms, the *value-aligned norm selection*.

The quantitative approach to align norms and values is grounded on the ethics literature. Specifically, from the study of the relations between norms, actions and values in the literature, we formally define how actions and values relate, through the so-called *value judgement functions*, and how norms and values relate, through the so-called *norm promotion functions*. We show that both functions provide the means to compute value alignment for a set of norms, and also that our norm selection problem can be cast as an optimisation problem: finding the set of norms that maximises value alignment. Furthermore, we provide an encoding to solve the value-aligned norm selection problem with off-the-shelf solvers. Finally, we illustrate our approach with a case study and provide an empirical analysis on the hardness of solving norm selection problems.

While utilitarian approaches are commonplace in multi-criteria decision making, utilities may not always be available or easy to specify. In the case of value-aligned norm selection, assessing numerically how a norm relates to a value may not be easy for a decision maker. In more general terms, decision makers can often be confronted with the need to select a subset of objects from a set of candidate objects by just counting on qualitative preferences regarding some criteria. In fact, this constitutes a family of problems, which we formalise as *dominant set selection problems*. We propose two approaches to solve the dominant set selection problem depending on how elements relate to the criteria. Both approaches are based on transforming the criteria

preferences to preferences over all possible sets of objects. We accomplish so by: (i) grounding the preferences over criteria to preferences over the objects themselves; and (ii) lifting these preferences to preferences over all possible sets of objects. Since the value-aligned norm selection problem is a particular instance of the dominant set selection problem, we can readily adapt the proposed qualitative approaches to perform value-aligned norm selection.

Our first qualitative approach supposes binary relations between elements and criteria. In the case of value-aligned norm selection, norms either promote or do not promote values. This approach relies on combining lex-cel (an existing method in the literature to ground preferences over criteria to preferences over elements) with our novel anti-lex-cel (a function that lifts preferences over elements to preferences over sets of these elements), which we formally (and thoroughly) study. Furthermore, we provide a binary integer program (BIP) encoding for the value-aligned norm selection problem to solve it with optimisation libraries.

Building on the first approach, we consider labelled relations between elements and criteria. For example, in the case of value-aligned norm selection, norms can promote or demote values with different degrees, we can capture these degrees of promotion and demotion through labels. This calls for a new decision making framework, which we formally introduce. Within such framework, we introduce a new method to ground preferences over criteria to preferences over single elements considering the labelled element-criterion relations: *multi-criteria lex-cel*. The resolution of the value-aligned norm selection problem in this case relies on the combination of *multi-criteria lex-cel* and anti-lex-cel. Here, we also provide a binary integer program encoding to solve the value-aligned norm selection problem. Furthermore, we formally establish that the contributions of this second approach generalise recent results in the social choice literature.

While we formalise both qualitative approaches in general, we thoroughly illustrate their application to the case of value-aligned norm selection.

Resum

Les normes s'han utilitzat àmpliament en societats tant d'humans com d'agents per regular les accions permeses als seus individus. Tanmateix, tot i que els legisladors poden estar considerant aspectes ètics de forma intrínseca quan defineixen normes, aquests aspectes no són usualment considerats de forma explícita. Aquesta tesi avança l'estat de l'art en sistemes multiagent normatius formalitzant mètodes quantitatius i qualitatius per seleccionar les normes d'una societat que millor s'alineïn als valors morals d'aquesta societat. Anomanem *selecció de normes alineades als valors morals* al problema de seleccionar aquestes normes.

La resolució quantitativa del problema de selecció de normes alineades als valors morals està basada en la literatura d'ètica. Arran de l'estudi de les relacions entre normes, accions i valors que es fa a la literatura, proposem una definició formal de les relacions entre accions i valors a través de les *funcions de judici*, i de les relacions entre normes i valors a través de les *funcions de promoció*. Aquests dos tipus de funcions poden ser utilitzades per computar quant d'alineat està un conjunt de normes amb els valors morals. D'aquesta manera, podem traduir el nostre problema de selecció de normes alineades als valors morals a un problema d'optimització: el de trobar el conjunt de normes de màxim alineament amb els valors. A més a més, definim una codificació del problema de selecció de normes alineades als valors morals. Finalment, il·lustrem aquesta resolució amb un cas d'estudi i fem una anàlisi empírica sobre la dificultat de resolució dels problemes de selecció de normes alineades als valors morals.

Tot i que les resolucions basades en utilitats són comunes en la presa de decisions, les utilitats no sempre són fàcils d'especificar. En el cas de la selecció de normes alineades als valors morals, avaluar numèricament l'impacte d'una norma sobre un valor pot no ser fàcil. En termes més generals, la selecció d'un subconjunt d'elements d'un conjunt de candidats, sol estar guiada per criteris de decisió. De fet, identifiquem aquesta família de problemes que anomenem *problemes de selecció del conjunt dominant*. Proposem dues

resolucions per a aquests problemes depenent en com s'especifiquen les relacions entre els elements i els criteris de decisió. Les dues resolucions es basen en transformar les preferències sobre criteris en preferències sobre conjunts d'elements. Ho fem en dos passos: (i) transformem les preferències sobre criteris en preferències sobre elements; i (ii) transformem les preferències sobre elements en preferències sobre conjunts d'aquests elements. Com que el problema de selecció de normes alineades als valors morals és una instància de la família de problemes de selecció del conjunt dominant, podem adaptar aquestes resolucions per a la selecció de normes.

La primera resolució qualitativa suposa que existeixen relacions binàries entre elements i criteris. En el cas de la selecció de normes alineades als valors morals, les normes promocionen o no promocionen els valors. Així, la resolució consisteix en combinar lex-cel (un mètode de la literatura que transforma les preferències entre criteris a preferències entre elements) amb l'anti-lex-cel (una nova funció per transformar les preferències sobre elements a preferències sobre conjunts d'aquests elements). A més a més, definim una codificació en programació en enters del problema de selecció de normes alineades als valors morals, per poder resoldre el problema mitjançant llibreries d'optimització.

Per millorar la primera resolució qualitativa, considerem etiquetes per expressar relacions graduades (en lloc de binàries) entre els elements i els criteris. Per exemple, en el cas de la selecció de normes alineades als valors morals, considerem que les normes poden promoure o descoratjar els valors en diferents graus. Per poder gestionar aquestes relacions més riques, formalitzem un nou entorn de presa de decisions. En aquest entorn, definim el multi-criteria lex-cel, una nova funció per convertir les preferències sobre els criteris en preferències sobre elements considerant les relacions etiquetades entre elements i criteris. D'aquesta manera, la segona resolució qualitativa del problema de selecció de normes alineades als valors morals consisteix a combinar el multi-criteria lex-cel amb l'anti-lex-cel. Per aquesta segona resolució qualitativa també definim una codificació en programació en enters del problema. A més, demostrem que les contribucions d'aquesta segona resolució qualitativa generalitzen resultats recents de la literatura.

Tot i que formalitzem les dues resolucions qualitatives en termes generals, és important remarcar que il·lustrem com s'han d'aplicar en el cas del problema de selecció de normes alineades als valors morals.

Acknowledgements

These last years have been a journey of knowledge, emotions, good moments and not so good moments. I would like to write a few words to thank all the people that have shared their knowledge or have supported me during this thesis.

First of all, I want to thank my advisors, Maite López-Sánchez and Juan Antonio Rodríguez-Aguilar. Even now I cannot fathom how lucky I have been of meeting you and having you as my guides along this journey.

Thank you Maite, for your brilliant teaching during my Bachelor's degree, which sparked my interest in AI. For believing in me and supervising my degree's final research project along with Jar, therefore making me realise that I wanted to research more. For your patience in addressing my many questions along all these years, and for your hard work in providing me with comments and ways to improve as a researcher every day.

Thank you Jar, for being an endless source of knowledge to learn from. For always being available to hear my doubts and always having a paper or a contact to answer them. For always looking at the bright side, for your optimism, and for always transforming problems into opportunities.

Thanks to you both, for having treated me as a college, a peer, a friend, even though I was and still am just a novice.

Thanks to the many collaborators who have shared their knowledge with me and have co-authored our papers. In particular, to Michael Wooldridge, for having first thought of the idea of value-aligned norm selection. To Stefano Moretti, for his insight in ranking functions, which was fundamental for the qualitative approaches. To Manel Rodríguez, for sharing his knowledge of the philosophy literature, which resulted in our definition of moral values.

To all the colleagues at the Institut d'Investigació en Intel·ligència Artificial (IIIA-CSIC) for having shared wonderful ideas, seminars, and moments. To the PhD students in the lab, thank you for the many chats and lunches we have shared. To the entrance and cleaning staff, for sparking conversations in the long afternoons. To the administration staff, for having always

helped me rapidly and with a smile on their faces. I cannot express how grateful I am of having had the honour of having my first job in such a wonderful work environment.

In more personal terms, I would like to thank those people around me who have put up with me during this thesis and have always believed in me and offered their most sincere support.

Thanks to my mother Jordina, gràcies mama per sempre haver-te preocupat per mi. Per ser com ets i per haver-me educat tan bé com has pogut. Per aguantar-nos dia a dia. T'estimo.

To my father Joan, gràcies papa perquè ni que siguis una mica seriós com jo, ens estimem molt, sempre has estat amb mi quan t'he necessitat i mai t'ha fet res canviar de plans per ajudar-me. T'estimo.

To my grandparents Ramona and Josep, gràcies padrins, per haver-me fet créixer i ensenyar-me la importància d'estudiar. Gràcies padrina, que com ma mare sempre m'has ajudat en tot el que he fet.

To my friend, "brother from another mother", and fellow PhD student, David Solé. Thank you for the many adventures we have shared during these years, and for understanding and supporting me with the PhD headaches and chores.

Thanks to all of you, you have been my daily energy to stand up to each challenge along this journey.

Marc Serramia
Barcelona, Juny 2021

Research supported by projects Collectiveware TIN2015-66863-C2-1-R (MINECO/FEDER), AI4EU (H2020-825619), MISMIS (PGC2018-096212B-C33), and "Artificial Intelligence applied to the Decidim Barcelona citizen participation platform" (code 20S02623-0018.30) from the Barcelona City Council through the Fundació Solidaritat de la UB.

Contents

List of Figures

Chapter 1

Introduction

Norms have been extensively established in both human and agent societies as a means to regulate societies [Boella et al., 2006; Sethi and Somanathan, 1996]. Within agent societies, problems such as norm synthesis [Shoham and Tennenholtz, 1995; Ågotnes and Wooldridge, 2010], norm emergence [Griffiths and Luck, 2010; Villatoro et al., 2011], norm learning [Savarimuthu et al., 2013; Campos et al., 2013; Riveret et al., 2014], or norm adoption [Castelfranchi, 1999] have been widely studied. One of the main research questions in normative multi-agent systems (NorMASs) research is how to engineer a normative system that regulates the actions the agents can perform in different situations. Thus, the literature in NorMASs has tackled the engineering of normative systems driven by a variety of goals.

An important aspect when regulating multi-agent systems (MASs) is to consider the fact that actions have ethical implications. Thus, along the lines of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems [IEEE Standards Association, 2016], with a committee devoted to "Embedding Values into Autonomous Intelligent Systems", here we take the stance that moral values must be a first-class criterion to consider when deciding on the regulation of a society. Therefore, note that by carefully selecting the norms to enact in a system, the system designer will ultimately constrain agents' ethical behaviour. This thesis aims at defining how to compose normative regulations for multiagent systems taking into account the moral values that norms promote and demote.

We structure this chapter as follows. First, in Section 1.1 we motivate our research. Second, Section 1.2 introduces the research questions we address in this thesis. Then, Section 1.3 outlines the contributions we provide to answer these research questions. Section 1.4 details how the thesis is structured.

Finally, in Section 1.5 we list the publications derived from this thesis.

1.1 Motivation

To motivate the work in this thesis, here we look with more detail into the general ideas that we have already discussed. We provide some examples to motivate the need for ethics in AI. Specifically, we show how ethics should also be taken into account when establishing regulations within a MAS. Thus, selecting norms considering the values they promote and demote becomes a useful tool to address this problem.

With the progressive influence of AI in our daily lives, it has become increasingly important to ensure that AI systems act in a way that is aligned with human values. As discussed by [Russell, 2019], we should seek to prevent that AI systems act in hostile ways towards humans. This quest to ensure AI systems act in a way aligned with human moral values is called the *value alignment problem*.

In multi-agent systems in particular, norms have been shown to be a useful means to regulate agent behaviour [Azar, 2004]. With norms we can regulate (permit, oblige, or prohibit) the actions that agents perform. Therefore, we should have to consider the value alignment problem when deciding on the norms to regulate a MAS. Namely, we have to enact norms that are aligned with the values of the society. By designing a normative system (the set of norms that regulate the MAS) while considering the moral values of the society, we ensure that when agents follow the norms, their actions will be aligned with these values.

There has been extensive study on the design of normative systems with a plethora of goals. Some examples of such goals include: avoiding norm conflicts [Kollingbaum et al., 2006; Vasconcelos et al., 2009], minimality and simplicity [Fitoussi and Tennenholtz, 2000; Morales et al., 2014], liberality [Morales et al., 2015b], compactness [Morales et al., 2015a], or stability [Sethi and Somanathan, 1996; Morales et al., 2018]. Nonetheless, to the best of our knowledge, the alignment of norms with moral values has never been considered as a goal when designing normative systems. This thesis is devoted to studying the problem of composing normative systems with the aim that the resulting norms are those that best align with the moral values of the society. We call this new problem *value-aligned norm selection* (VANS), which is the main problem that we address in this thesis.

Importantly, a fundamental component of value-aligned norm selection is

the formal representation of the moral values of the society. There has been some proposals on how to represent moral values and their preferences in AI [Bench-Capon and Atkinson, 2009; Luo et al., 2017; Serramia et al., 2018a]. The literature usually considers a set of moral values along with preferences over these values in what is called a *value system*. While the formalisation of this structure is enough for the purposes it has currently been applied to, we think its formal definition fails to capture the richness presented in the ethics literature. On the one hand, value systems (as defined in [Bench-Capon and Atkinson, 2009; Luo et al., 2017]) consider values as mere elements with no semantics. In the case of value alignment and norms, the ethics literature has long been studying the relation between values, actions and norms [Cooper, 1993; Hansson, 2018; Chisholm, 1963; von Wright, 1963; Hansson, 2001; McNamara, 2011]. Thus, based on these relations, we can better formalise value systems by providing semantics to values. On the other hand, we have to account for value preferences. Indeed, as shown in [Schwartz, 2012; Haerpfer et al., 2020] different societies have different priorities over which values they prefer. Currently, value systems use total orders to specify value preferences, but the reason for using this structure and not another one has not been discussed. This thesis also aims at better formalising value systems by giving semantics to moral values (taking into account the aforementioned relations in the literature) and arguing about the best way to represent value preferences.

Finally, we provide several examples to illustrate the usefulness of value-aligned norm selection. For instance, the *public civility game*, initially introduced in [Rodriguez-Soto et al., 2020]. This game provides a scenario through which to explore moral dilemmas. In short, the game represents a situation wherein two agents move daily from their initial positions (which can be their homes) to their respective target destinations (their workplaces, for instance). Along their journey, one of the two agents finds garbage on the floor that prevents it from progressing. Each agent in the game can deal with the garbage in different ways, like throwing the garbage aside or taking the garbage to the bin. These actions have different implications. Thus for example, throwing the garbage aside may hurt another agent, while taking the garbage to the bin may distract the agent, meaning that it would be late to work. Thus, depending on the society's values and their preferences, we could regulate this scenario differently. An individualist society would regulate for agents to be able to dispose of the garbage swiftly and to continue their journey even if this causes an inconvenience to the other agents. On the other hand, a collectivist society would regulate in favour of disposing

of the garbage so that it does not hurt other agents, even if this impacts the individual goals of the agent that encounters garbage.

In more particular terms, suppose the following three candidate norms (pictured as blue circles in Figure 1.1):

- a norm permitting agents to throw the garbage aside;
- a norm obliging them to take the garbage to the bin; and
- a norm prohibiting them to hurt another agent when throwing garbage.

Additionally, we consider two moral values to decide which norms to enact in this case, the values of civility and timeliness (shown as green squares in Figure 1.1)).

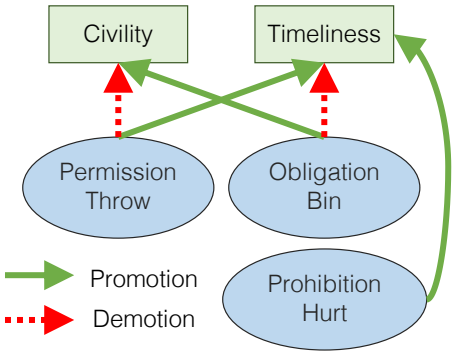


Figure 1.1: Example norms for the public civility game and their relations with the moral values of civility and timeliness.

As pictured in Figure 1.1, the candidate norms are related to moral values by promotion and demotion relationships. Thus, for example, throwing the garbage aside promotes timeliness (it is a fast way to deal with the problem), but demotes civility (since it can hurt another agent). Conversely, taking the garbage to the bin promotes civility (as it clears the path for other agents), but demotes timeliness (as the agent is distracted and can arrive late at its destination). Finally, prohibiting to hurt other agents when throwing garbage aside promotes timeliness (as it allows to throw garbage aside when nobody is there, thus disposing swiftly of the garbage) and neither promotes nor demotes civility (it does not hurt anybody, but the path remains dirty).

Depending on the preferences of the society over these values, we should enact different norms. A society that highly values civility will benefit from the norm obliging agents to take the garbage to the bin. On the other hand,

if timeliness is more important than civility (suppose for example the agents are part of an emergency service), then disposing swiftly of the garbage might be a more value-aligned norm. Note, though, that when selecting norms we should also take into account their relations to avoid incompatibilities or redundancies. For example, we should avoid selecting the norm permitting to throw the garbage aside together with the norm obliging to take the garbage to the bin because they are incompatible.

While the public civility game is only an illustratory example, it is important to stress the general idea we have motivated with it: by carefully composing the norms governing a MAS, we can regulate agent behaviour so that it becomes ethically aligned with the moral values of the society. Figure 1.2 provides a general picture of the value-aligned norm selection process identifying its input and output. Value-aligned norm selection considers a set of candidate norms N, moral value criteria, and the relation of promotion/demotion between the norms and the moral values. Then, this process aims at selecting the subset of the candidate norms $\Omega \subseteq N$ that best align with the moral values while maintaining soundness among the norms (avoiding incompatibilities and redundancies). Importantly, composing a value-aligned set of norms requires different approaches depending on the available information to the decision maker. If the decision maker is able to quantify or measure the value alignment of the norms we can approach this task quantitatively. Nonetheless, the decision maker might not be able to numerically assess the value alignment of norms, therefore, in that case, we have to resort to a qualitative approach.

This thesis explores the value-aligned norm selection problem, in particular, how to formalise it and its components, as well as the different approaches to solve it and their properties. In particular, we approach the problem from two different perspectives, namely the quantitative and the qualitative perspectives. To illustrate them, in later chapters we will revisit the public civility game and other similar examples.

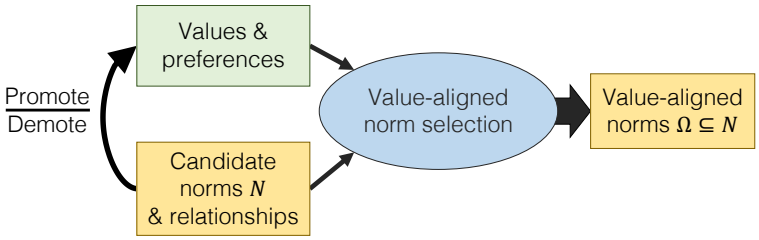


Figure 1.2: The value-aligned norm selection process.

1.2 Research questions

The problem of value-aligned norm selection opens questions on two fronts. The next two subsections are devoted to discussing research questions on i) the definition of the value-aligned norm selection problem and its components, and ii) on the proposed approaches for its resolution.

1.2.1 Norms, values and problem definition

As already discussed in the motivation, in the AI literature, moral values are commonly represented together with their preferences in value systems [Bench-Capon and Atkinson, 2009; Luo et al., 2017]. Nonetheless, the literature so far treats moral values as mere objects without any structure. To fully exploit the computational treatment of value systems, we require a more detailed formal definition that grounds value systems and its components on the philosophical literature. Hence, to that end, we consider the following research question:

Question Q1: How to formally define a value system? Due to the complexity of this question we further divide it into two sub-questions:

- **Question Q1.1**: How do we formally define moral values based on the philosophical literature?
- **Question Q1.2**: How do we represent the preferences over moral values of a value system?

Provided that we have grounded the notion of value system, the next front to address is the normative component of the problem. There is no consensus on how norms for MAS should be defined. However, the literature in normative Multi-Agent Systems has provided a number of alternative norm definitions, see for example [Dignum, 1999; López y López et al., 2002; Boella and van der Torre, 2004; Morales et al., 2015a]. Thus, we pose the following research questions:

Question Q2: How do we define norms and their relations?

With a formalisation of the value system and the normative component of the problem to decide about, we have to address how these two components are related and how this relation is formalised. The relation between

norms and values is one of the main subjects of research in ethics [Chisholm, 1963; von Wright, 1963; Hansson, 2001; McNamara, 2011]. Nonetheless, to the best of our knowledge, this has not been addressed to the same degree on the AI literature. Hence, we must tackle:

Question Q3: How are norms and values related?

With a clear idea of the definition of the structures of the problem and their relations, the question that follows is to define the problem itself:

Question Q4: How do we define the value-aligned norm selection problem?

1.2.2 The value-aligned norm selection process

As motivated previously, value-aligned norm selection requires different approaches depending on the information available to the decision maker. Thus, we explore a quantitative approach when the decision maker is able to quantify the relations between norms and values. On the other hand, we prospect a qualitative approach when such numerical information is not at hand. Each of these approaches has its different research questions, we first look into the ones pertaining to the quantitative approach.

Quantitative approach

The quantitative approach requires that the decision maker is able to numerically assess the relation between each norm and value. Considering both the value preferences together with these numerical assessments of the relations between each norm and value, the approach relies on computing how value-aligned each norm is with regards to all values. Thus, we pose the following research question:

Question Q5: How do we assess the value alignment of a single norm from its relation to each moral value, considering the value preferences?

Once we are able to assess the value alignment of a single norm, we have to address assessing the value alignment of a set of norms. From there, solving the value-aligned norm selection problem consists on finding the norm system with maximum value alignment. Nonetheless, we also have to take into

account the aforementioned norm relations, and ensure that the resulting norm system is free of conflict or redundancy. Thus, we must also consider:

Question Q6: How do we solve the value-aligned norm selection problem? In other words, how do we compose a norm system with maximum value alignment and taking into account norm relations to avoid norm conflicts or redundancy?

It is not obvious that solving the value-aligned norm selection problem is tractable or what its computational cost is. Furthermore, we have to study how different problem features affect its computational time. Note that, number of norms, types and number of norm relations, number of values, etc. are features that can impact differently the tractability of the problem. Thus, in order to ensure that the problem can be solved in a reasonable time and independently of the aforementioned features, we should study several instances varying these features to ensure that in all cases the problem is tractable. Hence, we propose to study:

Question Q7: Is solving the value-aligned norm selection problem computationally feasible and, if so, what factors affect the time required to solve it?

Qualitative approach

The decision maker may not be able to quantify the relationships between norms and values. Instead, it might be simpler to express such relationships qualitatively. This leads to further research questions:

Question Q8: How can we represent qualitatively the relations between norms and values?

Then, the idea of the qualitative approach is to exploit the known value preferences and cast them to norm preferences to afterwards use them to select value-aligned norms. So, we should consider how this process should be performed:

Question Q9: How do we solve the value-aligned norm selection problem qualitatively? This leads to two sub-questions:

- **Question Q9.1**: How do we transform preferences over values to

preferences over norms?

- **Question Q9.2**: How do we exploit preferences over norms to select the most value-aligned norms?

Analogously to the questions posed when following a quantitative approach, we are interested in studying these approaches theoretically as well as in exploring their use in practice. Hence, we propose to address:

Question Q10: Are qualitative approaches computationally feasible?

Furthermore, a qualitative approach for value-aligned norm selection can be useful for other similar decision making problems. Indeed, while we want to select norms based on their relations with moral values, this setting is not different to selecting elements based on their relation to given criteria. For example, building employee teams based on their capabilities taking into account some team-building criteria, building a diet based on the properties of food and following some general dietary guidelines, etc. In that regard, it seems a good idea to study whether it is possible to generalise the qualitative approach. Thus, we ask:

Question Q11: How can we generalise the qualitative approach to value-aligned norm selection to use it in other multi-criteria decision making problems?

1.3 Contributions

In relation to the research questions above, this thesis advances the state of the art by providing several models of the value-aligned norm selection problem and different approaches to solve it. The following subsections discuss the contributions with regards to the different research questions identified in Section 1.2.

1.3.1 Formalising the value-aligned norm selection problem

The first part of Chapter 3 is devoted to formalising the relations between actions, norms and values. The aim there is to answer research question

Q1. In particular, to establish a definition of moral values and their preferences. To that end, we study the relation between values and actions. We have detected two approaches in the AI literature to define the relations between values and actions. Firstly, [Tielman et al., 2018] proposes a straightforward view: the performance of actions promotes or demotes values in a measurable, commensurable, and comparable way. Other works see value promotion/demotion through state transitions [van der Weide et al., 2009; Bench-Capon, 2016; Luo et al., 2017]. While both of these approaches seem different, in the end they are based on the same idea: they consider value promotion/demotion happens through action performance (note that a state transitions when actions are performed). Nonetheless, these frameworks are not able to capture how not performing an action affects values. Take for example a state in which an agent finds itself in front of an accident with harmed people, and consider the value of solidarity. If the agent does not perform any action, it is hard to think that the value of solidarity is neither promoted nor demoted. In fact, in this case solidarity should be demoted. This is in line with the ethics literature, which considers that values judge how good or bad actions are to perform or to not perform [Chisholm, 1963]. Thus, we exploit this idea and formalise the concept of moral value through what we call a judgement function.

The other topic we address is that of preferences between values. Unlike the value systems in [Bench-Capon and Atkinson, 2009; Luo et al., 2017], which consider total orders (without an apparent reasoning), we define value preferences as rankings [Barberà et al., 2004]. Rankings are less strict than total orders since they allow for indifferently preferred values, while they still satisfy totality, a fundamental property when we need to compare values between them. With that, we formalise value systems as a set of moral values and a ranking over them. Thus, in terms of values and their preferences our contribution is:

Contribution C1: Formal definition of value system grounded on the ethics literature. Moral values judge how good or bad are actions whether they are performed or not. Exploiting this idea, we provide:

- **Contribution C1.1**: A novel formal definition of moral value. This definition provides semantics to moral values through the judgement function relating values and actions.

- **Contribution C1.2**: A formalisation of value preferences as rankings.

Chapter 3 addresses many of the research questions with regards to the value-aligned norm selection problem. When it comes to question Q2, we formally introduce a multi-agent system, and define actions and their context. Then, the concept of *normative domain* is presented, which can be regarded as the domain over which the process of composing a normative system takes place. This concept is later simplified in Chapters 4 and 5 as the norm net. Within a normative domain, we identify the fundamental relationships between norms. We characterise *sound norm systems* as those without norm conflicts or redundancy. Value-aligned norm selection will aim at finding a sound norm system that best aligns with a value system. Thus, we propose:

Contribution C2: Formalisation of the normative domain, and its simplified version, the norm net. Formalisation of *sound norm systems*, those that are the target of value-aligned norm selection.

Next, we address the relation between norms and values. There has been some research on this relation, especially with regards to environmentalism. In that regard, [Stern et al., 1999] proposes the value-belief-norm theory. This theory states that when individuals adhere to the values of a movement, they believe their actions matter towards those values, which ultimately leads to the activation of personal norms aligned with the values of the movement. Similarly, in the paradigm of socio-hydrology, [Roobavannan et al., 2018] also describe this forward loop where personal values activate personal norms. Furthermore, they also consider a backwards loop, where collective behaviour can spark change on personal norms. In more general terms (outside the area of environmentalism), [Sierra et al., 2019] characterises the value alignment of a norm considering the state transition paths available once we apply the norm. We exploit our novel definition of moral value to enter this discussion. Nevertheless, here we take a different stance, since we profit from the value judgement functions from contribution C1.1. As explained before, value judgement functions assess how good or bad is the performance or non-performance of actions with respect to some value. Since norms regulate actions, norms and values are also related. The relation between a value and a norm depends on the relation between the value and the norm's action (its judgement) and the relation between the norm and the action (how the norm regulates the action). Our notion of promotion function is based on these two relations. Therefore, we offer the following contribution to address research question Q3:

Contribution C3: Definition of the promotion function relating norms and values.

Finally, having formally defined all the components of the value-aligned norm selection problem and how they relate, we are ready to formalise it. We formalise the value-aligned norm selection problem as the problem of finding the sound norm system that best aligns with the value system. Hence, we address question Q4 with the following contribution:

Contribution C4: Formal definition of the value-aligned norm selection (VANS) problem.

1.3.2 Solving the value-aligned norm selection problem

As discussed by our research questions, we propose two different approaches to solve the value-aligned norm selection problem, depending on the information available to the decision maker. We develop quantitative and qualitative approaches. They are detailed in what follows.

Quantitative reasoning

The quantitative approach assumes that the decision maker has enough domain knowledge to numerically assess how each norm promotes or demotes each value. This approach to value-aligned norm selection follows the process outlined in Figure 1.3. Thus, this approach aims at combining all norm-value numerical assessments into an overall norm utility considering also value preferences (i.e., the norm value-alignment utility in the centre of Figure 1.3). The overall norm utility represents how much each norm is aligned with values. Therefore, the higher the utility of a norm, the more value-aligned the norm. Hence, the goal of this approach is to compose a set of norms with maximum overall utility. Note though, that this is not as straightforward as it seems, since we have to take into account norm relations (incompatibilities between norms and redundancies).

Chapter 3 explains in detail the quantitative approach that we propose to solve the value-aligned norm selection problem. This approach assesses the promotion/demotion relation between norms and values numerically. In this manner, it addresses research question Q5 with the following contribu-

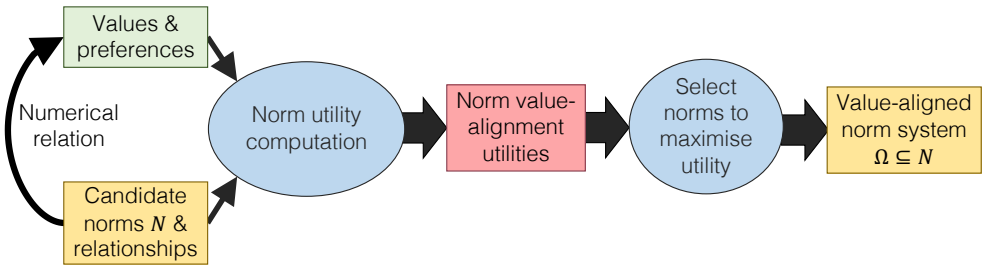


Figure 1.3: The quantitative approach to value-aligned norm selection.

tion:

Contribution C5: Definition of utility functions to compute the value-alignment of a norm with respect to a value system.

Using these utility functions, we tackle the problem of composing the norm system that maximises the overall norm utility while taking into account norm relations. This is addressed by casting the VANS problem as an optimisation problem. Thus, we answer question Q6 with the following contribution:

Contribution C6: An encoding of the value-aligned norm selection problem as a binary integer program (BIP).

Note that, although the BIP encoding allows us to solve the VANS problem, it is not guaranteed that a solver will be able to solve a VANS problem in a reasonable amount of time. Furthermore, the computational cost of solving a VANS problem may vary depending on its structure. For example, a large number of incompatibility relations between norms might make a problem harder to solve. Hence, we should be aware if the problem's features affect the problem's tractability.

In Chapter 3, we show that the VANS problem is NP-hard. Nonetheless, we prove empirically that off-the-shelf commercial solvers are able to handle large-scale VANS problem instances. Furthermore, we study the problem features that affect solving times. We see that some configurations of norm relations (e.g. a high number of incompatibility relations) make the problem harder. Nonetheless, while in these cases solving times increase, in all cases the problem can still be solved in a reasonable amount of time. Hence, with regards to Q7, we contribute with:

Contribution C7: Study of the VANS problem tractability and experimental evaluation of the VANS problem features affecting solving times.

The quantitative approach relies on the assumption that the decision maker is able to numerically quantify how norm promote/demote values. If this assumption does not hold, we propose to perform value-aligned norm selection through a qualitative approach.

Qualitative reasoning

Decision makers may not be able to numerically assert with precision the relation between norms and values. Furthermore, the additivity of utility functions prioritises quantity over quality, which may have unwanted consequences for ethical reasoning. In these cases, we propose qualitative approaches to solve the value-aligned norm selection problem.

Our qualitative approaches do not require the decision maker to provide numerical assessment of the relation between norms and values. Instead, we consider qualitative norm-value relations in two different levels of expressiveness. On the one hand, our first qualitative approach considers the bare minimum information to build the relations between norms and values. This is represented through binary promotion or no promotion relations between norms and values. On the other hand, our second qualitative approach allows for more expressiveness since it considers different degrees of norm-value promotion and demotion using labels. Independently of how the relation between norms and values is specified, both qualitative approaches follow the same idea. As pictured in Figure 1.4, the procedure consists on transforming the preferences over values to preferences over norms taking into account norm-value relations. We introduce a mechanism for inferring preferences over norms so that these embody the norms' value-alignment. Hence, the more preferred a norm, the more it aligns with values. Once we obtain the preferences over individual norms, our approach proceeds to lift these preferences to preferences over sets of norms. Then, solving the VANS problem amounts to selecting the more preferred sound norm system, that is the most preferred set of norms in the ranking that satisfies soundness.

Chapters 4 and 5 describe the two qualitative approaches (considering binary and graded norm-value relations respectively). With regards to question Q8 each approach answers it differently, hence leading to different contributions:

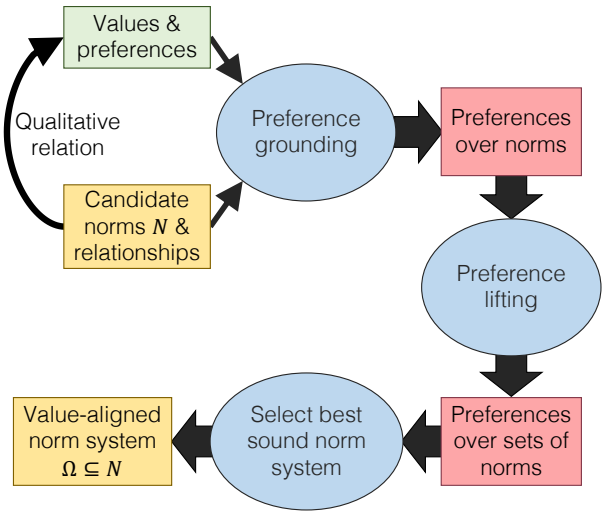


Figure 1.4: The qualitative approach to value-aligned norm selection.

Contribution C8: A qualitative definition of the norm-value relation:

- Chapter 4 considers a binary promotion or no promotion relation, in other words, each norm either promotes or does not promote each value.

- Chapter 5 considers a graded relation that allows for different degrees of norm promotion or demotion towards values.

Our proposed qualitative approaches transform preferences over values to preferences over sets of norms. To accomplish that, we resort to composing ranking functions, a novel approach that to the best of our knowledge has not been previously studied in the literature. As already outlined in Figure 1.4, our approaches consist on two steps, namely preference grounding and preference lifting. In particular our contributions to answer question Q9 are:

Contribution C9: We design a general process for tackling value-aligned norm selection by counting on the composition of two transformations: preference grounding (from preferences over values to preferences over norms); and preference lifting (from preferences over norms to preferences over sets of norms).

- **Contribution C9.1**: Depending on how norm-value relations are defined we use different functions to perform preference grounding:

  - In Chapter 4 we consider binary norm-value relations. To perform preference grounding we exploit a recently introduced family of ranking functions called social rankings [Moretti and Öztürk, 2017]. In particular, we perform preference grounding through an adaptation of the lex-cel function, a social ranking function introduced in [Bernardi et al., 2019].
  - In Chapter 5, norm-value relations are graded. Thus, we introduce a novel category of ranking functions named multi-criteria ranking functions. We show that this family generalises the family of social ranking functions of [Moretti and Öztürk, 2017]. In particular, we formalise a novel ranking function called MC lex-cel, which we use to perform preference grounding. This function is a generalisation of the lex-cel function of [Bernardi et al., 2019].

- **Contribution C9.2**: With regards to preference lifting, we introduce the novel anti-lex-cel ranking function (a function that generalises that of [Bossert et al., 1994]). Anti-lex-cel builds a ranking over all sets of norms embodying value alignment. The solution to the VANS problem is the most preferred norm system in this ranking that satisfies soundness.

Chapters 4 and 5 prove that the solution provided by these transformations is indeed the solution of the VANS problem. Nonetheless, these transformations are computationally costly, which is a concern, as noted by research question Q10. Indeed, building the ranking for all sets of norms requires to compute preferences for $2^{|N|}$ sets. Conversely, the use of a binary integer programs has proved to be useful in the quantitative approach. Therefore, we use a BIP to solve the problem here and avoid the computational cost of building the whole norm system ranking. Thus, we propose:

Contribution C10: A BIP encoding of the problem to obtain the solution of the qualitative approach avoiding the computational cost of the grounding and lifting transformations. Importantly, we prove that this BIP encoding produces the same solution obtained through the grounding and lifting functions.

Much like with value-aligned norm selection, some actual-world decision making problems require to select a (sub)set of elements despite decision makers only counting on preferences over some decision criteria instead of

the elements themselves. Examples of such problems are committee selection, coalition formation, product line composition, budget allocation, or college admissions [Fishburn, 1992; Gale and Shapley, 1962; Roth and Sotomayor, 1992]. Interestingly, we can think of many other similar set selection problems, such as selecting the team of players for a match (where we prefer some types of players over others), personnel selection (where some capabilities may be preferred over others), etc. Considering this last example, picture the following situation. A school head master must decide on which students to grant admission to. For that, the head master leverages on the admission policy of the school, which, for instance, prioritises some minorities, or fosters impoverished neighbourhoods. Such policies can be cast as preferences over student selection criteria. Nonetheless, the head master lacks of a straightforward manner to rank all possible sets of students. Moreover, there is a further dimension of complexity: some sets may not be eligible (e.g. because of limited budget, or unfulfilment of minority quotas). And yet, despite only counting on preferences over criteria and not sets, the head master must select the most preferred set of students. While we can readily adapt utilitarian approaches in the literature to solve these problems, decision makers may not have enough domain knowledge to quantify how elements relate to the criteria. Thus, by generalising our qualitative approach to solve the VANS problem we can adapt it to suit this family of problems.

Hence, for research question Q11, Chapters 4 and 5 describe respectively binary and graded qualitative approaches to select elements with regards to some decision criteria. While ultimately both chapters explain how to apply the described approaches to the value-aligned norm selection problem, they previously define both the goals of the problem as well as the methodology to solve it in a general context. Chapter 4 provides a definition of the dominant set selection problem (DSSP), the generalisation of the VANS problem that considers elements instead of norms and decision criteria instead of moral values. The DSSP formalisation relies heavily on the notion of dominance. Chapter 4 explains and formalises this concept assuming binary relations between elements and criteria (i.e., elements either align with a criterion or not). Conversely, Chapter 5 supposes a more complex relation between elements and criteria, which counts on degrees of alignment or unalignment.

Thus, we answer research question Q11 with the following contribution:

Contribution C11: Formalisation of the dominant set selection problem with two different degrees of expressiveness, namely considering binary and

graded element-criteria relations. Resolution of both versions of the problem by means of the novel composition of ranking functions generalising the previous contribution C9. We also provide a BIP encoding to solve the problem avoiding the whole computation of the set ranking, thus generalising contribution C10.

1.4 Dissertation outline

Following the concepts introduced in this chapter, the rest of this thesis is structured as follows:

- **Chapter 2** outlines the related work to this thesis and compares the contributions of this work to similar proposals in the literature.

- **Chapter 3** introduces a quantitative approach to value-aligned norm selection. This approach assumes the decision maker can assess how norms relate to values numerically. Then, value-alignment is treated as a utility function and thus, the solution is the set of norms that maximises this utility.

- **Chapter 4** introduces a qualitative approach to value-aligned norm selection. In this case, the decision maker does not have to numerically asses the relation between norms and values. Thus, this approach transforms the preferences over values to preferences over norm systems in terms of value alignment, then the solution can be selected as the most preferred norm system.

- **Chapter 5** builds on the previous qualitative approach by allowing for more expressiveness. While the qualitative approach in Chapter 4 only counts on norms promoting or not promoting values, the approach in this chapter allows for different degrees of norm promotion or demotion specified through the use of labels.

- **Chapter 6** is devoted to discuss the conclusions of the thesis, the lessons learned along this work, and to provide future research paths.

After the conclusions we provide a List of Notation and Symbols, which outlines the notation used within each chapter.

1.5 Publications derived from this thesis

Some results in this thesis have been already published, as we detail in this section. First, the following publications helped to ground the ethical concepts of this work and provided a first attempt at the quantitative approach:

- Lopez-Sanchez, M., Serramia, M., Rodriguez-Aguilar, J. A., Morales, J., and Wooldridge, M. (2017). Automating decision making to help establish norm-based regulations. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS'17)*, pages 1613–1615. International Foundation for Autonomous Agents and Multiagent Systems

- Serramia, M., López-Sánchez, M., Rodríguez-Aguilar, J. A., Morales, J., Wooldridge, M., and Ansotegui, C. (2018a). Exploiting moral values to choose the right norms. In *Proceedings of the 1st Conference on artificial intelligence, ethics and society (AIES'18)*, pages 1–7

- Serramia, M. (2018). Ethics in norm decision making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 378–379, New York, NY, USA. Association for Computing Machinery

We further explored the quantitative approach of Chapter 3 in:

- Serramia, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J. A., Rodriguez, M., Wooldridge, M., Morales, J., and Ansotegui, C. (2018b). Moral values in norm decision making. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS'18)*, pages 1294–1302. International Foundation for Autonomous Agents and Multiagent Systems

As for the qualitative approach described in Chapter 4, we first explored it in:

- Serramia, M., Lopez-Sanchez, M., and Rodriguez-Aguilar, J. A. (2020). A qualitative approach to composing value-aligned norm systems. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1233–1241

Our paper [Serramia et al., 2020] was considered a premier paper of AAMAS2020, hence we were invited to submit an extended journal version to JAAMAS:

- (In press) Serramia, M., Lopez-Sanchez, M., Moretti, S., and Rodriguez-Aguilar, J. A. (2021a). On the dominant set selection problem and its application to value alignment. *Autonomous Agents and Multi-Agent Systems (JAAMAS)*

Additionally, implementations to solve the value-aligned norm selection problem quantitatively and qualitatively, and also the dominant set selection problem, can be found in:

- Serramia, M., López-Sánchez, M., and Rodríguez-Aguilar, J. A. (2021d). Algorithm to generate the BIP encoding of a VANS problem with quantitative input. https://gitlab.iiia.csic.es/marcserr/vans-quant
- Serramia, M., López-Sánchez, M., and Rodríguez-Aguilar, J. A. (2021c). Algorithm to generate the BIP encoding of a VANS problem with qualitative input. https://gitlab.iiia.csic.es/marcserr/vans-problem
- Serramia, M., López-Sánchez, M., and Rodríguez-Aguilar, J. A. (2021b). Algorithm to generate the BIP encoding of a DSSP problem. https://gitlab.iiia.csic.es/marcserr/dssp

More applied publications derived from the work in this thesis:

- Serramia, M., Ganzer-Ripoll, J., López-Sánchez, M., Rodríguez-Aguilar, J. A., Criado, N., Parsons, S., Escobar, P., and Fernández, M. (2019a). Citizen support aggregation methods for participatory platforms. In Sabater-Mir, J., Torra, V., Aguiló, I., and Hidalgo, M. G., editors, *Artificial Intelligence Research and Development - Proceedings of the 22nd International Conference of the Catalan Association for Artificial Intelligence, CCIA 2019, Mallorca, Spain, 23-25 October 2019*, volume 319 of *Frontiers in Artificial Intelligence and Applications*, pages 9–18, Amsterdam. IOS Press
- Serramia, M., López-Sánchez, M., Rodríguez-Aguilar, J. A., and Escobar, P. (2019b). Optimising participatory budget allocation: The decidim use case. In Sabater-Mir, J., Torra, V., Aguiló, I., and Hidalgo, M. G., editors, *Artificial Intelligence Research and Development*

- Proceedings of the 22nd International Conference of the Catalan Association for Artificial Intelligence, CCIA 2019, Mallorca, Spain, 23-25 October 2019, volume 319 of *Frontiers in Artificial Intelligence and Applications*, pages 193–202. IOS Press

1.5.1 Publications under review

With regards to the graded qualitative approach described in Chapter 5, we have submitted this journal paper:

- (Under review) Serramia, M., Lopez-Sanchez, M., Moretti, S., and Rodriguez-Aguilar, J. A. (2021a). Building rankings to consider multiple decision-making criteria and its application to ethical reasoning. *Information sciences*.

After attending the AAAI 2021 Spring Symposium on Implementing AI Ethics, the organisers invited the attendees to submit an extended abstract for a special issue of the Journal of Philosophy and Technology. We submitted the following paper discussing the lessons learnt from the work in this thesis:

- (Under review) Serramia, M., Lopez-Sanchez, M., and Rodriguez-Aguilar, J. A. (2021f). Value-aligned AI: Lessons learnt from value-aligned norm selection. *Journal of Philosophy and Technology*.

Chapter 2

Related work

As previously mentioned in the introduction, this thesis tries to answer research questions from several areas. This chapter is devoted to briefly introducing the related work on each of these areas. Firstly, we will look into the literature on values and value alignment to show that this literature has not been conveniently exploited to formalise values in AI. The formalisation of values will be paramount to found our quantitative method. Then, we discuss the literature in normative multi-agent systems to show that composing norm systems with regards to their value alignment is a novel problem that has not yet been thoroughly addressed. Finally, we also provide an insight on the literature of social choice, rankings and preferences. This will be useful to address one of the aims of this thesis, which is to explore the novel methods that we propose for qualitative reasoning.

2.1 Moral values in AI

The AI research community has been increasingly active in the study of moral agency. Thus, besides the work by Wallach and Allen on moral machines [Wallach and Allen, 2008] and that of Moniz-Pereira and Saptawijaya on machine ethics [Pereira-Moniz and Saptawijaya, 2016], a number of research papers focus on moral values. Just to mention a few, Murukannaiah et al. [Murukannaiah et al., 2020] provide an excellent roadmap to guide research on ethics and multi-agent systems. Ajmeri [Ajmeri, 2018] tackles the question of engineering agents that can reason about values and act ethically. Floridi and Sanders [Floridi and Sanders, 2004] use values as thresholds: an agent is morally good if all its actions respect that threshold; and it is morally evil if some action violates it. Kohler et al. [Kohler et al., 2014] include ar-

tificial moral agents in multi-agent institutions to accomplish fair resource allocation. Abel et al. [Abel et al., 2016] formalize the ethical learning and decision-making problem as solving a partially observable Markov decision process. Cointe et al. [Cointe et al., 2016] propose a judgement ability for agents to be able to evaluate the rightness and/or goodness of both its own behaviour and those of others. In [Cranefield et al., 2017] the authors select agent plans by optimising (minimising) the sum of the current importance of all values.

Considering the wider perspective of decision making support, we find different ethical decision making systems aimed at aiding humans with ethical dilemmas. Thus, Ethos [Harris Jr et al., 2013] and Dax Cowart [Anderson et al., 1996] correspond to two educational programs that challenge students to solve ethical dilemmas. Alternatively, Truth-Teller [McLaren and Ashley, 1995] compares pairs of given cases (together with their corresponding reasons), presenting ethical dilemmas about whether to tell the truth or not. Finally, SIROCCO [McLaren, 2006] supports the decision maker with both a list of ethical codes and practical cases that are relevant to the dilemma at hand. Instead, our approach deals with ethical principles to automatically provide the decision maker with a solution. Conversely, there are works, such as the one by Pitt et al. [Pitt et al., 2014] or Petruzzi et al. [Petruzzi et al., 2015] that operationalize ethical considerations in resource allocation settings by proposing metrics on fairness and social capital respectively. In fact, when it comes to fairness, there is a plethora of works in machine learning considering it, see [Friedler et al., 2019] for a survey.

Argumentation constitutes another research area that has studied values. Some representative examples include the work of Bench-Capon et al. [Bench-Capon and Atkinson, 2009; Atkinson et al., 2006] and Modgil [Modgil, 2006], which use different "Value-based Argumentation Framework" to decide if a statement is true or to evaluate the goodness of certain actions. In terms of agents (moral agency), Bench-Capon extends agent reasoning with values in [Bench-Capon, 2016]. Specifically, value promotion or demotion is associated to changes in system states when agents perform actions. In [Luo et al., 2017], this idea is further explored as authors introduce agents with an opportunistic behaviour that take advantage of less informed agents to reach those state transitions that further promote their individual values. Although both approaches take into account the impact of values and their preferences, these works consider decision making as an individual process, while we take a system-wide perspective.

More related to this thesis, moral values have also been studied together

with norms. Thus, Fan-Yun et al. [Sun et al., 2019] propose a regulation enforcement mechanism based on ethical considerations. Kasenberg et al. [Kasenberg et al., 2018] infer values (expressed as norms) by behaviour observation. Tielman et al. [Tielman et al., 2018] present a method to derive norms from actions, values and domain. Mercuur et al. [Mercuur et al., 2019] compare human behaviour with agents endowed with moral values and norms, which are expressed with the "non-standard" deontic operator *should*. Finally, [Montes and Sierra, 2021] synthesise the parameters of norms with value-alignment in mind.

Nevertheless, there is still room for advancing the state of the art in the formalisation of value alignment. Particularly, we see that a clear formalisation of values is missing. Hence, for example, while [Luo et al., 2017; Bench-Capon and Atkinson, 2009] consider moral values and some preferences over those values, they could be further detailed. In order to formalise moral values in the framework of normative multi-agent systems, we look into the philosophy, psychology and sociology literature.

When it comes to the specification of values, the literature proposes various views. Firstly, [Timmons, 2012] consider moral values as ethical principles that guide the evaluation of actions. Alternatively, Hartman [Hartman, 1967] formalises goodness not related to actions but to the descriptive properties of entities. Thus, for example, a pen that cannot write is considered as a bad pen. It is also worth mentioning those research works diving into proscriptive systems [Janoff-Bulman et al., 2009], which are based on behavioural activation and positive outcomes – such as desirable states – as opposed to prescriptive systems that are based on behavioural inhibition and negative outcomes. Moreover, positive ethics [Lopez et al., 2012; Boddington, 2017] also constitutes a good example of such ethical systems, since it shifts the emphasis from following rules to aspire to the highest ethical ideals. It does so by integrating values to improve decision making. Finally, the communitarian approach [Etzioni and Etzioni, 2016] highlights those social moral values that the community seeks to foster but are beyond those required by law (e.g. most communities expect parents to care for their children far beyond what the law commands).

From a psychological or sociological perspective, Schwartz [Schwartz, 2012] provides an overview of ten basic moral values that are recognised in cultures around the world. Cultural differences affect the prioritisation assigned to these values. This has also been studied in the World Values Survey [Haerpfer et al., 2020], which keeps track of the evolution and maps the value priorisations of various cultures. In much the same way, we advo-

cate for taking into consideration the relative priority of values when facing complex decision making processes that involve several values. This perspective is also aligned with other contributions in the AI literature, which take into account these preferences in specific domains, such as argumentation [Luo et al., 2017; Bench-Capon and Atkinson, 2009], elderly care [Cranefield et al., 2017], as well as in general settings, such as intelligent systems design [Dignum, 2017].

Thus, in this work, we take inspiration from the literature to formalise moral values (and their relative preferences). We provide a more thorough study and formalisation in Section 3.2. Since our work focuses on actions, norms and values, we look into the literature studying the relations between these three elements. Briefly, we take the point of view of [Timmons, 2012] and consider that moral values judge the goodness of actions. From this, knowing that norms regulate actions, we can infer norm-value promotion/demotion from the value judgement of the regulated action. Finally, as discussed before, the sociology literature considers preferences between values [Schwartz, 2012], something which has already been adopted in the AI literature [Cranefield et al., 2017; Dignum, 2017; Luo et al., 2017; Bench-Capon and Atkinson, 2009]. Thus, we formalise the value system as a structure containing moral values and preferences between them.

2.2 Normative multi-agent systems

Within societies, norms have long been used as a coordination mechanism [Azar, 2004]. We refer to the set of norms enacted in a society as a *norm system*.

Engineering norm systems has been thoroughly studied, thus the literature has proposed many methods to enact norm systems in a multi-agent system. Some examples of these techniques are:

- **Norm emergence**: A bottom-up approach. Thus, in this case, agents themselves build norms to regulate the situations they encounter, and if these norms prove to be useful, they are propagated throughout the multi-agent system. This method is studied in [Axelrod, 1986; Shoham and Tennenholtz, 1997; Sen and Airiau, 2007; Savarimuthu et al., 2007; Sugawara, 2011].

- **Norm synthesis**: A top-down approach. In this case, the system designer or the policy-making authority provide the norms for the agents

to follow. Example works using this method are [Shoham and Tennenholtz, 1995; Ågotnes et al., 2007; Morales et al., 2013; Morales et al., 2015a].

- **Norm programming**: Norm engineering by programming the norms of the multi-agent system with a norm-oriented language or environment. For example, [Arcos et al., 2005; Hubner et al., 2007; Garcia-Camino et al., 2009; Sensoy et al., 2012; Dybalova et al., 2013].

Furthermore, these techniques have a further dimension, namely the time at which norms are designed. We can classify some of the previously mentioned works into the following two groups:

- **Off-line approaches**: Formal approaches that engineer norms before the multi-agent system is initialised. For example, [Shoham and Tennenholtz, 1995; Ågotnes et al., 2007].

- **On-line approaches**: Empirical approaches, in the sense that they engineer the norms considering the feedback received from the multi-agent system, while it is being run. For example, [Morales et al., 2013; Morales et al., 2015a]

Apart from techniques to engineer norm systems, the literature has also addressed engineering norm systems with several goals in mind. Some examples are:

- **Avoiding conflicts**: Norms may conflict with one another, for example a norm prohibiting an action conflicts with a norm obliging the same action. Conflicting norms are useless as agents are unable to comply with one or the other, hence avoiding norm conflicts is a desirable goal when designing norm systems. Some works that tackle this problem are [Kollingbaum et al., 2007; Vasconcelos et al., 2009].

- **Minimality and simplicity**: Minimality aims at designing norm systems that do not overregulate or contain superfluous norms. On the other hand, simplicity aims at engineering norms such that they are computationally easy to process for the agents. These goals have been studied in the works of [Fitoussi and Tennenholtz, 2000; Morales et al., 2014].

- **Compactness**: The compactness of a norm system refers to its size. In terms of [Morales et al., 2015a], the overall number of predicates of its norms. Note that, compactness is closely related to minimality and simplicity as requiring minimality implies a small number of norms and simplicity implies norms with a small number of predicates.

- **Liberality**: Composing norms systems aiming at respecting the autonomy of the agents as much as possible [Morales et al., 2015b].

- **Stability**: The stability of a norm system refers to how this system adapts to unforeseen situations. Thus, a norm system is stable when agents cannot benefit from deviating from these norms [Sethi and Somanathan, 1996; Morales et al., 2018].

Even though the literature has thoroughly studied norms in multi-agent systems, along with methodologies to compose norm systems, and several goals to consider when building these norm systems, moral values have traditionally not been considered. It has not been until recently that norms have started to be related to *moral values* [Hansson, 2001]. Here, we understand moral values as the moral objectives worth striving for [van de Poel and Royakkers, 2011]. In the previous section we have outlined some works studying moral values in multi-agent systems. Recall, for example, [Cranefield et al., 2017] where the authors select agent plans by optimising (minimising) the sum of the current importance of all values. The current importance of a value is computed as the salience of the value in a situation times the difference between the target amount of that value and the current value state [Di Tosto and Dignum, 2012]. Another example is that of [Montes and Sierra, 2021], which synthesises parametric norm systems with value-alignment in mind. Both of these works are closely related to this thesis (and the corresponding papers [Serramia et al., 2018a; Serramia et al., 2018b; Serramia et al., 2020]), since here, likewise [Montes and Sierra, 2021], we work on value-aligned norms, and like [Cranefield et al., 2017] we apply optimisation techniques to accomplish so.

Importantly, choosing norms that promote ethical behaviour (i.e., moral values) naturally induces this ethical behaviour in the society. Moreover, if different moral values can be promoted, then it seems reasonable to prioritise those most preferred ones. Consider, for example, a government that enacts norms limiting pollution. Then, we can easily guess that this government prefers sustainability over other values such as development.

However, the problem of selecting the regulatory norms that best align with the ethical principles of a society (or, in other words, the *most value-aligned norm system*) is not straightforward. In addition to the different values and the preferences over them that a society may have, we must also consider whether norms actually promote or demote those values as well as the degree of promotion or demotion. Some literature in Philosophy discusses some of these aspects [Hansson, 2018]. Nonetheless, in the Artificial Intelligence literature, while value promotion and demotion are commonly considered, degrees of such relations are typically disregarded (e.g. [Luo et al., 2017], [Bench-Capon and Atkinson, 2009], [Serramia et al., 2020]). In fact, to the best of our knowledge such aspects have only been considered in legal cases [Bench-Capon et al., 2013].

To summarise, composing value-aligned norm systems has started to be studied recently, but still needs further research. In this thesis, we formalise and study off-line approaches to norm selection with the goal of value-alignment.

2.3 Rankings for qualitative reasoning

A ranking of objects establishes how the objects compare to each other. Thus, rankings are usually considered in many decision making problems, such as committee selection, coalition formation, product line composition, budget allocation, etc. [Fishburn, 1992]. For example, in college admissions [Gale and Shapley, 1962]: considers each student has a ranking of the colleges they prefer and each college has a ranking of the students they want admitted and the paper aims at studying what they call stable and optimal student-college assignments. Similarly, [Roth and Sotomayor, 1992] provides a more detailed and thorough study of this problem.

It is then no surprise that rankings and ranking functions have been long investigated in the social choice literature. Without aiming for completeness, we highlight three different bodies of work.

- Firstly, from the seminal works on social choice and voting in [Arrow, 2012; Sen, 2017], some works as e.g. [Brandt et al., 2016] propose voting mechanisms for aggregating rankings to obtain a ranking of individual elements.

- Second, [Barberà et al., 2004] study functions that transform rankings over individual elements into rankings over sets of these elements.

Maxmin and minmax [Arlegi, 2003] or leximin and leximax [Pattanaik and Peleg, 1984] constitute some examples of such functions.

- Thirdly, [Moretti and Öztürk, 2017] introduce the social ranking as a mapping that transforms a ranking over sets of elements into a ranking over the individual elements of these sets. Recently, social rankings have attracted much attention and several functions have been proposed: [Haret et al., 2018] base their social ranking function on the *ceteris paribus* majority principle; [Khani et al., 2019] base their work on the notion of marginal contribution; [Bernardi et al., 2019] on lexicographical preferences handled by the *lex-cel* ranking function; and [Allouche et al., 2020] introduce two rankings based on the analysis of majority graphs and minmax score.

Hence, the literature has explored ranked voting, ranking lifting (transform rankings over elements to rankings over sets of the elements), and ranking grounding (transform rankings over sets of elements to rankings over the elements themselves). This allows to solve many decision making problems, such as the ones discussed previously. Nonetheless, more often than not, decision makers do not count directly on preferences over the individual elements (or sets of these elements). Instead, they count on decision criteria and preferences over these criteria. For example, when designing a diet, we may consider different criteria over the food to choose –such as healthiness or tastiness– and preferences over these criteria —e.g., healthiness preferred over tastiness. Or in the case of personnel selection, if we count on a large set of candidates we may not have a clear idea of the preferences over all the candidates, but we may just know those criteria that are most important to us and how the candidates relate to these criteria.

Unfortunately, to the best of our knowledge, the social choice literature has not yet studied how to ground the preferences over criteria to preferences over the candidate elements taking into account how the elements relate to the criteria.

Additionally, while the literature previously cited introduces and studies the properties of many types of ranking functions, studying compositions of such functions and their properties remains an open problem.

Addressing both of these concerns can be useful for a myriad of problems. For example, a function that is able to ground criteria preferences to element preferences composed with a lifting function can transform preferences over criteria to preferences over sets of elements. This can be useful for the diet or personnel selection problem, where we are not deciding on an individual

food or candidate but on a diet (set of foods) or a team (set of candidates). This is also the case of value-aligned norm selection, note that we want to select a set of norms, but we do not have preferences over the norms directly. Instead, we have preferences over moral values (our decision criteria) and we know how norms relate to the values.

2.4 Conclusions

In conclusion, and as argued in the introduction of this thesis, the literature lacks some research in several areas which might be of interest.

Firstly, while ethics have started to become an important topic in AI, we have seen that decision making applications have adopted moral values as mere decision criteria elements. There is no formal notion of value, nor of value system, grounded on the literature. This in turn means that we cannot build upon these formalisations to better define concepts such as the promotion or demotion of values. Some of the cited works may have frameworks in which formalising values might not be straightforward. In our case, though, the literature provides a clear relation between actions, norms and values. Thus, we can profit from these relations to formalise values and value systems.

Secondly, we have seen that the literature has extensively researched normative multi-agent systems in terms of goals and methodologies. Nonetheless, the goal of value-alignment has only started to be considered very recently. Hence, composing norm systems with value-alignment in mind is still an open problem.

Finally, with regards to the social choice and preferences literature, we see that there has been extensive study on ranking transformations from elements to sets of elements and from sets of elements to the elements themselves. Nonetheless, decision makers may not readily know neither of these preferences. It may be the case that they have preferences over decision criteria instead, knowing also how the elements relate to the criteria. On the other hand, while the study of these ranking transformations individually is quite thorough, the study of the composition of these functions has not been looked into. As previously explained, both of these points can be useful for many decision making problems, including value-aligned norm selection.

Chapter 3

Quantitative value-aligned norm selection

3.1 Introduction

As previously discussed, norms have been extensively established as a means to regulate both human and agent societies [Boella et al., 2006; Sethi and Somanathan, 1996]. Within agent societies, problems such as norm synthesis [Shoham and Tennenholtz, 1995; Ågotnes and Wooldridge, 2010], norm emergence [Griffiths and Luck, 2010; Villatoro et al., 2011], norm learning [Savarimuthu et al., 2013; Campos et al., 2013; Riveret et al., 2014], or norm adoption [Castelfranchi, 1999] have been widely studied. One of the main questions in normative multi-agent systems (NorMASs) research is how to engineer a normative system that regulates the actions the agents can perform in different situations. Furthermore, the literature in NorMASs has tackled the engineering of normative systems driven by a variety of goals: lack of conflicts [Kollingbaum et al., 2006; Vasconcelos et al., 2009], minimality and simplicity [Fitoussi and Tennenholtz, 2000; Morales et al., 2014], liberality [Morales et al., 2015b], compactness [Morales et al., 2015a], or stability [Sethi and Somanathan, 1996; Morales et al., 2018].

An important aspect when regulating MASs is to consider the fact that actions have ethical implications. Therefore, by carefully selecting the norms to enact in a system, the system designer ultimately constrains agents' ethical behaviour. Thus, along the lines of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems [IEEE Standards Association, 2016], with a committee devoted to "Embedding Values into Autonomous Intelligent Systems", here we take the stance that moral values must be a first-class

criterion to consider when deciding on the regulation of a society. However, the ethical dimension of norms has started to be considered by MAS research only very recently. The usual approach is to consider the existence of a value system composed of moral values and a relationship between norms and values, the so-called promotion/demotion function [Bench-Capon and Atkinson, 2009; Atkinson et al., 2006; Luo et al., 2017; Lopez-Sanchez et al., 2017; Serramia et al., 2018b]. This function tells whether a given norm promotes (/demotes) a given value, and eventually the degree of promotion (/demotion). Thus, a norm promotion function encodes the *value alignment* of each norm, and hence it can be further employed to compute the value alignment of a *normative system* (i.e., the set of norms enacted in the society [Serramia et al., 2018b]). Although norm promotion functions are often used in the MAS literature, they are seldom formally defined. An initial proposal can be found in [Sierra et al., 2019]: Considering values as preferences over states of the world, the authors then assess the value alignment of a norm in terms of the preference increase for those state transitions affected by the norm. Nonetheless, the ethics literature has long studied the relationship between norms and values. Indeed, in ethics, typically a norm is considered to promote a moral value depending on how it regulates an action and how this action is considered with respect to the moral value [Urmson, 1958; Hansson, 2018]. Therefore, the ethics literature counts on the means to set the foundations for a mathematical definition of such promotion function, and, ultimately, of value alignment for a norm system. It is worth noticing that henceforth, we use the terms moral and ethical interchangeably (without differentiation) as it is common practice in the philosophy literature [Frankena, 1973; Audi, 1999; Fieser and Dowden, 2021].

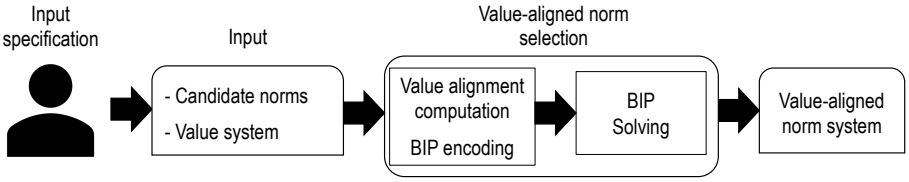


Figure 3.1: The value-aligned norm selection process.

Against this background, in this chapter we investigate the problem of computing the norm system that is best aligned with a value system. As Figure 3.1 shows, this endeavour assumes that a decision maker is tasked with solving that problem that takes as input: (i) a set of candidate norms that have been deemed beneficial for agent coordination, and (ii) a value

system containing a collection of moral values together with preferences over them. From these, value-aligned norm selection is defined as a two-stage process. The first stage is devoted to specifying (building) the problem at hand. For doing so: (i) we propose a methodology to mathematically produce a norm promotion function upon which the value alignment of a norm system can be computed; and (ii) we cast the decision maker's problem as an optimisation problem that can be encoded as a binary integer program (BIP) [Lieberman and Hillier, 2005], and hence optimally solved by any state of the art solver. In this way, we are able to automate the engineering of value-aligned norm systems, namely, of regulations aligned with moral values.

The contributions of this chapter are:

- We formally introduce the concept of *normative domain*, which can be regarded as the domain over which norm reasoning takes place. Within this domain, we identify the fundamental relationships between norms that must be considered to compose a so-called *sound norm system*.

- We establish a formal relationship between actions and moral values. Thus, we introduce the so-called *value judgement function* as the means to evaluate the "goodness" of actions in different situations with respect to a given moral value. This allows us to formally characterise moral values, hence grounding the informal definitions that typically appear in the ethics and AI literature [Dignum, 2017]. In philosophical terms, our notion of value judgement function follows a meta-ethical approach: it is a mechanism for deciding whether actions are "good" (supporting the value), or "bad" (detrimental to the value). Furthermore, we also formalise our notion of value system, which unlike the value systems in [Serramia et al., 2018b; Luo et al., 2017; Bench-Capon and Atkinson, 2009], also includes value judgement functions as a core element.

- We establish a formal relationship between norms and moral values based on the notion of value judgement. First, we propose a formal, general characterisation of the properties that a *norm promotion function* ought to satisfy. After that, we propose two particular norm promotion functions.

- We show how to compute the value alignment of a set of norms with a value system by means of a norm promotion function and the preferences over moral values in the value system.

- We introduce the so-called *value-aligned norm selection problem* (VANS) as the optimisation problem of finding a sound norm system maximising value alignment. We also show how to encode the VANS problem as an binary integer program (BIP).

- We analyse our approach. First, we provide a case study to investigate the VANS problem in the realm of the so-called public civility game to illustrate the different norm systems obtained when considering different value systems. Second, we conduct an empirical hardness analysis to investigate the applicability of our approach. Overall, we observe that the structure of norm relationships drives hardness. More precisely, the density of conflicting norms (norms that are incompatible) and general norms (that represent a collection of norms) among the candidate norms drive hardness. However, we show that a state-of-the-art solver like CPLEX [IBM, 1988] can solve medium hardness problem instances with 5000 norms in around 50 seconds.

Although we treat the value-aligned norm selection problem from a theoretical point of view, the framework presented in this chapter has many practical applications. For example, budget allocation in participatory systems [Serramia et al., 2019b] (where given a budget, proposals are accepted or rejected based on their alignment with common moral values), moderation of online communities through norms [Morales et al., 2015c], or value-driven modelling of public policies [Perello-Moragues and Noriega, 2020].

We first framed the problem of selecting the set of norms to enact in a society in [Lopez-Sanchez et al., 2017]. Nevertheless, selection in this work just took into account norm relationships and deployment costs. Subsequently, we advanced towards the consideration of moral values by reformulating the problem as to "choosing the *right* norms to establish" in [Serramia et al., 2018a; Serramia et al., 2018b]. Specifically, in [Serramia et al., 2018a] we proposed moral values as additional (explicit) preference criteria and discussed how norms can be established in new-born or highly dynamic social groups. Then, in [Serramia et al., 2018b] we cast this initial approach as an optimisation problem and studied its empirical hardness. In this chapter we build upon this background and extend it. Firstly, we discuss the philosophical foundations of value-aligned norm selection. Secondly, we use this philosophical basis to better formalise our theoretical approach and computational methods to selecting value-aligned norm systems. Finally, we provide a further detailed empirical study by considering a wider range of decision scenarios and a more fine-grained analysis.

The chapter is structured as follows. Initially, Section 3.2 delves into the ethics and MAS literature to learn how norms and values relate at an abstract level. Then, Section 3.3 introduces an illustrative case study we employ along the chapter. Subsequently, Section 3.4 studies the fundamental relationships between norms, characterises norm systems and introduces the notion of normative domain. From these basic concepts, Section 3.5 first defines the so-called value judgement function together with the value system structure and, then, Section 3.6 specifies norm promotion functions, which characterise how norms promote moral values. Thereafter, Section 3.7 defines the value-aligned norm selection problem and Section 3.8 describes its encoding into a binary integer program. Next, Section 3.9 conducts an empirical analysis to learn the solving times required to find value-aligned norm systems. Finally, Section 3.10 draws conclusions and discusses some limitations of our approach. Recall that, for the ease of readability, we have included a List of Notation and Symbols.

3.2 On norms, values, and norm value alignment

As noted above, one of our core goals in this chapter is to formally define the notion of norm value alignment, or in other words, what it means for a set of norms to be aligned with certain moral values. However, this necessarily requires considering the basic concepts of norms, actions, and values. Based on the literature on ethics and multi-agent systems, the purpose of this section is to analyse how these concepts are related so as to ground the notion of value alignment.

The relationship between norms and actions has been long studied in the literature. On the one hand, ethics is the branch of philosophy that reflects on what is moral[1], right or good [Frankena, 1973; Audi, 1999; Fieser and Dowden, 2021] in order to know what we ought to do [Cooper, 1993] within our society. On the other hand, norms have been widely studied in the multi-agent systems (MAS) literature [Boella et al., 2006; Kollingbaum et al., 2006; Morales et al., 2015b; Dignum, 1999] as a means to achieve coordination by regulating which actions can be performed (permissions), which ones ought to be performed (obligations), and which ones are forbidden (prohibitions).

[1]Morality here refers to the codes of conduct that, given some conditions, would be adopted by all rational people [Gert and Gert, 2020] .

The relationship between the norms enacted in a society and the values that this very same society is aligned with has only recently been addressed in the MAS literature [Wallach and Allen, 2008; Cointe et al., 2016; Tielman et al., 2018; Kasenberg et al., 2018; Mercuur et al., 2019]. However, this relationship is one of the main subjects of research in ethics [Chisholm, 1963; von Wright, 1963; Hansson, 2001; McNamara, 2011]. Within ethics, moral values (also called ethical principles) express the moral objectives *worth striving for* [van de Poel and Royakkers, 2011][2]. Examples of values include justice, happiness and autonomy [Audi, 1999]. Every ethical theory considers one or more moral values that should guide our behaviour [Cooper, 1993]. From these considered values, an ethical theory can prescribe a series of norms as means to realise them [van de Poel and Royakkers, 2011].

Moreover, since norms regulate actions, we need to judge actions ethically in order to determine which norms to prescribe. For that reason, it is argued in [Cointe et al., 2016; Cooper, 1993; Hansson, 2018] that the central theme that unites norms and values is the moral consideration (judgement) of actions. Specifically, an action can be judged as being either good or bad to perform (or skip) with respect to a given moral value [Chisholm, 1963]. This relationship between norms and values being influenced by actions implies that, if a society considers an action to be good to perform from the perspective of a given moral value, then, any norm permitting or obligating such an action would be considered as a norm that *promotes* that value [Cooper, 1993; Hansson, 2018]. Conversely, a norm prohibiting the same action would *demote* that moral value. Classically, a norm is considered to promote a moral value depending on how it regulates an action and how this action is considered with respect to the moral value [Urmson, 1958; Hansson, 2018]: (i) Obligation (if the action is good to perform and bad to skip); (ii) Permission (if the action is good to perform); (iii) Prohibition (if the action is good to skip and bad to perform).

It is clear then that to assess the value alignment of a norm system we must not only consider the relationship between norms and values, but also the ethical dimension of the actions being regulated. Figure 3.2 depicts the relationships that we have identified between norms, values and actions: (i) Norms *regulate* actions; (ii) Moral values *judge* actions; and (iii) Norms *promote/demote* moral values. Figure 3.2 offers a very similar structure to the diagrams in [Cooper, 1993; Hansson, 2018], which also show a relationship between values, norms, and actions.

[2]Moral values are often very high ideals/imperatives that can seldom be achieved

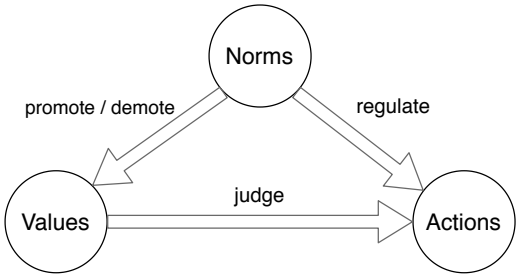


Figure 3.2: Relationship between norms, values and actions.

The formalisation of these relationships (regulation, judgement and promotion/demotion) will provide the foundations for a mathematical definition of value alignment for a norm system. Thus, action regulations are formalised in Section 3.4.2, action judgements are formalised in Section 3.5, and finally value promotion is formalised in Section 3.6. Based on that, we will finally introduce our notion of value alignment for a norm system in Section 3.7.

3.3 Case study: the public civility problem

To illustrate the concepts that will be introduced along this chapter we use the *public civility game*. Recall from Section 1.1, that this game was initially introduced in [Rodriguez-Soto et al., 2020], and provides a scenario through which to explore moral dilemmas. In short, the game represents a situation wherein two agents move daily from their initial positions (which can be their homes) to their respective target destinations (their workplaces, for instance). Along their journey, one of the two agents finds garbage on the floor that prevents it from progressing. Figure 3.3 represents the game scenario. Each agent in the game can deal with the garbage only in one of the following ways:

- By throwing the garbage aside to unblock his way. If the agent throws the garbage at the location where the other agent is, it will hurt the other agent.
- By taking the garbage to the bin. This option is safe for all agents. However, it will impede the progress of the agent performing the action.

perfectly, though this does not preclude from pursuing them.

This scenario can be regulated in different ways depending on the society's preferences. If the value of civility is preferred to the value of timeliness, the regulation will promote picking the garbage to bring it to a bin. Alternatively, if timeliness is preferred, the regulation should allow carelessly throwing the garbage aside disregarding others.

Figure 3.3: Possible initial state of a public civility game. The agent on the left must deal with a garbage obstacle ahead.

In summary, depending on the norms governing the agents, an agent will perform in a way that either promotes the value of civility, the value of timeliness or some combination of the two; and the selected norm system will depend on the value preferences of the regulator. In the following sections we will refer back to the public civility game to illustrate how the preferences of the regulator (the decision maker) in a value system lead to selecting a different norm system.

3.4 Formalising normative domains and norm systems

We first focus on formalising the domains of actions and norms. We call these *action domain* and *normative domain*. As Figure 3.4 shows, norms and actions are related through the regulation relation. We formalise the normative domain based on a given action domain, thus its norms will regulate the actions in the action domain.

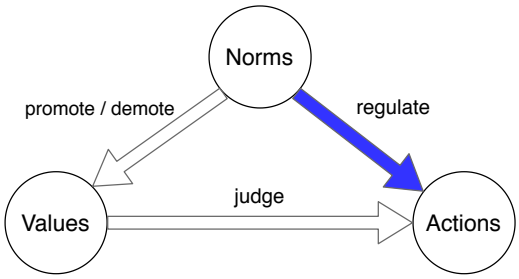


Figure 3.4: Norms regulate actions.

3.4.1 Contextualised actions and action domains

First, we start by considering a multi-agent system composed of a set of agents Ag; a set of actions the agents can perform A; a propositional language $\mathcal{L}$ (with propositions in $\mathcal{P}$ and the logical operator "and"); and a set of states S. Like [Morales et al., 2015b; Morales et al., 2015a], we consider a state transition function that changes the state of the world when agents perform actions. While agents can perform any action in A, only some of the actions in A may be feasible to perform, depending on the state of the multi-agent system. Thus, we refer to a *context* as a subset of the propositions of the language $\varphi \subseteq \mathcal{P}$ describing the conditions that must hold in the state of the multi-agent system for these actions to be feasible. Propositions in the context are connected with *and* semantics. Moreover, it is worth noticing that, since we will ethically judge actions, it is important to also consider the context where they are performed when doing so. For instance, although smoking (action) is blameworthy to preserve public health (value), the context where it takes place dictates how blameworthy it is: smoking in a hospital is more blameworthy, in terms of public health, than smoking at home. Due to the strong link between actions and contexts we consider them together by means of contextualised actions. The norms that we consider in this chapter will aim at regulating such contextualised actions.

Definition 1 (Action)**.** *Given a context $\varphi \subseteq \mathcal{P}$ and an action $a \in A$, we call the tuple (φ, a) a contextualised action, we note a set of contextualised actions as $\mathbb{A} \subseteq \mathcal{P} \times A$.*

Henceforth, since we always consider contextualised actions, we simply call them actions and note them as $a \in \mathbb{A}$.

Example 1. *In the scenario of the public civility game, we consider a propositional language $\mathcal{L}$ with propositions $\mathcal{P} = \{garbage\_in\_front,\ no\_agent\_nearby\}$ and actions $A = \{bring\_to\_bin,\ throw\_nearby\}$. Then, some possible (contextualised) actions are[3]: (i) $bin = (\{garbage\_in\_front\},\ bring\_to\_bin)$; (ii) $thr = (\{garbage\_in\_front\},\ throw\_nearby)$; (iii) $safe = (\{garbage\_in\_front,\ no\_agent\_nearby\},\ throw\_nearby)$. Thus, bin corresponds to the action of bringing garbage to a bin if the agent finds garbage in front; thr represents the action of throwing garbage nearby if the agent finds garbage in front; and finally, safe is the action of throwing garbage nearby if the agent finds garbage in front and knows there are no other agents nearby. As to action safe, notice that our notation for the context $\{garbage\_in\_front, no\_agent\_nearby\}$*

[3]We name these contextualised actions for easy referencing in subsequent examples.

is interpreted under the "and" semantics: both predicates garbage_in_front and no_agent_nearby must hold in the current state.

Actions are often related, in the sense that they may interact with each other. For our purposes, we consider two types of interactions, namely *action incompatibility* (referring to those actions that cannot be performed simultaneously) and *action generalisation* (actions that include other, superfluous actions). Action *incompatibility* is a binary relation $R_i \subseteq \mathbb{A} \times \mathbb{A}$, such that when $(a, a') \in R_i$, we say that a and a' are incompatible, that is when both φ and φ' apply, if the agent performs a, then a' cannot be performed. Note that, R_i is an irreflexive, symmetric, and intransitive relation. The other relation, *action generalisation*, is a binary relation $R_g \subseteq \mathbb{A} \times \mathbb{A}$, where $(a, a') \in R_g$ means that a generalises a' (e.g. when $\varphi \subseteq \varphi'$. We consider generalisations to be atomic steps: if $(a, a') \in R_g$, then $\nexists a^{mid}$, such that $(a, a^{mid}), (a^{mid}, a') \in R_g$. With that in mind, R_g is irreflexive, antisymmetric, and intransitive. Notice that R_g defines *direct* generalisations between two actions. Based on this relation, we can capture the notion of *indirect* generalisation through the so-called *ancestors* and *successors* of an action. Given two actions, a',$a \in \mathbb{A}$, we say that a is an ancestor of a' (and that a' is a successor of a) if there is a subset of actions $\{a^1, \ldots, a^p\} \subseteq \mathbb{A}$ such that $(a^1, a^2), \ldots, (a^{p-1}, a^p) \in R_g$, $a^1 = a$, and $a^p = a'$. Henceforth, given an action $a \in \mathbb{A}$, we will note its ancestors as $A(a)$ and its successors as $S(a)$. Contextualised actions and their relations form an action domain, this structure represents the basis upon which we build norms to regulate action performance, action relations will then dictate how norms are related.

Definition 2 (Action domain)**.** *An action domain is a tuple $\langle \mathbb{A}, R \rangle$, where $\mathbb{A}$ is a set of actions; and $R = \{R_i, R_g\}$ is a set of action relations, where R_i is a set of incompatibility relations and R_g a set of generalisation relations.*

Example 2. *We build an action domain following Example 1, firstly we consider the set of actions $\mathbb{A} = \{bin, thr, safe\}$. Secondly, in terms of action relations, note that the actions of bringing the garbage to the bin and throwing it nearby are incompatible because the agent has to do one or the other. On the other hand, the action of throwing the garbage nearby generalises the action of throwing it nearby if there is no agent nearby. Thus, the action relations would be $R_i = \{(bin, thr), (bin, safe)\}$ and $R_g = \{(thr, safe)\}$.*

Throughout our work, we assume that the decision maker has sufficient knowledge to compose the action domain correctly. In particular, note that the generalisation relation should not have cycles.

The action domain represents one of the three components in Figure 3.2 and the basis upon which we will build regulation. The next step is to provide the normative component, which we call the *normative domain*.

3.4.2 The normative domain

Agents will typically have many actions to perform. We use norms to regulate which actions they must, should, or must not perform. Action relations induce norm relations, in this section we define the structure containing norms and their relations, the normative domain. Firstly, we tackle the norm formalisation. Although norms have been extensively studied in the multi-agent systems literature as a means to regulate agent behaviour, [Boella et al., 2006; Dignum, 1999; Andrighetto et al., 2013], there is no consensus on their formal definition. Here our notion of norm is based on a simplification of the one in [Morales et al., 2015a]. Our notion of norm establishes obligations, permissions, and/or prohibitions [Meyer and Wieringa, 1993] of agent's actions. Formally,

Definition 3 (Norm)**.** *Given an action domain $\langle \mathbb{A}, R \rangle$, a norm is given by an expression of the form $\theta(a)$, where $a \in \mathbb{A}$ is an action and $\theta \in \{Obl, Per, Prh\}$ is a deontic operator.*[4] *We will let N denote the set of norms.*

Obligations and permissions allow the performance of actions, whereas prohibitions forbid them. We capture this distinctive feature of norms by means of the sign of the norm, that is the function $sgn : N \to \{-1, 1\}$ defined as:

$$sgn(n) = \begin{cases} 1 & \text{if } \theta \in \{Obl, Per\} \\ -1 & \text{if } \theta = Prh \end{cases} \tag{3.1}$$

Example 3. *From the action domain in Example 2, we consider the following norms: $Per(thr)$, $Obl(thr)$, $Prh(thr)$, $Per(safe)$, $Obl(safe)$, $Per(bin)$, $Obl(bin)$. Moreover, $sgn(Per(thr)) = 1$, $sgn(Obl(thr)) = 1$, and $sgn(Prh(thr)) = -1$.*

Since the decision maker has knowledge of the action domain, we rely on them to provide candidate norms to regulate a multi-agent system. We

[4]Note that, since $a \in \mathbb{A}$ contains both the action and the context $\varphi \subseteq \mathcal{P}$ our definition of norm is equivalent to the usual definition $\langle \varphi, \theta(a) \rangle$.

assume that norms at hand are considered beneficial (for example, because they align with the goal of the decision maker). We aim at selecting the best norms out of these candidate norms. Note though that, since norms regulate actions and actions are related, these relations actually extend to norms. Norm relations have been previously studied in the literature. Thus, for example, [Grossi and Dignum, 2005] studies the relation between abstract and concrete norms, whereas [Kollingbaum et al., 2006; Vasconcelos et al., 2009] focus on norm conflicts —and solve them based on first-order unification and constraint solving techniques. Here, we induce norm relations from action relations, therefore we consider *norm incompatibility* and *norm generalisation*.

Informally, we say that two norms are incompatible when they cannot be enacted at once. For example, two norms that allow incompatible actions are considered incompatible norms. Formally:

Definition 4. *We say norms $n = \theta(a)$, $n' = \theta'(a')$ are incompatible iff either:*

- *The actions are incompatible $(a, a') \in R_i$ and neither of them is forbidden, namely $sgn(n) = sgn(n') = 1$, or*
- *Either an action is more general than the other ($(a, a') \in R_g$) or they are the same action ($a = a'$) and one action is forbidden and the other is not, namely $sgn(n) \neq sgn(n')$.*

We note as $\mathfrak{R}_i$ the norm incompatibility relation.

As for norm generalisation, a norm generalises another if enacting the general norm deems the other one redundant. This happens when two norms both oblige, permit or prohibit actions that are one more general than the other. Also, obliging an action deems the permission redundant, because obliging an action implies permitting it. Thus, we formalise norm generalisation as:

Definition 5. *We say that norm $n = \theta(a)$ generalises $n' = \theta'(a')$ iff either:*

- *$(a, a') \in R_g$ and $\theta = \theta'$ or $\theta = Obl$ and $\theta = Per$.*
- *The norms regulate the same action $a = a'$, the first is an obligation and the second a permission $\theta = Obl$ and $\theta = Per$.*

We note as $\mathfrak{R}_g$ the norm generalisation relation. Abusing notation of action's ancestors and successors, we note the ancestors and successors of n as $A(n)$, $S(n)$

Proposition 1. *$\mathfrak{R}_i$ is an irreflexive, symmetric, and intransitive relation and $\mathfrak{R}_g$ is an irreflexive, anti-symmetric, and intransitive relation.*

Proof (Proposition 1)**.** *Immediate from the respective definitions.* □

Having formalised norms and their relations, we now formalise the structure representing them and which we will use to perform norm selection: the normative domain.

Definition 6 (Normative domain)**.** *A normative domain is a tuple $\langle D, N, \mathfrak{R} \rangle$ such that:*

- *D is an action domain with well-defined action relations;*
- *N is a set of candidate norms regulating actions in D; and*
- *$\mathfrak{R} = \{\mathfrak{R}_i, \mathfrak{R}_g\}$ is a set of norm relations over N.*

Example 4. *We build the normative domain with the action domain in Example 2 and the set of candidate norms in Example 3 and the resulting norm relations (applying Definitions 4 and 5) represented in Figure 3.5.*

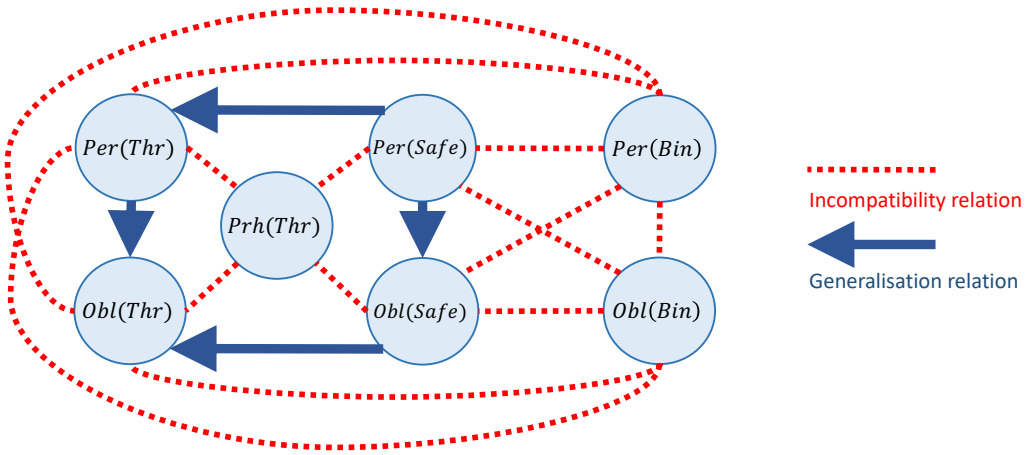


Figure 3.5: Example normative domain of the public civility game.

3.4.3 Characterising norm systems

The goal of the process depicted in Figure 3.1 is to obtain a norm system. Put simply, we will refer to any subset of the norms in a normative domain as a *norm system*.

Definition 7 (Norm system)**.** *Given a normative domain $\langle D, N, \mathfrak{R}\rangle$, any subset of the norms in the normative domain $\Omega \subseteq N$ is a norm system.*

Since norm systems are just subsets of candidate norms, a norm system can contain incompatible norms or redundant norms (due to generalisation relationships). Thus, when selecting norms we desire that the resulting norm system does not contain incompatible nor redundant norms.

Definition 8 (Compatible norm system)**.** *Given a normative domain $\langle D, N, \mathfrak{R}\rangle$, we say that a norm system $\Omega \subseteq N$ is compatible iff for each $n_i, n_j \in \Omega$, $(n_i, n_j) \notin \mathfrak{R}_i$.*

Definition 9 (Non-redundant norm system)**.** *Given a normative domain $\langle D, N, \mathfrak{R}\rangle$, a norm system $\Omega \subseteq N$ is non-redundant iff for each $n_i, n_j \in \Omega$, $n_j \notin A(n_i)$ and $n_j \notin S(n_i)$, where $A(n_i)$ are the ancestors of n_i and $S(n_i)$ its successors.*

Definition 10 (Sound norm system)**.** *Given a normative domain $\langle D, N, \mathfrak{R}\rangle$, we say that a norm system $\Omega \subseteq N$ is sound iff it is both compatible and non-redundant.*

Example 5. *The set $\{Per(thr),\ Per(bin),\ Prh(thr)\}$ is a non-redundant norm system, but it is not compatible because $(Per(thr),\ Per(bin))$, $(Per(thr),\ Prh(thr)) \in \mathfrak{R}_i$. On the other hand, $\{Per(thr),\ Per(safe)\}$ is a compatible norm system but it is redundant because $(Per(thr),\ Per(safe)) \in \mathfrak{R}_g$. Finally, $\{Per(bin),\ Prh(thr)\}$, $\{Obl(bin),\ Prh(thr)\}$, $\{Per(thr),\ Obl(safe)\}$ and the singletons containing each one of the norms are all sound norm systems in this normative domain.*

To be more precise, the goal of the process in Figure 3.1 is to yield a particular type of sound norm system, namely one that is aligned with the moral values specified by the decision maker.

3.5 Value-based judgement of actions

As introduced in Section 3.2, ethics is the branch of philosophy that reflects on what is moral[1], right or good [Frankena, 1973; Audi, 1999; van de Poel

and Royakkers, 2011]. The philosophical discipline of ethics is eminently practical because we do not want to know what is good or bad out of mere curiosity, but because we want to know what we ought to do [Cooper, 1993; Wallach and Allen, 2008]. To answer that question, the field of normative ethics is devoted to prescribe us what is the correct action to do at any given situation [Fieser and Dowden, 2021]. Of course, we cannot provide guidelines on how to do the good if we do not first define what is good to begin with. This and other foundational problems are the subject of the field of metaethics, which attempts to clarify the ethical methodology and terminology [Beauchamp and Childress, 2009].

Within ethics, moral values (also called ethical principles) bridge normative ethics and metaethics. In the AI literature, values are seen as criteria to discern which actions are right and which are wrong [Charisi et al., 2017; Dignum, 2017]. Examples of values include justice, happiness and autonomy. We formally characterise moral values following these informal definitions. As shown in Figure 3.6, values and actions are related. Specifically, a value judges the extent to which the performance (or non-performance) of actions is beneficial or detrimental. Thus, we formally characterise moral values through their judgement of actions as follows.

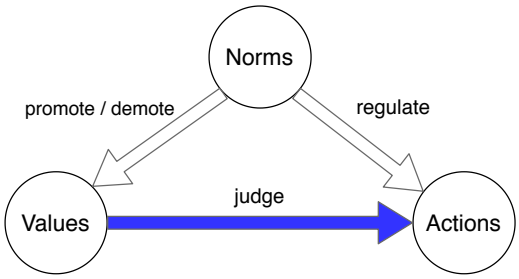


Figure 3.6: Values judge actions

Definition 11 (Moral value)**.** *We characterise a moral value through a pair of value judgement functions* $v = \langle \alpha_v^+, \alpha_v^- \rangle$*. Given a set of actions* $\mathbb{A}$*, each of these functions takes an action and returns its evaluation* $\alpha_v^+, \alpha_v^- : \mathbb{A} \to [-1, 1]$*. Function* α_v^+ *evaluates the praiseworthiness of performing the action, while* $\alpha_v^-(a)$ *evaluates the praiseworthiness of not performing the action*[5]*. These evaluations are real numbers in the interval* $[-1, 1]$*: a positive number means that the moral value is being promoted, whereas a negative*

[5]Note that, $\alpha_v^+(a)$ and $\alpha_v^-(a)$ are independent so $\alpha_v^-(a)$ is not necessarily equal to $-\alpha_v^+(a)$.

one stands for demotion. We require that an action cannot be praiseworthy (or blameworthy) both to perform and to skip with respect to the same moral value. Thus, for a moral value to be well-defined, its value judgement functions have to satisfy:

$$\alpha_v^+(a) \cdot \alpha_v^-(a) \leq 0, \forall a \in A \tag{3.2}$$

Value judgement functions allow us to quantify the moral praiseworthiness of performing/skipping actions. Note that the condition in Equation 3.2 dictates that, if an action is praiseworthy to perform, then it must be either blameworthy or neutral to skip. Similarly, if the action is blameworthy to perform, it then must be praiseworthy or neutral to skip. Overall, these value judgement functions within our characterisation of moral values allows us to adhere to previous literature [Charisi et al., 2017; Dignum, 2017] and use moral values as criteria for discerning right (praiseworthiness) from wrong (blameworthiness).

Example 6. *We judge the actions of the action domain in Example 2 with respect to two values: civility* $Civ = \langle \alpha_{civ}^+, \alpha_{civ}^- \rangle$*; and timeliness* $Tim = \langle \alpha_{tim}^+, \alpha_{tim}^- \rangle$*. In terms of civility, the action of bringing garbage to a bin is highly praiseworthy to perform, but neutral to skip since the garbage is not the agent's property. In terms of timeliness though, the action is slightly blameworthy to perform as it will take time to go to the bin and slightly praiseworthy to skip. Thus, the judgement functions of both moral values for the action bin may, for instance, be as follows:*

$$\alpha_{civ}^+(bin) = 1 \quad \alpha_{civ}^-(bin) = 0 \quad \alpha_{tim}^+(bin) = -0.5 \quad \alpha_{tim}^-(bin) = 0.5$$

Regarding civility, throwing the garbage aside is blameworthy to perform and praiseworthy to skip as there are chances of hitting other agents. On the other hand, throwing the garbage aside saves time because it frees the agent's path towards the target. Thus, this action is highly praiseworthy to perform and highly blameworthy to skip in terms of timeliness. Therefore, the judgement functions for thr could be defined as:

$$\alpha_{civ}^+(thr) = -0.8 \quad \alpha_{civ}^-(thr) = 0.8 \quad \alpha_{tim}^+(thr) = 1 \quad \alpha_{tim}^-(thr) = -1$$

Finally, we judge the safe action, which refers to throwing garbage nearby (i.e., clearing the path of the agent that encountered it) when there are no other agents nearby. In terms of civility, we consider safe to be praiseworthy to perform and neutral to skip, as we do for bin. However, safe is

slightly less praiseworthy than bin, since bin is the only action that actually removes the garbage from the street. As for the timeliness moral value, safe is much faster to perform than bin but still takes more time than thr, since it requires checking that no agent is nearby. Thus, assuming it is slightly praiseworthy to perform and slightly blameworthy to skip, we can define the value judgement functions as follows:

$$\alpha^{+}_{civ}(safe) = 0.8 \quad \alpha^{-}_{civ}(safe) = 0 \quad \alpha^{+}_{tim}(safe) = 0.5 \quad \alpha^{-}_{tim}(safe) = -0.5$$

Ethical reasoning typically involves not a single moral value, but multiple moral values along with value preferences [Bench-Capon and Atkinson, 2009; Luo et al., 2017; Serramia et al., 2018b] conforming a *value system*. Value systems can be individual or shared by a society. In this work we will suppose we know the society's value system in order to select norms accordingly. Recall that in Figure 3.1 we considered a value system as one of the two main inputs of our value-aligned norm system engineering process. As depicted in Figure 3.6, values judge actions via their judgement functions. As these value judgement functions characterise moral values, they also implicitly constitute an integral part of the value system, which is explicitly composed by the values and their preferences.

Definition 12 (Value system)**.** *A value system is a tuple $\langle V, \succeq \rangle$, where: V stands for a non-empty set of moral values; and $\succeq$ is a ranking*[6] *over the moral values in V. If $v \succeq v'$ we say that v is more preferred than v', and if also $v' \succeq v$ we say that v and v' are indifferently preferred, and note it as $v \sim v'$.*

Notice that unlike our value system, the definitions in [Bench-Capon and Atkinson, 2009] and [Serramia et al., 2018b] do not consider the link between values and actions. Moreover, although [Luo et al., 2017] considers the relation between actions and values, it does not quantify it. Furthermore, in terms of the ordering structure used, we favour rankings as they are more flexible than the total orders used in [Luo et al., 2017] and [Bench-Capon and Atkinson, 2009], though they are stricter than the partial order used in [Serramia et al., 2018b]. We do so because partial orders would require us to make arbitrary assumptions when values are not related (in the order).

Example 7. *The values of* civility *and* timeliness*, $V = \{Civ, Tim\}$, together with the ranking $Civ \succeq Tim$, constitute a value system.*

[6] In particular, a ranking is irreflexive, transitive and total. Note that, being irreflexive and transitive, this relation disallows the existence of cycles over preferences: $\nexists v_1, \ldots, v_k$, s.t. $v_1 \succeq \cdots \succeq v_k \succeq v_1$ and $v_1 \nsim \cdots \nsim v_k \nsim v_1$.

3.6 Promotion of moral values through norms

Once established the relation between actions and values, as well as our formal definition of value system, we now focus on the relation between norms and values. Recalling the relations triangle (see Figure 3.7), norms promote values: we capture this relationship by means of the so-called *norm promotion function*. Specifically, this norm promotion function evaluates how much each norm promotes each value, taking into account the norm's deontic operator and the praiseworthiness of its regulated action. In this section, we first characterise the properties the norm promotion function ought to satisfy and then propose two alternative functions.

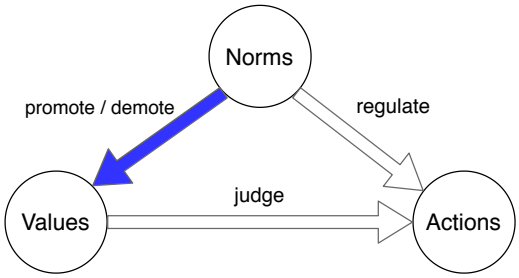


Figure 3.7: Norms promote/demote values.

3.6.1 Characterising norm promotion

We require that a norm promotion function will satisfy the following key properties:

- **Deontic and judgement dependency**: The promotion function only depends on the deontic operator of the norm and the value judgement of its regulated action with respect to the value.
- **Deontic coherence**: Norms that regulate the same action but have different deontic operators should have coherent promotions. For example, if permitting an action promotes a value, prohibiting the action should demote it.
- **Coherence (or correlation) with value judgements**: Norm promotion and value judgement must be aligned. This property is divided into three cases:

- **Neutrality**: If an action is neutral to a value, norms regulating the action should also be neutral to the value.
- **Praiseworthiness pursuit**: If an action is praiseworthy to a value, permitting or obliging the action should promote the value, while prohibiting the action should demote the value.
- **Blameworthiness avoidance**: If an action is blameworthy to a value, permitting or obliging the action should demote the value, while prohibiting the action should promote the value.

Formally, we include these requirements in the promotion function definition:

Definition 13 (Promotion function)**.** *Let $\langle D, N, \mathfrak{R}\rangle$ be a normative domain with candidate norms N over the actions in $\mathbb{A}$, $\langle V, \succeq\rangle$ a value system, and $\pi : V \times N \to [-1, 1]$, a function over pairs of values and norms. We say that π is a promotion function if it satisfies the conditions below. If so, we will say that π assigns a degree of promotion (when positive) or demotion (when negative) from the norms in N to the values in V. Thus, $\pi(v, n)$ is the degree of promotion/demotion of n to v. A promotion function must satisfy the following properties:*

- ***Deontic and judgement dependency**: The promotion function is broken into three continuous cases π^{Obl}, π^{Per}, and π^{Prh} relating to the three deontic operators. Furthermore, for each of these cases, given a norm and a value, the promotion degree solely depends on the value judgement over the action regulated by the norm. Hence, $\pi^{\theta} : [-1, 1] \times [-1, 1] \to [-1, 1]$, where: $\theta \in \{Obl, Per, Prh\}$; $[-1, 1] \times [-1, 1]$ is the domain of a vector containing both value judgements of the regulated action (α_v^+, α_v^-); and the result is in $[-1, 1]$, where -1 means total norm demotion and 1 total norm promotion of v.*

$$\pi(v, n) = \begin{cases} \pi^{Obl}(\alpha_v^+(a), \alpha_v^-(a)) & \text{if } \theta = Obl, \\ \pi^{Per}(\alpha_v^+(a), \alpha_v^-(a)) & \text{if } \theta = Per, \\ \pi^{Prh}(\alpha_v^+(a), \alpha_v^-(a)) & \text{if } \theta = Prh, \end{cases}$$

- ***Deontic coherence**: The promotion function should be coherent in terms of deontic operators, that is, for a given action, if the norm permitting it promotes (demotes) a value, then the norm obliging it should also promote (demote) the value. On the other hand, if the norm obliging the action promotes (demotes) the value, then the norm prohibiting*

the action should demote (promote) it. The following conditions must hold:

- *The promotion degrees of a norm permitting an action and of a norm obligating the very same action must have the same sign, one of them be 0, or both of them be 0, namely* $\pi(v, Per(a)) \cdot \pi(v, Obl(a)) \geq 0$.
- *The promotion degree of a norm prohibiting an action and of a norm obliging the very same action must have different signs, one of them be 0, or both of them be 0 (i.e., their signs cannot be both positive nor both negative), namely* $\pi(v, Obl(a)) \cdot \pi(v, Prh(a)) \leq 0$.

- ***Neutrality****: If an action is neutral to a value, then the promotion of a norm regulating the action to that value is neutral. Formally, given* $n = \theta(a)$ *then if* $\alpha_v^+(a) = \alpha_v^-(a) = 0 \Rightarrow \pi(v, n) = 0$.

- ***Praiseworthiness pursuit****: Consider* $a \in \mathbb{A}$*, an action that is praiseworthy to perform (*$\alpha_v^+(a) > 0$*). The promotion degree of norms permitting or obliging the action must be positive, while it must be negative for those norms prohibiting it. Formally,* $\pi(v, Prh(a)) \leq 0$*;* $\pi(v, Per(a)) \geq 0$ *and* $\pi(v, Obl(a)) \geq 0$.

- ***Blameworthiness avoidance****: Consider* $a \in \mathbb{A}$*, an action that is blameworthy to perform (*$\alpha_v^+(a) < 0$*). The promotion degree for a norm prohibiting the action must be positive, while it must be negative for those norms permitting or obliging the action. Formally,* $\pi(v, Prh(a)) \geq 0$*;* $\pi(v, Per(a)) \leq 0$ *and* $\pi(v, Obl(a)) \leq 0$*. Notice that in fact it is worse to oblige the action than to permit it, therefore we also require* $\pi(v, Per(a)) \geq \pi(v, Obl(a))$ *(we allow equality to encompass the case of actions that are so damaging to the value that both permitting or obliging them should have promotion -1).*

3.6.2 Defining norm promotion functions

Considering the characterisation of the family of norm promotion functions in Definition 13, this work proposes two example norm promotion functions. A simple norm promotion function called *base norm promotion function* and a more complex function called *supererogatory norm promotion function*.

The base promotion function

First, our aim is to define a linear norm promotion function that can be both readily used as a simple promotion function or can be used as the basis to create other more complex promotion functions. The rationale behind its design is that obligations will promote the value proportionally (increasing linearly) to the praiseworthiness to perform – and blameworthiness to skip – their regulated action. Conversely, the more blameworthy a regulated action is to perform – and praiseworthy to skip –, the more a prohibition norm will promote the corresponding value.

Finally, the promotion of permissions must be between that of obligations and that of prohibitions whilst having the same sign as that of obligations (due to deontic coherence). Therefore, we assess the promotion of permitting an action as a fraction $\epsilon \in [0, 1]$ of the promotion of obliging it. Although establishing this fraction remains a task of the decision maker, it is worth noticing that ϵ values close to 1 will favour the selection of permission norms, whereas ϵ values close to 0 will favour obligations. Thus, we define the base promotion function as follows:

Definition 14 (Base promotion function)**.** *Given a normative domain with a set of candidate norms N over the actions in $\mathbb{A}$ and a value v with value judgement functions α_v^+ and α_v^-, we define $\pi_{base} : V \times N \to [-1, 1]$, such that for a value $v \in V$ and a norm $n = \theta(a) \in N$:*

$$\pi_{base}(v, n) = \begin{cases} \frac{\alpha_v^+(a) - \alpha_v^-(a)}{2} & \text{if } \theta = Obl, \\ \epsilon \cdot \frac{\alpha_v^+(a) - \alpha_v^-(a)}{2} & \text{if } \theta = Per, \\ \frac{-\alpha_v^+(a) + \alpha_v^-(a)}{2} & \text{if } \theta = Prh, \end{cases} \tag{3.3}$$

where $\epsilon \in [0, 1]$. Note that ϵ ranges from $\epsilon = 0$, meaning that permissions always have 0 promotion (thus, they are disregarded), to $\epsilon = 1$, meaning that permissions have the same promotion as obligations.

Theorem 1. *π_{base} is a promotion function.*

Proof (Theorem 1)**.** *Deontic and judgement dependency holds as the function is defined in three cases as required and is defined in $[-1, 1]$. Deontic coherence holds because, $\frac{\alpha_v^+(a) - \alpha_v^-(a)}{2} \cdot \frac{\alpha_v^+(a) - \alpha_v^-(a)}{2} \geq 0$ always and $\epsilon \in [0, 1]$, hence in the case of an obligation and a permission $\pi(v, Obl(a)) \cdot \pi(v, Per(a)) = \frac{\alpha_v^+(a) - \alpha_v^-(a)}{2} \cdot \epsilon \frac{\alpha_v^+(a) - \alpha_v^-(a)}{2} \geq 0$, and in the case of an obligation and a prohibition we have that $\operatorname{sgn}(\frac{\alpha_v^+(a) - \alpha_v^-(a)}{2}) = -\operatorname{sgn}(-\frac{\alpha_v^+(a) - \alpha_v^-(a)}{2})$.*

Hence, it follows that $\pi(v, Obl(a)) \cdot \pi(v, Prh(a)) \leq 0$. *Neutrality is satisfied because, if* $\alpha_v^+(a) = \alpha_v^-(a) = 0$, *then* $\frac{\alpha_v^+(a)-\alpha_v^-(a)}{2} = \epsilon\frac{\alpha_v^+(a)-\alpha_v^-(a)}{2} = -\frac{\alpha_v^+(a)-\alpha_v^-(a)}{2} = 0$. *When it comes to praiseworthiness pursuit, if* $\alpha_v^+(a) > 0$, *due to Equation 3.2 we have* $\alpha_v^-(a) \leq 0$, *therefore* $\frac{\alpha_v^+(a)-\alpha_v^-(a)}{2} \geq \epsilon\frac{\alpha_v^+(a)-\alpha_v^-(a)}{2} \geq 0$ *and* $-\frac{\alpha_v^+(a)-\alpha_v^-(a)}{2} \leq 0$, *so praiseworthiness pursuit holds. Conversely, if* $\alpha_v^+(a) < 0$, *then* $\alpha_v^-(a) \geq 0$, *and therefore* $0 \geq \epsilon\frac{\alpha_v^+(a)-\alpha_v^-(a)}{2} \geq \frac{\alpha_v^+(a)-\alpha_v^-(a)}{2}$, *and* $-\frac{\alpha_v^+(a)-\alpha_v^-(a)}{2} \geq 0$, *which proves blameworthiness avoidance.*□

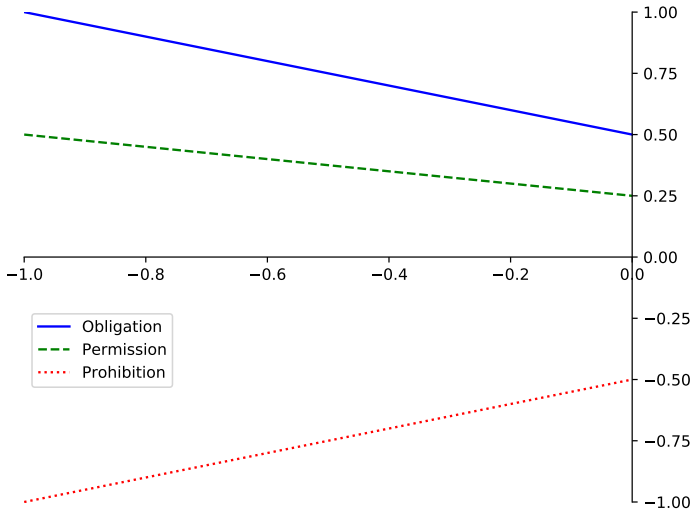


Figure 3.8: Base promotion function π_{base} for a fully praiseworthy action $\alpha_v^+(a) = 1$ and $\epsilon = 0.5$. x-axis (from 0 to -1): blameworthiness to skip action $\alpha_v^-(a)$. y-axis: value promotion degree of obligation, permission, and prohibition of the action.

Figure 3.8 shows an example of the base promotion function π_{base} for an action that is totally praiseworthy to perform ($\alpha_v^+(a) = 1$) and $\epsilon = 0.5$. The x-axis represents the range of possible blameworthiness of skipping the action $\alpha_v^-(a)$ and goes from -1 in the left to 0 in the right (being a praiseworthy action, positive values of $\alpha_v^-(a)$ cannot be considered due to Eq. 3.2). The three lines represent the promotion degrees (in the y-axis from -1 to 1) for the three possible norms regulating the action: obligation, permission and prohibition. Note that, since the action is praiseworthy, prohibiting it always results in a negative promotion, while permitting or obliging it always implies a positive promotion. In particular, obliging this praiseworthy

action has always greater promotion than permitting it independently of how blameworthy it is to skip.

In addition to the 2D representation of the base norm promotion function π_{base} in Figure 3.8, we can further inspect π_{base} in the 3D space. We do so in Figure 3.9, which depicts Obligations, Permissions, and Prohibitions for $\epsilon = 0.5$. In each of these cases, π_{base} is a two-variable function depending on possible different values of $\alpha_v^+(a)$ and $\alpha_v^-(a)$. Because of the definition of the moral value's judgement functions (see Equation 3.2), the promotion function is only defined when value judgements are of opposite sign or zero, thus $\alpha_v^+(a) \cdot \alpha_v^-(a) \leq 0$ must hold.

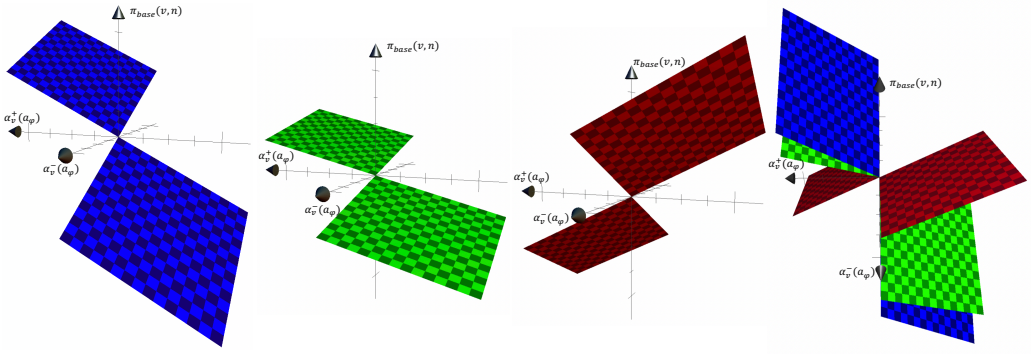
The surface in Figure 3.9a represents the promotion function for obligations. It is a plane that is positive for actions praiseworthy to perform and blameworthy to skip, $\alpha_v^+(a) > 0$ and $\alpha_v^-(a) < 0$. On the other hand, the plane is negative for actions blameworthy to perform and praiseworthy to skip, $\alpha_v^+(a) < 0$ and $\alpha_v^-(a) > 0$. Note that in particular, given a value v and a norm $n = Obl(a)$:

- If $\alpha_v^+(a) = 1$ and $\alpha_v^-(a) = -1$, then it has maximum promotion, $\pi_{base}(v, n) = 1$
- If $\alpha_v^+(a) = -1$ and $\alpha_v^-(a) = 1$ then it has maximum negative promotion (or demotion), $\pi_{base}(v, n) = -1$.
- If $\alpha_v^+(a) = 0$ and $\alpha_v^-(a) = 0$ then it is neutral to the value, $\pi_{base}(v, n) = 0$.

On the other hand, the surface in Figure 3.9c represents the promotion degrees for prohibitions. Note that, the promotion takes the opposite value than the promotion function does for obligations. Thus, given $n = Prh(a)$, $n' = Obl(a)$, then $\pi_{base}(v, n) = -\pi_{base}(v, n')$.

Finally, the surface in Figure 3.9b represents the promotion function for permissions. Notice that the formula of this surface is the same as the one for obligations, but scaled by $\epsilon \in [0, 1]$ (in this case $\epsilon = 0.5$). Thus, the promotion degree for a permission will be lower than for an obligation when $\alpha_v^+(a) > 0$ and $\alpha_v^-(a) < 0$, whereas it will be larger when $\alpha_v^+(a) < 0$ and $\alpha_v^-(a) > 0$, as shown in the combined plot of all cases in Figure 3.9d. Note that, ϵ marks the upper bound of the promotion function when evaluating permissions (and $-\epsilon$ the lower bound), that is, $\forall n = Per(a) \in N$, $\pi_{base}(v, n) \in [-\epsilon, \epsilon]$. Therefore, a smaller ϵ must be used in cases where the decision maker prefers enforcing norms (obligations and prohibitions), while

a larger ϵ must be used if the decision maker wants to set larger promotion degrees for permissions.



(a) Obligations (b) Permissions (c) Prohibitions (d) Combination

Figure 3.9: Plots of each of the cases of the base promotion function using $\epsilon = 0.5$. All axis represent values in $[-1, 1]$, the arrow on all axis marks point 1.

The supererogatory promotion function

Consider a, a praiseworthy action to perform ($\alpha_v^+(a) > 0$), but not very harmful to skip ($\alpha_v^-(a)$ close or equal to 0). The ethics literature refer to these actions as *supererogatory* [Urmson, 1958; Chisholm, 1963; Montague, 1989; Horgan and Timmons, 2010; Hansson, 2013]. Using the base promotion function, a norm obliging a would have larger (if $\epsilon \neq 1$) or equal (if $\epsilon = 1$) promotion degree than a norm permitting it. Thus, using this base promotion function may seem to imply that obliging this action is better than permitting it. However, this is not desirable since, for example, although taking care of a piece of garbage found on the street is a good thing to do, we would hardly expect it to be compulsory. Indeed, as noted in [Heyd, 2019], an ethical system would be impoverished if it just considered obligations (duties).

Since deontic logic limits norms to obligations, permissions and prohibitions, when considering supererogation, we advocate for regulating supererogatory actions as permissions but associating them the semantics of recommendations.

To further illustrate supererogation consider, for instance, the moral value of solidarity. In this context, giving money to a charity represents a paradigmatic example of a supererogatory action, since giving money to a charity is highly praiseworthy but skipping it cannot be considered to be

particularly harmful (it can be somehow considered as neutral whereas stealing money from the charity would certainly be harmful). Therefore, since we cannot be forced to give money to charities, we advocate for regulating it as permission with the associated semantics of recommendation.

Overall, considering the nature of supererogation, we aim at ensuring that permissions of supererogatory actions have larger promotion degrees than those norms obliging them. Formally, we define a supererogatory promotion function as follows.

Definition 15. *In a normative domain, given a set of candidate norms N over the actions in $\mathbb{A}$ and a value system $\langle V, \succeq \rangle$, we say that a promotion function π is supererogatory if the following conditions hold:*

(C1) If a is praiseworthy to perform ($\alpha_v^+(a) > 0$) and neutral to skip ($\alpha_v^-(a) = 0$), we consider it a supererogatory action and assign positive promotion degree to a norm permitting it, but 0 promotion degree to a norm obliging it ($\pi(v, Per(a)) > \pi(v, Obl(a)) = 0$).

(C2) For each action a that is praiseworthy to perform ($\alpha_v^+(a) > 0$) and totally blameworthy to skip ($\alpha_v^-(a) = -1$), the promotion degree of the norm obliging it has to be greater than the promotion of the norm permitting it, and both have to be positive ($\pi(v, Obl(a)) > \pi(v, Per(a) > 0$).

Suppose we have a value v and a praiseworthy action to perform a. Hence, following Eq. 3.2 we know that $\alpha_v^+(a) > 0$ and $\alpha_v^-(a) \leq 0$. The first condition (C1) in Definition 15 demands that if action a is totally supererogatory ($\alpha_v^-(a) = 0$), then a norm making it obligatory to do the action has zero promotion value, whereas a norm permitting the action has positive promotion value. The second condition (C2) in Definition 15 states that if a is clearly not supererogatory ($\alpha_v^-(a) = -1$), obliging to perform the action has more promotion value than permitting it. Anyhow, both obliging and permitting the action have positive promotion values. These conditions ensure[7] that the domain of $\alpha_v^-(a)$ can be split into supererogatory values and non-supererogatory values. More precisely, there exists a threshold value $\beta \in [-1, 0]$ such that: (i) if $\alpha_v^-(a) \in [\beta, 0]$, action a is supererogatory, and; (ii) if $\alpha_v^-(a) \in [-1, \beta)$, action a is not supererogatory.

[7] By applying Bolzano's theorem considering the two conditions and the continuity of π^{Obl} and π^{Per}.

Now, to handle supererogatory actions, we must give greater promotion value to norms permitting them than to norms obliging them. This is captured by the supererogatory promotion function that we introduce next as an extension of the base promotion function presented above.

Definition 16 (Supererogatory promotion function)**.** *Given a value system* $\langle V, \succeq \rangle$ *and the base promotion function* π_{base} *with* $\epsilon \in (0,1)$[8]*, we define the supererogatory promotion function* π_{sup} *as:*

$$\pi_{sup}(v,n) = \begin{cases} -\alpha_v^-(a) \cdot \pi_{base}(v,n) & \text{if } \theta = Obl,\ \alpha_v^+(a) \geq 0,\ \text{and } \alpha_v^-(a) \leq 0 \\ \pi_{base}(v,n) & \text{otherwise} \end{cases} \tag{3.4}$$

Note that π_{sup} is based largely in π_{base}. In fact, for non-supererogatory actions ($\alpha_v^-(a)$ closer to -1), the promotion value of π_{sup} is similar or equal to that obtained with π_{base}. The difference comes with supererogatory praiseworthy actions ($\alpha_v^+(a) \geq 0$, $\alpha_v^-(a)$ close to 0). In this case, the promotion of a norm obliging it is close to zero, while a norm permitting the action will have greater promotion.

Note that the change of preference between obligations and permissions happens when $\epsilon \frac{\alpha_v^+(a) - \alpha_v^-(a)}{2} > -\alpha_v^-(a) \frac{\alpha_v^+(a) - \alpha_v^-(a)}{2} \Leftrightarrow \alpha_v^-(a) > -\epsilon$

Therefore, in the supererogatory promotion function ϵ not only allows us to give more promotion degree to permissions, but also marks the boundary of supererogation. When an action is less harmful to skip than $-\epsilon$ we will consider it supererogatory and prefer the permission over the obligation.

Theorem 2. *π_{sup} is a supererogatory promotion function*

Proof (Theorem 2)**.** *We first prove that π_{sup} is a promotion function. We have already proved π_{base} is a promotion function. For π_{sup}, deontic and judgement dependency holds as the function defines three cases (those of π_{base}) depending on the deontic operator and is defined in $[-1,1]$. Deontic coherence holds because it holds for π_{base} and in case $\alpha_v^+(a) \geq 0$ and $\alpha_v^-(a) \leq 0$, we have that $-\alpha_v^-(a) \geq 0$ and $\pi_{base}(v, Obl(a)) \cdot \pi_{base}(v, Per(a)) \geq 0$ because π_{base} is a promotion function, hence $\pi_{sup}(v, Obl(a)) \cdot \pi_{sup}(v, Per(a)) =$*

[8]Note that although the base promotion function considers an $\epsilon \in [0,1]$, here we demand that $\epsilon \in (0,1)$, because the supererogatory property does neither allow i) to assign a promotion of 0 to a permission; ii) to assign the same promotion to the permission and obligation of the same action.

$-\alpha_v^-(a) \cdot \pi_{base}(v, Obl(a)) \cdot \pi_{base}(v, Per(a)) \geq 0$. *Neutrality is satisfied because, if* $\alpha_v^+(a) = \alpha_v^-(a) = 0$, *then* $-\alpha_v^-(a) \cdot \pi_{base}(v, n) = \pi_{base}(v, n) = 0$. *If* $\alpha_v^+(a) > 0$, *we have* $\pi_{base}(v, n) \geq 0$ *and* $\alpha_v^-(a) \leq 0$, *so* $-\alpha_v^-(a) \cdot \pi_{base}(v, n) \geq 0$, *therefore praiseworthiness pursuit holds. On the other hand, if* $\alpha_v^+(a) < 0$, *then* $\pi_{sup}(v, n) = \pi_{base}(v, n)$, *which proves blameworthiness avoidance. Now we prove that* π_{sup} *satisfies the supererogatory property (see Definition 15). Suppose a supererogatory action with* $\alpha_v^+(a) > 0$ *and* $\alpha_v^-(a) = 0$, *then we have* $\pi_{sup}(v, Obl(a)) = -\alpha_v^-(a) \cdot \pi_{base}(v, Obl(a)) = 0$ *and* $\pi_{sup}(v, Per(a)) = \pi_{base}(v, Per(a)) > 0$, *because we have demanded an* $\epsilon \in (0, 1)$ *(see Definition 16), hence* $\pi_{sup}(v, Per(a)) > \pi_{sup}(v, Obl(a)) = 0$, *which means that* π_{sup} *satisfies condition* $(C1)$. *Now suppose* $\alpha_v^+(a) > 0$ *and* $\alpha_v^-(a) = -1$, *in this case* $\pi_{sup}(v, n) = \pi_{base}(v, n)$ *and since we are considering an* $\epsilon \in (0, 1)$, *we have* $\pi_{sup}(v, Obl(a)) > \pi_{sup}(v, Per(a)) > 0$, *hence satisfying condition* $(C2)$. *Thus,* π_{sup} *is supererogatory.* □

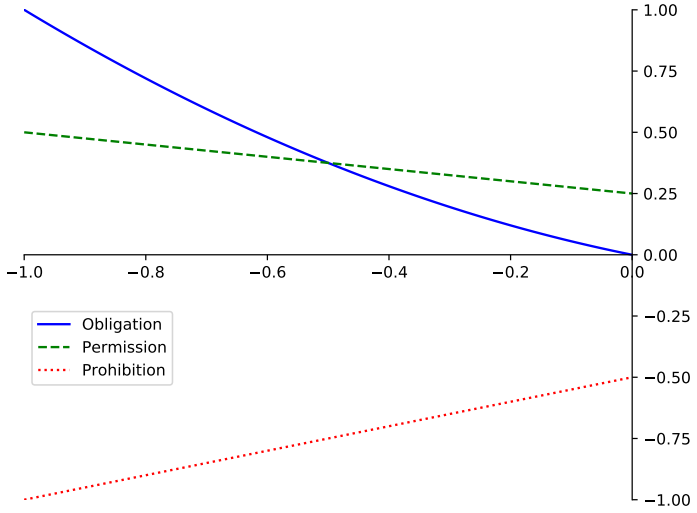
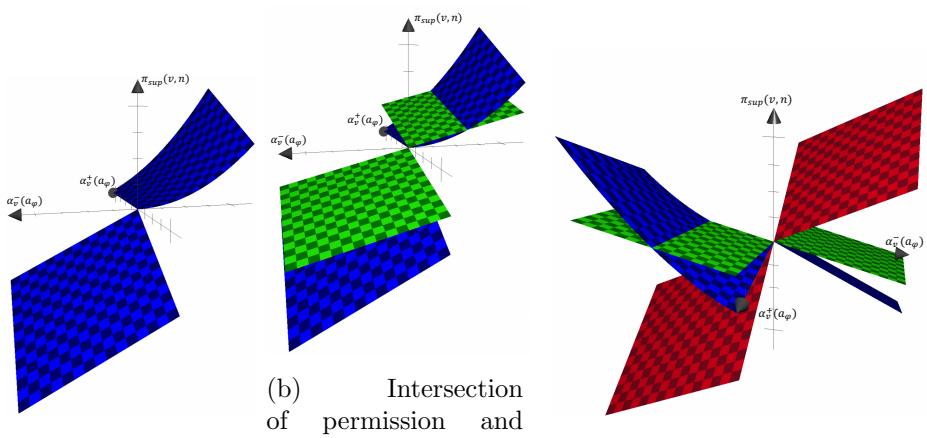


Figure 3.10: Supererogatory promotion function π_{sup} for a fully praiseworthy action $\alpha_v^+(a) = 1$ and $\epsilon = 0.5$. x-axis (from 0 to -1): blameworthiness to skip action $\alpha_v^-(a)$. y-axis: value promotion degree of obligation, permission, and prohibition of the action.

Figure 3.10 shows the plot of the supererogatory promotion function π_{sup} supposing an action that is totally praiseworthy to perform $(\alpha_v^+(a) = 1)$ and $\epsilon = 0.5$. Likewise Figure 3.8, the x-axis represents the blameworthiness of skipping the action $\alpha_v^-(a)$. The three plots represent the promotion degrees

(in the y-axis from -1 to 1) for the three possible norms regulating the action, namely: obligation (in blue), permission (in green) and prohibition (in red). Note that, since the action is praiseworthy, prohibiting it has always negative promotion, while permitting or obliging it has always positive promotion. In particular, notice that when the action is supererogatory and therefore is not very blameworthy to skip $\alpha_v^-(a) > -0.5$ the permission has greater promotion than the obligation. On the other hand, if the action is very blameworthy to skip $\alpha_v^-(a) < -0.5$, the obligation has greater promotion than the permission.

In general, Figure 3.11 shows plots considering any possible $\alpha_v^+(a)$ and $\alpha_v^-(a)$ for obligations only (Figure 3.11a), obligations and permissions (Figure 3.11b) and all deontic operator cases (Figure 3.11c). Figure 3.11b shows the intersection of the plots of π_{sup} representing the permission and obligation cases for $\epsilon = 0.5$. Again, when $\alpha_v^-(a) > -0.5$ (i.e., $\alpha_v^-(a) > -\epsilon$) the permission has higher promotion degree than the obligation, and when $\alpha_v^-(a) < -0.5$ the obligation has higher promotion degree. Figure 3.11c shows the 3D plots for the three cases of π_{sup}, here you can see the intersection of the plots of π_{sup}^{Obl} and π_{sup}^{Per} from another point of view.



(a) Obligations (b) Intersection of permission and obligation cases (c) Combined

Figure 3.11: Plots of different combinations of cases of the supererogatory promotion function using $\epsilon = 0.5$. All axis represent values in $[-1, 1]$, the arrow on all axis marks point 1.

3.7 Computing value-aligned norm systems

At this point, we are ready to formally pose our central problem. Given a set of candidate norms and a value system, recall that our goal, as outlined in Figure 3.1, is to compute the *most value-aligned* sound norm system.

3.7.1 Computing value alignment

To reason about norm systems based on moral value preferences, we must be able to compare them in terms of the moral values that they promote. The key principle that we adopt for this is: *the more preferred the moral values promoted by a norm system and the higher the promotion degree, the more value-aligned the norm system.* Thus, a decision maker will opt for sound norm systems that promote the most preferred moral values, and hence are more aligned with the value system on hand.

Let $\langle D, N, \mathfrak{R} \rangle$ be a normative domain. In order to quantitatively compute the value alignment of a norm system (out of the candidate norms N) with a value system $VS = \langle V, \succeq \rangle$, we will proceed as follows. First, we obtain the relevance of each moral value in VS from the value ranking $\succeq$. The relevance of a value is a numerical utility to encompass how preferred the value is (see the following paragraph). Second, we compute the value alignment of any norm system using the norm promotion of its norms to the values and the relevance of the promoted values.

To compute quantitative preferences over the moral values in VS, we define a relevance function $r : V \to \mathbb{R}$ that translates the qualitative preferences expressed by $\succeq$ to value relevance. Specifically, we require that, for $v, v' \in V$, if v is more preferred than v', then its relevance $r(v)$ must be greater than $r(v')$. Following the same reasoning, if v and v' are indifferently preferred, then they have equal relevance $r(v) = r(v')$. Ultimately, by setting a relevance for each moral value, we will be able to compare all the moral values in V.

Thus, we consider the value equivalence classes in $V/\sim$ and their quotient order $\succ$. All the values in an equivalence class $\eta \in V/\sim$ have the same relevance. Furthermore, the more preferred the equivalence class the more relevance their values have. Then, we can define the relevance of a value as the relevance of its equivalence class as follows. Say that v is a value in

equivalence class η. Then, we compute the relevance of v as:

$$r(v) = r(\eta) = \sum_{\eta \succ \eta'} r(\eta') + 1 = 2^{|\eta' : \eta \succ \eta'|} \tag{3.5}$$

Example 8. *The values in the value system of Example 7 would have the following relevance (applying Equation 3.5):* $r(Tim) = 1$ *and* $r(Civ) = 1 + r(v_{tim}) = 2$.

By using value relevance we can calculate the value alignment of a norm system by aggregating the relevance of the moral values each norm it promotes/demotes, being the relevance of each moral value weighed by the degree of promotion/demotion from the norm to the moral value. Formally:

Definition 17 (Value alignment)**.** *Given a norm system* $\Omega \subseteq N$*, a value system* $VS = \langle V, \succeq \rangle$*, and a promotion function* π*, we define the value alignment of* Ω *as:*

$$va(\Omega) = \sum_{n \in \Omega} \sum_{v \in V} \pi(v, n) \cdot r(v) \tag{3.6}$$

The following example illustrates how to compute the value alignment of some norm systems in our running example.

Example 9. *Considering the value judgements of Example 6, and the supererogatory promotion function* π_{sup} *(with* $\epsilon = 0.5$*), we obtain the following norm promotions:*

$\pi(Civ, Per(safe)) = 0.2$	$\pi(Tim, Per(safe)) = 0.25$	$\pi(Civ, Per(bin)) = 0.25$
$\pi(Civ, Obl(safe)) = 0$	$\pi(Tim, Obl(safe)) = 0.25$	$\pi(Tim, Per(bin)) = -0.25$
$\pi(Civ, Per(thr)) = -0.4$	$\pi(Tim, Per(thr)) = 0.5$	$\pi(Civ, Obl(bin)) = 0$
$\pi(Civ, Obl(thr)) = -0.8$	$\pi(Tim, Obl(thr)) = 1$	$\pi(Tim, Obl(bin)) = -0.5$
$\pi(Civ, Prh(thr)) = 0.8$	$\pi(Tim, Prh(thr)) = -1$	

Considering these norm promotions and the values' relevance of Example 8, we now assess the value alignment of the sound norm systems in Example 5:

$va(\{Per(thr)\}) = -0.3$	$va(\{Obl(thr)\}) = -0.6$
$va(\{Per(safe)\}) = 0.65$	$va(\{Obl(safe)\}) = 0.25$
$va(\{Per(bin)\}) = 0.25$	$va(\{Obl(bin)\}) = -0.5$
$va(\{Prh(thr)\}) = 0.6$	$va(\{Prh(thr), Obl(bin)\}) = 0.1$
$va(\{Per(thr), Obl(safe)\}) = -0.05$	$va(\{Prh(thr), Per(bin)\}) = 0.85$

3.7.2 Problem formalisation

In Section 3.7.1 we learned how to compute the value alignment of a norm system in terms of the values it promotes. Now we are ready to define the so-called *value-aligned norm selection problem* as an optimisation problem as follows:

Problem 1 (Value-aligned norm selection problem (VANS))**.** *Given a normative domain $\langle D, N, \mathfrak{R}\rangle$, a value system $\langle V, \succeq\rangle$, and a promotion function π, the value-aligned norm selection problem is that of finding a sound norm system $\Omega \subseteq N$ maximising value alignment. This amounts to solving:*

$$\max_{\Omega \subseteq N} \Big(va(\Omega) \Big) \, s.t. \; \Omega \text{ is sound.} \tag{3.7}$$

Theorem 3. *The value-aligned norm selection problem is NP-Hard.*

Proof (Theorem 3)**.** *We prove the theorem by reduction of the maximum independent set problem, a classic NP-Hard optimisation problem [Karp, 1972], to the value-aligned norm selection problem. Consider a graph $G = \langle Vt, E\rangle$, where Vt is a set of vertices and E is a set of (both directed and undirected) edges connecting the vertices in Vt. We say a set of vertices $S \subseteq Vt$ is independent if no two vertices in S are connected through an edge in E. Then, the maximum independent set problem amounts to finding the independent set S of maximum cardinality. Here we focus on a simpler class of the general value-aligned norm selection problem. Consider the VANS problem consisting of a normative domain $\langle D, N, \mathfrak{R}\rangle$ such that $\mathfrak{R} = \{\mathfrak{R}_i, \emptyset\}$ only contains incompatibility relationships, and a value system $\langle V, \succeq\rangle$ such that all norms have the very same value alignment, namely $va(\{n\}) = k$ for all $n \in N$. Now consider the graph $G = (Vt, E)$, where each vertex in Vt stands for a norm in N and each edge in E stands for an incompatibility relationship in $\mathfrak{R}_i$. From this follows that finding the maximum independent set of G amounts to solving the value-aligned norm selection problem for the above-defined normative domain and value system. Indeed, note that since we have supposed that all norms have the same value alignment, the solution to the VANS problem is the largest set of norms that is sound, which is exactly the maximum independent set of G (because we only have incompatibility relations, which are represented by edges in G). Therefore, in general, solving the value-aligned norm selection problem is at least NP-Hard.* □

3.8 A binary integer program to compute value-aligned norm systems

Notice that solving the VANS problem amounts to solving the optimisation problem expressed in equation 3.7. Next, we show how to solve such optimisation problem as a binary integer program. A binary integer program (BIP) [Lieberman and Hillier, 2005] encodes an optimisation problem in which the decision variables take values in $\{0, 1\}$. A VANS problem can be encoded as a BIP where each decision variable represents a norm. Thus, we would have the binary decision variables $\{x_1, \ldots, x_{|N|}\}$[9], where each x_i encodes the decision on whether a norm $n_i \in N$ is selected (taking value 1) or not (taking value 0). Thus, the VANS problem can be solved by the following binary integer program:

$$\max_{x_i \in \{0,1\}} \sum_{i=1}^{|N|} x_i \cdot va(\{n_i\}) \tag{3.8}$$

Subject to the following constraints:

- *Incompatibility constraints* preventing that two incompatible norms are jointly selected to be part of a norm system. Thus, the following constraints must hold:

$$x_i + x_j \leq 1 \quad \text{for each } (n_i, n_j) \in \mathfrak{R}_i. \tag{3.9}$$

- *Generalisation constraints* ensuring that a norm cannot be simultaneously selected together with any of its ancestors, namely:

$$x_i + x_k \leq 1 \quad n_k \in A(n_i) \quad 1 \leq i \leq |N| \tag{3.10}$$

- Non-aligned norm constraints discarding those norms that are not aligned with the moral values (in other words, norms with negative or neutral value alignment):

$$x_i = 0 \quad \forall n_i \in N, \text{ s.t. } va(\{n\}) \leq 0 \tag{3.11}$$

[9] While theoretically a VANS problem can be defined with a non-finite set of norms N, in order to encode it as a BIP we require a finite number of decision variables, hence N has to be finite.

We also need constraints for decision variables $x_i \in \{0, 1\}$. The BIP encoding of the VANS problem requires $|N|$ binary decision variables; $|\mathfrak{R}_i| + \sum_{n \in N} |A(n)|$ pairwise constraints (Equations 3.9 and 3.10); and $|\{n : va(n) \leq 0, n \in N\}|$ non-negativity constraints (Equation 3.11).

Notice that the specification above corresponds to a maximization problem whose constraints are all inequalities. Hence, it is in standard form and it can be solved with state-of-the-art solvers such as CPLEX [IBM, 1988] or Gurobi [Gurobi Optimization, 2010].

Example 10. *Considering the normative domain of Example 4, the value system in Example 7 and the value alignements of Example 9. The optimisation function would be:*

$$-0.3x_{thr}^{Per} - 0.6x_{thr}^{Obl} + 0.6x_{thr}^{Prh} + 0.65x_{safe}^{Per} + 0.25x_{safe}^{Obl} + 0.25x_{bin}^{Per} - 0.5x_{bin}^{Obl}$$

where variable $x_\theta^a \in \{1, 0\}$ represents norm $\theta(a)$. Some of the constraints to consider in this case are: $x_{thr}^{Per} + x_{thr}^{Obl} < 1$ (due to Equation 3.9), $x_{thr}^{Per} + x_{safe}^{Per} < 1$ (due to Equation 3.10) or $x_{thr}^{Obl} = 0$ (due to Equation 3.11). With this optimisation formula and all the constraints the resulting most value-aligned sound norm system is $\{Prh(thr), Per(bin)\}$. Note that, for this small problem we could have found this solution manually, in fact in Example 9 we have assessed the value alignment of all sound norm systems and $\{Prh(thr), Per(bin)\}$ has the maximum (0.85).

In [Serramia et al., 2021d], we provide the implementation of an algorithm for encoding a VANS problem into a BIP and solve it subsequently, more details about this implementation can be found in Appendix A.

3.8.1 Example: Analysing the public civility problem

Different value rankings may vary the selection of the most value-aligned norm system. In previous examples we solved the public civility game for the case that civility is preferred to timeliness ($Civ \succeq Tim$). This section explores how the solution changes for different value rankings (namely, that timeliness is preferred to civility or that both are equally valued).

Timeliness preferred to civility

Considering the normative domain of Example 4, the norm promotions of Example 9 and the value system in Example 7 but changing the ranking to

$Civ \succeq Tim$. The optimisation function would be:

$$0.6x_{thr}^{Per} + 1.2x_{thr}^{Obl} - 1.2x_{thr}^{Prh} + 0.7x_{safe}^{Per} + 0.5x_{safe}^{Obl} - 0.25x_{bin}^{Per} - 1.0x_{bin}^{Obl}$$

where variable $x_\theta^a \in \{1, 0\}$ represents norm $\theta(a)$. With this optimisation formula and all the constraints the resulting most value-aligned sound norm system is $\{Obl(thr)\}$.

Civility and timeliness equally valued

Considering the normative domain of Example 4, the norm promotions of Example 9 and the value system in Example 7 but changing the ranking to $Civ \sim Tim$. The optimisation function would be:

$$0.1x_{thr}^{Per} + 0.2x_{thr}^{Obl} - 0.2x_{thr}^{Prh} + 0.45x_{safe}^{Per} + 0.25x_{safe}^{Obl} + 0x_{bin}^{Per} - 0.5x_{bin}^{Obl}$$

where variable $x_\theta^a \in \{1, 0\}$ represents norm $\theta(a)$. With this optimisation formula and all the constraints the resulting most value-aligned sound norm system this time is however $\{Per(safe)\}$.

Analysis

As we have seen on the running example as well as on this section, given a normative domain, different value system's preferences will yield different solutions. In our running example we have considered that the value of civility is preferred over the timeliness of the agents. Thus, the resulting norm system both allows the agents to clean the garbage they might encounter on their path and prohibits to throw the garbage (as it may hurt another agent). Importantly, note that even though civility is the most preferred value in this case, the norm system does not oblige agents to clean the street, since this is a supererogatory action. On the other hand, in this case study, we have seen that when we prefer timeliness over civility, the resulting norm system obliges agents to throw the garbage nearby anytime they find it. In this case, cleaning the street is not a priority. Instead, the norm obliges the agents to get rid of the garbage swiftly to arrive in time. Finally, if we equally prefer both values, the resulting outcome is in between both previous outcomes, since now the resulting norm system allows to throw the garbage nearby after the agent has ensured that it is safe to do so (i.e. that it will not harm another agent).

3.9 Experimental evaluation

In this section we present an experimental evaluation to assess the hardness of solving the VANS. Our goal is twofold. On the one hand, we aim at finding the factors that make VANS problems hard to solve. On the other hand, we want to investigate the scope of applicability (in terms of scalability) of BIP solving for our problem instances. First, in Section 3.9.1 we describe the methodology that we followed to run our experiments. Thereafter, we analyse our results in Sections 3.9.2, 3.9.3, and summarise the results in Section 3.9.4.

3.9.1 Experimental methodology

Our empirical analysis follows three stages: (1) generation of synthetic instances of VANSs; (2) encoding of synthetic problem instances as BIPs following the methodology in sections 3.7 and 3.8; (3) solving the resulting binary integer programs using a BIP solver. Next, we focus on the generation of synthetic problem instances of VANSs. For that, we follow these steps (the necessary input parameters to generate a problem instance are highlighted):
(1) Firstly, we generate a **number of norms**.
(2) We then generate norm relations. For each pair of norms we randomly decide if they must be related or not based on a **relation density** parameter. If two norms are deemed to be related, we use the **incompatibility percentage** to decide if the relation should be an incompatibility relation or a generalisation relation. Note that the incompatibility percentage sets the number of incompatible relations, while the remaining relations will be generalisation relations.
(3) Afterwards, we generate the values (whose number is fixed to 10% of the number of norms) and their preferences. Notice though that in [Serramia et al., 2018b], we already observed that the number of values does not affect solving times. For each $i \in [1, |V|]$, we set v_i and v_{i+1} as indifferently preferred with a 20% fixed probability.
(4) Finally, we generate norm promotions. For each norm-value pair, we decide whether the norm promotes/demotes the value considering a fixed probability of 20%. If the norm is deemed to promote/demote the value, the norm promotion is generated randomly from $[-1, 1]$, otherwise is 0.

We provide an algorithm detailing how to generate VANS problems in Appendix B. Figure 3.12 shows an example of a synthetically-generated

VANS problem instance.

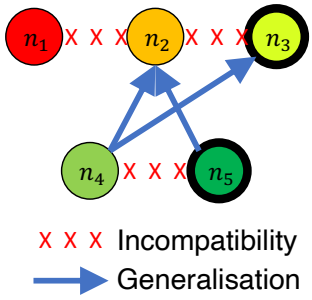


Figure 3.12: VANS problem instance for 5 norms, 3 values, 30% relation density, 50% incompatibility percentage. Edges represent norm relations. Norms' colours represent overall value alignment (w.r.t. all values): from green (most positive alignment) to red (most negative alignment). Hence, if we order the norms in terms of their value alignment we would have that $va(n_5) > va(n_4) > va(n_3) > va(n_2) > va(n_1)$. The solution is: $\{n_3, n_5\}$.

At this point the normative domain and value system are fully generated, thus the BIP can be built according to Section 3.8, and then solved using an off-the-shelf commercial solver[10]. Next, we divide our analysis in two fronts, namely how solving time is affected by: (1) norm relations; (2) the number of norms. In [Serramia et al., 2018b] we concluded that the ratio of generalisation relations to incompatibility relations (there noted as exclusivity relations) was a variable that greatly affected time. While here we define the problem and norm relations differently, it is still similar enough to use it as a lead for the empirical analysis.

3.9.2 Effect of norm relations on solving times

To study the sensitivity to norm relations, we specify the following generation parameters: number of norms, fixed to 500; relation density, from 1% to 100% in steps of 1; and incompatibility percentage, also from 1% to 100% in steps of 1. Specifically, we generate 10000 problem instances varying both the relation density and the incompatibility percentage.

Figure 3.13 shows a scatter plot of the solving times for these 10000 problem instances. The x-axis represents the incompatibility percentage

[10]We used IBM ILOG CPLEX Interactive Optimizer 12.10.0.0 on an Ubuntu 16.04 box with an Intel(R) Core(TM)i7-8700K CPU @ 3.70GHz, with 31GiB system memory, and 8th Gen CoreProcessor Host Bridge / DRAM R.

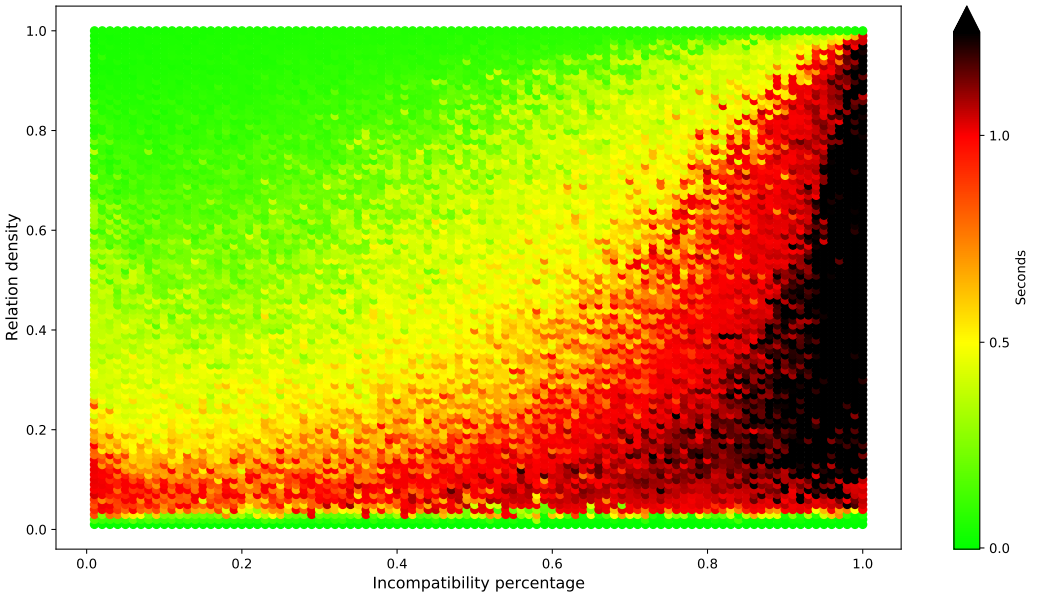


Figure 3.13: Solving times of problem instances with different relation densities and incompatibility percentages.

and the y-axis represents relation density. The colour of each dot in the plot represents the solving time of a problem instance.

Firstly, we observe that the vast majority of problem instances were solved in less than 1 second (indeed, most of these took less than 0.5 second). Problem instances that took more than 1 second usually had much larger solving times: in fact, all problem instances coloured in black took 10 seconds or more. Secondly, in terms of relation density, we observe that: when relation density is very close to 0 or 100%, problems are easy to solve either because no checks have to be performed (when close to 0), or because not many solutions (if any) are plausible (when close to 100%). Interestingly, the problem instances become harder to solve with relation densities at around 10%. We think this is a consequence of these problem instances being at a point where they have a number of constraints large enough so that many checks have to be made, but not large enough to discard many possible solutions. Thus, problem instances with less relation density become easier because less checks have to be made, while problems with more relation density also become easier because there are less possible solutions. Finally, in line with the findings in [Serramia et al., 2018b], solving times increase continuously with the percentage of incompatibilities over all relations.

We classify problem instances in three sets: *low hardness* (in green), *medium hardness* (in yellow or orange) and *high hardness* (in red or black). For a fixed incompatibility percentage of 50%, low hardness problem instances would happen between 100% and 75% relation density, medium hardness between 75% and 25% and high hardness for percentages lower than 25%. We think this happens because generalisation relation constraints depend on norm ancestors which creates interrelated generalisation networks. Therefore, if a norm in the network is clearly beneficial, the other norms are discarded immediately. Furthermore, incompatibility relations are more dispersed, thus creating less interrelated networks. Thus, when a norm is clearly beneficial fewer – incompatible – norms will be discarded.

3.9.3 Effect of the number of norms on solving time

Now we look at how norms affect solving times considering (previously described) problem hardness. Specifically, we study four types of problem instances: a scenario without norm relations that we use as baseline, and the low, medium, and high hardness problems described above. For each of these problem types we generate 400 instances. In particular we generate 10 instances for each number of norms ranging from 0 to 5000 norms in steps of 500.

Figures 3.14 and 3.15 show the plots relating the number of norms (x-axis) and the solving times in seconds (y-axis).

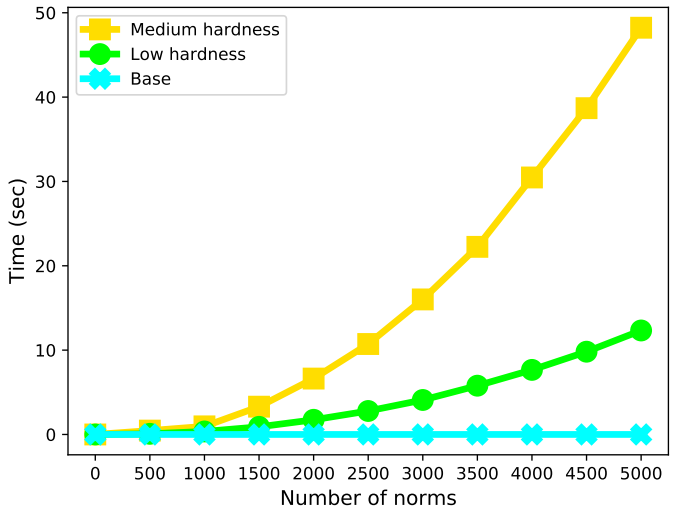


Figure 3.14: Solving times for base, low and medium hardness

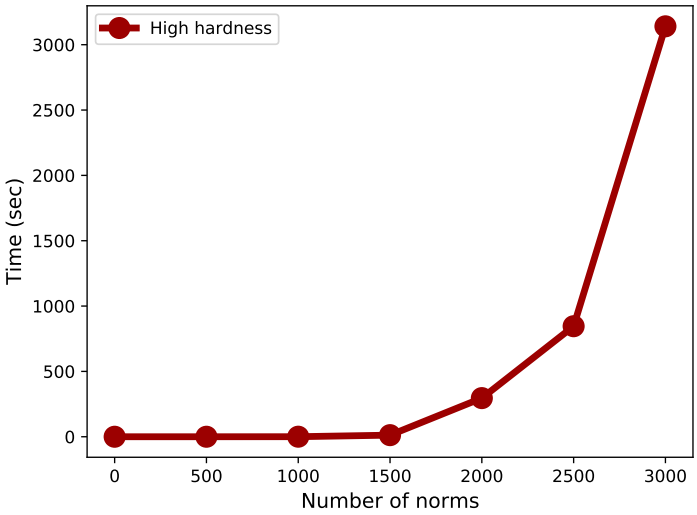


Figure 3.15: Solving times for high hardness

Figure 3.14 shows solving times for baseline, easy and medium problems. We observe that solving times increase as the number of norms increases. Furthermore, for a fixed amount of norms, the solving time depends on the hardness of the problem instance at hand. We can observe that base and low-hardness problems are solved in less than 10 seconds even for 5000 norms. Medium problems have reasonable solving times of under a minute for 5000. Finally, Figure 3.15 shows that the solving time for hard instances exponentially grows with the number of norms[11].

3.9.4 Summary of the empirical analysis

Overall, we have seen that solving times increase in three cases:

- **When incompatibility relations significantly outnumber generalisation relations**. When fixing the number of norms and considering different relation densities and incompatibility percentages, we observe that the solving times for problems with larger percentages of incompatibility relations are much larger (see Figure 3.13). This is mainly due to the different relation structures generated by incompatibility and generalisation. On the one hand, generalisation forms

[11]Note that Figure 3.15 only shows results up to 3000 norms, since solving problem instances with more norms required more than 1 hour, which is the maximum time set for our solver.

(ancestor) trees. Therefore, the selection of a norm will automatically imply discarding all its sibling norms, and thus, reducing the solving time. On the other hand, incompatibility relations are uniformly distributed. This implies that incompatible norms are disperse, so that choosing a norm causes discarding few incompatible norms. Hence, for a fixed number of relations, the larger the percentage of incompatibility over generalisation, the longer the solving times.

- **When there is low density of relations (around 10% of all possible norm relations)**. Figure 3.13 also shows that problem instances with low relation densities are larger. Relation densities close to 0 do not require checking constraints. However, relation densities at around 10% require to check a number of constraints that is large enough to require quite some solving time but it is not large enough for the solver to discard vast amounts of solutions.

- **When increasing the number of norms considered**. Quite trivially, when all else is fixed, solving times increase as the number of norms increases. This is shown in Figures 3.14 and 3.15. Of course, increasing the number of norms of a problem also increases the number of possible solutions, hence solving problems with larger number of norms takes longer.

While each of the points discussed above alone is enough to increase solving times, problems combining many of them produce longer solving times. This effect is noticeable when comparing figures 3.14 and 3.15. The slope of the curve of solving times in Figure 3.15 is steeper than that of the curves in figure 3.14, and also its curvature is more pronounced. This is because the problem instances used to produce Figure 3.15 were hard problems (with 50% incompatibility percentage and 25% relation density). Hence, these problems are already hard with regards to the first two points discussed above, but when increasing the number of norms, their solving times increase even more.

3.10 Conclusions and limitations

This chapter provides both the theoretical foundations and practical mechanisms for the selection of norm systems that promote most preferred moral values in a society. We do so by posing the so-called *value-aligned norm selection problem* (VANS) grounded in two structures: the normative domain,

defining norms and their relationships; and the value system, containing prioritised sets of moral values. We connect these structures via the norm promotion function, which is grounded on the praiseworthiness (or value judgement) of actions and allows us to quantify the value alignment of norm systems. Then solving the problem amounts to finding the sound norm system (i.e., without conflicting nor redundant norms) that maximises value alignment. In order to find the solution, we encode the VANS problem as a binary integer program and solve it with a state-of-the-art solver. We illustrate our proposal with a case study over the public civility game. Finally, we conduct empirical analysis and conclude that problems dealing with thousands of candidate norms can be solved within reasonable times.

This amounts to the first seven contributions from Section 1.3, which in turn answer the corresponding research questions in Section 1.2:

- Question Q1: How to formally define a value system? Grounded on the Ethics literature, we formally define a moral value based on the judgement of actions (contribution C1.1) and formalise value preferences as rankings (C1.2).
- Question Q2: How do we define norms and their relations? As the normative domain (C2).
- Question Q3: How are norms and values related? Through the promotion function (C3).
- Question Q4: How do we define the value-aligned norm selection problem? As that of selecting the (sound) norm system that promotes the most preferred moral values in the value system of the society (C4).
- Question Q5: How do we assess the value alignment of a single norm? By means of the value alignment function, an utility function which considers both norm-value promotion and value preferences (C5).
- Question Q6: How do we solve the value-aligned norm selection problem? By encoding it as a binary integer program (BIP) (C6).
- Question Q7: Is solving the value-aligned norm selection problem computationally feasible? Yes (C7).

Nonetheless, the quantitative approach described in this chapter has some limitations worth discussing.

Firstly, the approach described in this chapter requires the decision maker to numerically assess the relation between actions and values. In other words, we assume that the value judgement functions are known to the decision maker. This is a strong assumption that may not always hold. Decision makers may only count on minimal domain knowledge, and therefore, may not be able to provide a numerical assessment.

Secondly, our utility approach is not excellence rewarding. It favours quantity of norms over quality of norms. Thus, depending on the VANS problem, the approach will favour the selection of a large set of mediocre norms (those with low utility, namely low value alignment) instead of a smaller set of excellent norms (those with high utility, namely high value alignment). This may happen when the overall cumulative utility of the mediocre norms is greater than the utility of the excellent norms, and the mediocre norms cannot be jointly selected with the excellent ones. In this case, the mediocre norms would be selected. This has undesirable consequences for value-alignment, since this means that several norms slightly value-aligned will be selected in front of a single strongly value-aligned norm. Thus, for value-aligned norm selection, we prefer an approach that is excellence rewarding, where even a single excellent norm will always be selected in front of any arbitrary large number of mediocre norms.

In the next chapter, we tackle both of this shortcomings.

Chapter 4

Qualitative value-aligned norm selection

4.1 Introduction

Some actual-world decision making problems require to select an array of elements despite decision makers only counting on preferences over the elements' features. Note that, in this chapter, we consider the elements' features as our decision criteria. As previously mentioned, some examples of such problems are committee selection, coalition formation, product line composition, budget allocation, or college admissions [Fishburn, 1992; Gale and Shapley, 1962; Roth and Sotomayor, 1992]. Considering this last example, picture the following situation. A school head master must decide on which students to grant admission to. For that, the head master leverages on the admission policy of the school, which, for instance, prioritises some minorities, or fosters impoverished neighbourhoods. Such policies can be cast as preferences over the students' features. Nonetheless, the head master lacks of a straightforward manner to rank all possible sets of students, since these features somehow pose a multi-criteria problem. Moreover, there is a further dimension of complexity: some sets may not be eligible (e.g. because of limited budget, or unfulfilment of minority quotas). And yet, despite only counting on preferences over features and not sets, the head master must select the most preferred set of students. Interestingly, we can think of many other, similar set selection problems, such as selecting the team of players for a match (where we prefer some types of players over others), personnel selection (where some capabilities may be preferred over others), or the problem discussed in this thesis, that of selecting regulatory

norms (where we prefer norms that are aligned with moral values), etc. The goal of this chapter is to design the tools to help decision makers select the "most preferred" set in this type of problem, which hereafter we will refer to as *dominant set selection problem* (DSSP). Dominance characterises maximal preference in a formal (and particular) way.

In more general terms, assuming that we have sets of objects representing alternatives in a decision making process, the problem that we tackle is that of finding the most preferred set of objects. This decision must be made based on preference information over the *features* characterising the objects. For instance, in our admission example, ethnic group, neighbourhood, and studied subjects constitute some possible features. Furthermore, as noticed above, when dealing with decisions, preferences are not the only aspect to consider. Thus, we also require that the selected set does comply with some *feasibility constraints*, be them structural –due to relationships between the objects–, or inherent to the application domain.

In order to solve the *dominant set selection problem*, we propose to proceed as follows: (1) extract preferences over single objects based on preferences over objects' features; (2) rank all possible sets of objects; and (3) select the most preferred and feasible set of objects. For that, we resort to recent, seminal work in the realm of decision making and social choice theory, namely social rankings [Moretti and Öztürk, 2017] and its solutions [Bernardi et al., 2019; Haret et al., 2018; Khani et al., 2019; Allouche et al., 2020]. By adapting lex-cel, a ranking method introduced in [Bernardi et al., 2019], we are able to obtain a ranking over single objects from the feature preferences. Ultimately, our goal is to rank all sets of objects considering this element ranking, in other words, lifting the element ranking to a set ranking. This lifting procedure is very similar to the *ranking sets of objects* problem, which has been extensively studied in the social choice literature [Barberà et al., 2004]. Example solutions to this problem are the maxmin and minmax [Arlegi, 2003] or leximin and leximax [Pattanaik and Peleg, 1984] functions. Unfortunately, this problem considers a total order of elements instead of an element ranking. Hence, for the purpose of this chapter, we cannot readily use any of these approaches. Instead, here we design a novel *ranking function*, the so-called *anti-lex-cel*. This function receives as an input a ranking over single objects (obtained through lex-cel), and builds a ranking over all possible sets of these objects such that the most preferred feasible set in the ranking is the solution to the dominant set selection problem. The combination of the *lex-cel* ranking described in [Bernardi et al., 2019] with our novel *anti-lex-cel* ranking helps us produce our intended rank-

ing over all possible sets of objects, and hence solve the core of the dominant set selection problem.

From a pragmatic perspective, building a ranking over all sets of objects turns out to be computationally costly. Hence, we show how to solve the dominant set selection problem while avoiding the cost of explicitly building a whole ranking. In particular, we show how to to encode it as a binary integer program (BIP) so that it can be solved with the aid of off-the-shelf solvers. Importantly, we prove that the proposed encoding adheres to the ranking produced after *lex-cel* and *anti-lex-cel*, and that the solution to our BIP is equivalent to that of the dominant set selection problem. We illustrate the application of our method to a value-alignment problem initially introduced in [Serramia et al., 2018b] and subsequently investigated from a qualitative perspective in [Serramia et al., 2020]. In particular, and as discussed in previous chapters, given a collection of candidate norms, we investigate the selection of the (sub)set of norms, the so-called norm system[1], that is best *aligned* with the moral values[2] in a value system. The dominant set selection problem in this case is performed according to the following principle: the more preferred the moral values promoted by a norm system, the more preferred the norm system, or, in other words, the more dominant with respect to value alignment. Here the decision maker must consider: the preferences over moral values in the value system, the promotion relationship between norms and moral values (which can be interpreted as norm features), and the feasibility conditions based on the relationships between norms.

Notice that the approach of this chapter differs from the norm selection method proposed in Chapter 3 , which follows a quantitative approach despite the decision maker counting on qualitative information (i.e. value preferences). The approach in Chapter 3 forces the decision maker to quantify the relations between norms and actions by defining value judgement functions, which are then used to quantify the relation between norms and values. We argue that defining judgement functions is hard to ascertain and, as noted in [Santhanam, 2016], transforming qualitative information (i.e. value preferences) into numerical data (i.e. numerical value alignment) is prone to errors and biases (see the limitations discussed in Section 3.10). In fact, this is a general claim that can be applied when solving the dominant set selection problem. Therefore, in this chapter we opt for a qualitative ap-

[1]Norms provide the means to regulate the behaviour of individuals within a society, and a norm system is a set of norms to enact in that society for regulatory purposes.

[2]Moral values can be described as principles that a society deems valuable.

proach with the aim of keeping the decision making process as intuitive as possible.

The contributions of this chapter are:

- Formalisation of a novel qualitative decision-making problem, the so-called *dominant set selection problem* (DSSP).
- Formalisation and study of a novel preference lifting function called *anti-lex-cel*. We provide an axiomatic characterisation of *anti-lex-cel*, and we show that it generalises former results in the social choice literature in [Bossert et al., 1994].
- Development of a novel method for solving the DSSP based on the combination of the *lex-cel* ranking function in [Bernardi et al., 2019] with our novel *anti-lex-cel* ranking function.
- A binary integer program (BIP) encoding that is proven to solve the DSSP while avoiding the cost of explicitly building a whole ranking over all possible sets of objects.
- An application of the general methodology explained in this chapter to the value-aligned norm selection problem.

This chapter significantly extends our previous work in [Serramia et al., 2020] in two main respects. First, here we present a general formalisation and solving method for the DSSP, hence going beyond [Serramia et al., 2020], which solely focused on composing value-aligned norm systems, namely on a particular DSSP. Second, here we add with respect to [Serramia et al., 2020] a thorough axiomatic characterisation of *anti-lex-cel*, a formal proof of its uniqueness, and results that show the generality of anti-lex-cel with respect to existing results in the social choice literature.

The chapter is structured as follows. Next, Section 4.2 motivates the usefulness of the dominant set selection problem and provides an informal definition. Then, in Section 4.3 we introduce some necessary background on order theory to subsequently formalise the dominant set selection problem in Section 4.4, where we also introduce a simple running example to illustrate the technicalities along the chapter. Section 4.5 outlines the resolution of the DSSP. We base the solution of the DSSP on two operators: lex-cel (in Section 4.6) and anti-lex-cel (in Section 4.7). In Section 4.8 we detail their use to solve the DSSP along with a BIP encoding to compute its solution with the aid of state-of-the-art solvers. In more practical terms, Section 4.9

we exploit the tools developed to solve DSSPs to show how to undertake value-aligned norm selection. Section 4.9 also illustrates how value-aligned norm selection depends on the actual preferences over the value system at hand. Finally, Section 4.10 draws conclusions and discusses the limitations of this approach. Recall that, for the ease of readability, we have included a List of Notation and Symbols.

4.2 Problem motivation: Value-aligned norm selection

As mentioned above, there is a number of problems that require to select the most preferred set of objects considering preferences over their (qualitative) features. Thus, we have discussed that a decision maker may need to choose: students to award grants to; players to form teams; personnel to undertake tasks; projects to be funded; or norms to be enacted.

In fact, that last example will help us to illustrate the characterisation of the problem at hand. Specifically, we assume that there is a set of candidate norms N and we aim to find the set of norms that better aligns with the moral values of the society. Our paper [Serramia et al., 2020] introduces some norm examples in an airport border context, where a norm "Permission to cross the border" is aligned with the moral value of "freedom of movement" whereas the norm "Obligation to show passport" is aligned with the value of "security" and is incompatible with the previous norm (i.e., they cannot be simultaneously enacted). Overall, to assess value-alignment we count on a set of moral values, preferences among these values, and a function relating norms to the values that they promote (i.e., specifying norms' features). For instance, consider: four norms $\{n_1, \ldots, n_4\}$; three values $\{v_1, v_2, v_3\}$, being v_1 more preferred than v_2 and v_3, which are indifferent between them; and a feature function that specifies that norm n_1 promotes the three values, and that the remaining norms only promote one value each (n_2 promotes v_1, n_3 promotes v_2 and n_4 promotes v_3). Then, the principle we adhere to is: The more preferred the values promoted by a norm, the more preferred the norm and the more preferred the norms in a set the more value-aligned the set. Thus, we consider $\{n_2\}$ aligns more with moral values than $\{n_4\}$ because n_2 is preferred over n_4 since it promotes a more preferred value. Furthermore, when considering larger sets of these norms, value alignment only grows larger. Following our example, set $S_1 = \{n_1, n_2\}$ is more value-aligned than $S_2 = \{n_3, n_4\}$ because n_1 alone is more preferred than any of the norms in

S_2, and adding n_2 only strengthens the value alignment of S_1. Additionally, while the more preferred values have greater impact on assessing which set is more aligned, whenever possible, we still will prefer to select additional norms even if they promote less preferred values (e.g., we favour $\{n_1, n_2, n_3\}$ over S_1). Finally though, since not all norm sets are feasible (norms may be incompatible or redundant between them), the decision maker counts on a function to check if a norm set is feasible or not.

In these terms, the value-aligned norm selection problem consists on finding a set of norms $S \subseteq N$, such that:

- S is feasible;
- S contains the most preferred norms possible (the norms that promote the most preferred values): If we change any norm of S for a more preferred one, the set becomes unfeasible.
- S is *maximal*, namely it is the largest feasible set: adding any further norms to S makes it unfeasible.

We say this S dominates all other feasible sets and therefore we call the problem of finding it *dominant set selection problem* (DSSP). As discussed before many selection problems that count on preferences among the features of the elements can be cast into a DSSP. For example, selecting players to play in a match (where we prefer some types of players to others), awarding research teams (where we prefer to award excellency teams over regular teams), personnel selection (where some capabilities may be preferred over others), etc.

In Section 4.4 we provide a general formalisation of the dominant set selection problem that encompasses the particular case described above. Before that, we introduce some necessary background on order theory in the following section.

4.3 Background

Let X be a set of objects. A binary relation $\succeq$ on X is said to be: *reflexive*, if for each $x \in X$, $x \succeq x$; *transitive*, if for each $x, y, z \in X$, ($x \succeq y$ and $y \succeq z$) $\Rightarrow x \succeq z$; *total*, if for each $x, y \in X$, $x \succeq y$ or $y \succeq x$; *antisymmetric*, if for each $x, y \in X$, $x \succeq y$ and $y \succeq x \Rightarrow x = y$. We can define preferences among the elements of X by means of binary relations. Moreover, we can

categorise the type of preferences depending on the properties they hold as follows.

Definition 18 (Preorder, ranking, linear order and partial order)**.** *A preorder (or quasi-ordering) is a binary relation $\succeq$ that is reflexive and transitive. A preorder that is also total is called total preorder or ranking. A total preorder that is also antisymmetric is called a linear order. A preorder that is antisymmetric but not total is called a partial order.*

Note that neither preorders nor rankings are necessarily antisymmetric relations. Thus, given a ranking (or a preorder) $\succeq$ if $x, y \in X$, such that $x \succeq y$ and $y \succeq x$, then we cannot conclude that $x = y$, instead we say these two elements are indifferently preferred and note it as $x \sim y$.

Example 11. *Given a set $X = \{x_1, x_2, x_3\}$, an example of ranking would be: $x_1 \succeq x_2 \sim x_3$ (with a ranking we know how all elements are related).*

Notation 1. *We note all possible rankings over X as $\mathcal{R}(X)$.*

Using the indifference relation we can consider the quotient set $X/\sim$, which contains the equivalence classes of $\succeq$. Thus, given the ranking $x_1 \sim \cdots \sim x_s \succeq \cdots \succeq x_{r-k} \sim \cdots \sim x_r$, with $x_1, \ldots, x_s, \ldots, x_{r-k}, \ldots, x_r \in X$, then we can consider the quotient set $X/\sim$ with quotient order $\succ$: $\Sigma_1 \succ \cdots \succ \Sigma_n$, where $\Sigma_1 = \{x_1, \ldots, x_s\}, \ldots, \Sigma_n = \{x_{r-k}, \ldots, x_r\} \in X/\sim$ are equivalence classes.

4.4 Formalising the dominant set selection problem

The goal of this section is to formalise the dominant set selection problem. Informally, and in short, this problem is that of finding a set $S \in \mathcal{P}(X)$ that is both *feasible* and *more preferred than any other set*, and hence *dominates* other sets. Notice that feasibility is an internal property of each set that captures the compatibility of its elements. However, dominance refers to a preference relation of each set with others that is not initially known since the preferences at hand are those over the features of the elements of a set S. In what follows, we start formally characterising the objects in a dominant set selection problem. Thereafter, we show how to gradually build our formal notion of dominance over sets from the preferences at hand,

namely those over features of elements. Finally, we offer a formal definition of the dominant set selection problem.

To start with, we go back to our value-alignment problem in Section 4.2, from which we can generalise to identify the objects that formally characterise the input of a dominant set selection problem as follows:

- a set of elements X;
- a set of features F;
- a ranking $\succeq_F$ over the features in F;
- a function $\mathfrak{f} : X \to \mathcal{P}(F)$ that outputs the features of each element in X; and
- a feasibility function $\phi : \mathcal{P}(X) \to \{\top, \bot\}$, which checks if a set $S \in \mathcal{P}(X)$ is feasible ($\phi(S) = \top$ means that it is feasible, and $\phi(S) = \bot$ means that it is not).

At this point, it is important to remark that throughout this chapter we consider that $\emptyset$ is not a set in $\mathcal{P}(X)$ ($\emptyset \notin \mathcal{P}(X)$). Therefore, we note as $\mathcal{P}(X)$ the set containing the $2^{|X|} - 1$ different non-empty subsets of X.

Informally, solving the dominant set selection problem amounts to selecting a feasible set $S \in \mathcal{P}(X)$ that is more preferred than any other set and includes as many elements as possible. We will say that such set *dominates* the other sets. To select such dominant set we must first formalise our notion of *dominance*. First, we will only consider a single element and define element dominance in: (i) a (equivalence) class of features; and (ii) a whole ranking over features. Once we have established how element dominance works, we will build upon it to define set dominance.

Given a ranking over features $\succeq_F$, we define element dominance within the scope of an equivalence class of features as follows:

Definition 19. *Given two elements $x, y \in X$ with features in F, a ranking over features $\succeq_F$, and a feature equivalence class $\Psi \in F/\sim_F$, we say that x is Ψ-dominant over y if*

$$|\mathfrak{f}(x) \cap \Psi| > |\mathfrak{f}(y) \cap \Psi|.$$

If $|\mathfrak{f}(x) \cap \Psi| = |\mathfrak{f}(y) \cap \Psi|$, we say that x and y are Ψ-indifferent.

Back to our example in Section 4.2, the dominant set selection problem would be characterised by: $X = \{n_1 \dots n_4\}$; $F = \{v_1, v_2, v_3\}$; $\mathfrak{f}(n_1) = \{v_1, v_2, v_3\}$, $\mathfrak{f}(n_2) = \{v_1\}$, $\mathfrak{f}(n_3) = \{v_2\}$, $\mathfrak{f}(n_4) = \{v_3\}$; and $v_1 \succeq v_2 \sim v_3$. In the quotient order of $F/\sim_F$, this results in two feature equivalence classes: $\Psi_1 = \{v_1\} \succ_F \Psi_2 = \{v_2, v_3\}$. With this in mind, n_1 is Ψ_1-dominant over n_4 because n_1 promotes v_1 but n_4 does not. n_4 is Ψ_2-dominant over n_2 since n_4 promotes v_3 and n_2 does not promote any value in Ψ_2. Finally n_1 and n_2 are Ψ_1-indifferent as they both promote v_1.

Next, we exploit the definition of element Ψ-dominance to define element dominance considering all the features in F and their ranking $\succeq_F$. Formally:

Definition 20. *Given two elements $x, y \in X$ with features in F and a ranking over features $\succeq_F$, we say that x is dominant over y if there is a feature equivalence class $\Psi \in F/\sim_F$, such that:*

- *x is Ψ-dominant over y; and*

- *$\forall \Psi' \in F/\sim_F$, such that $\Psi' \succ_F \Psi$, x and y are Ψ'-indifferent.*

If neither x dominates y nor vice versa, we say that x and y are indifferent.

Note that the first condition in Definition 20 implies that the dominant element x has more of the features of Ψ than y ($|\mathfrak{f}(x) \cap \Psi| > |\mathfrak{f}(y) \cap \Psi|$). As for the second condition, it demands that x and y are indifferent for any other equivalence classes that are more preferred than Ψ. Hence, the most preferred feature equivalence class for which x and y differ, is the class that marks dominance between them.

Back to our example: n_1 is dominant over n_4 because n_1 is Ψ_1-dominant over n_4 and Ψ_1 is the most preferred feature class; n_1 is also dominant over n_2 because even though they are Ψ_1-indifferent, n_1 is Ψ_2-dominant over n_2.

With the definition of element dominance we now consider dominance between sets in $\mathcal{P}(X)$. Given a set $S = \{x_1, \dots, x_t\}$, $S \in \mathcal{P}(X)$, we can order its elements in a sequence $(x_{\sigma(1)}, \dots, x_{\sigma(t)})$ according to dominance, where σ is a permutation of the indexes, such that $\sigma(i)$ is the index in S of the i-th element in the sequence. According to such ordering, $x_{\sigma(i)}$ is indifferent or dominated by $x_{\sigma(1)}, \dots, x_{\sigma(i-1)}$ while being indifferent or dominating $x_{\sigma(i+1)}, \dots, x_{\sigma(t)}$. With this in mind we define set dominance as follows.

Definition 21. *Given two sets* $S = \{x_1, \ldots, x_t\}$ *and* $S' = \{x'_1, \ldots, x'_r\}$ *in* $\mathcal{P}(X)$ *and their orderings according to dominance* $(x_{\sigma(1)}, \ldots, x_{\sigma(t)})$ *and* $(x'_{\sigma(1)}, \ldots, x'_{\sigma(r)})$ *respectively, we say that* S *is dominant over* S' *if* $\exists j \in \{1, \max(t, r)\}$*, such that:*

- $x_{\sigma(j)}$ *dominates* $x'_{\sigma(j)}$ *or* $j > r$*; and*
- $x_{\sigma(i)}$ *and* $x'_{\sigma(i)}$ *are indifferent* $\forall i < j$*.*

Notice that the notion of dominance that we propose rewards element *excellence* in a set: the more preferred (excellent) the features of the elements in a set, the more dominant the set. Therefore, a set containing a few excellent elements (with regards to their features) will be preferred over larger sets with *mediocre* elements (i.e. related to less preferred features). This will be the case even if the mediocre elements in a larger set are related to many more features.

Continuing with our example, the set $S_1 = \{n_1, n_2\}$ is *dominant* over $S_2 = \{n_3, n_4\}$ because n_1 is the most dominant element in S_1, n_3 is the most dominant element in S_2, and n_1 is dominant over n_3.

With the definition of set dominance we can now tackle the formalisation of the dominant set selection problem. Formally:

Problem 2 (Dominant set selection problem)**.** *Given a set of elements* X*, a set of features* F*, a ranking* $\succeq_F$ *over* F*, a function* $\mathfrak{f} : X \to \mathcal{P}(F)$ *linking the elements in* X *with their features, and a feasibility function* $\phi : \mathcal{P}(X) \to \{\top, \bot\}$ *that checks if a set* $S \in \mathcal{P}(X)$ *is feasible, then the dominant set selection problem (DSSP) is that of finding a set* $S \in \mathcal{P}(X)$ *such that:*

- S *is feasible, that is,* $\phi(S) = \top$*; and*
- *no other feasible set dominates* S *, that is, if* $S' \in \mathcal{P}(X)$*, such that,* S' *is dominant over* $S \Rightarrow \phi(S') = \bot$*.*

At this point, we remind the reader that our notion of dominance above is meant to reward element excellence. Hence, dominant set selection problems model decision problems for which element excellence is the main decision criterion. This is the case in the examples mentioned in the introduction (granting admissions, committee selection), other examples are awarding prizes or scholarships. Of course, other decision criteria are possible. For example, a decision maker could consider avoiding incompetence as the main

decision criterion (in this case elements with more features would be preferred over elements with few more preferred features).

Indeed, it is worth stressing that the problem of how to aggregate different attributes or variables is typically studied in the field of *Multiple Criteria Decision Making* (MCDM), where it is relevant to ask to a decision maker the question on whether the compensation of bad performances on some criteria by good performances on other criteria is acceptable or not [Roy and Słowiński, 2013]. As pointed out in [Bouyssou, 1986], the notion of compensation in general boils down to that of 'tradeoffs' among criteria. For instance, a possibility of compensation is provided by additive utility-based approaches, but there are plenty of other methods offering different levels of compensation, or using non-compensatory aggregation techniques (see, for instance, the article [Roy and Słowiński, 2013] for a discussion about the question guiding to the choice of an appropriate MCDM method and the articles [Bouyssou, 1986; Słowiński et al., 2002] for an axiomatic analysis of MCDM methods in situations with multicriteria non-compensatory preferences; see also [Greco et al., 2019] for an updated review of compensatory and non-compensatory approaches). Therefore, although this issue is the subject of considerable debate in the MCDM literature, here we define a specific dominance notion that rewards element excellence, and argue that its applicability is strongly dependent on the context.

Note also that the dominant set selection problem may have multiple solutions when multiple sets satisfy the conditions of the problem and do not dominate one another. However, it may also be worth mentioning that, by construction, these solutions will always have the same number of elements (see Section 4.5).

To illustrate the problem and its resolution we use the following problem as a running example in the following sections.

Example 12. *Consider four elements* $X = \{x_1, x_2, x_3, x_4\}$*, two features* $F = \{f_1, f_2\}$*, the feature ranking* $f_1 \succeq_F f_2$ *and the feature function* $\mathfrak{f}(x_1) = \mathfrak{f}(x_2) = \{f_1\}$ *and* $\mathfrak{f}(x_3) = \mathfrak{f}(x_4) = \{f_2\}$*. In terms of feasibility, we know that any set containing both* x_1 *and* x_3*, or both* x_2 *and* x_4*, is not feasible (e.g.* $\phi(\{x_1, x_2, x_3\}) = \bot, \phi(\{x_3, x_4\}) = \top$*). These elements conform an example of dominant set selection problem.*

The next section outlines how we actually proceed to solve the dominant set selection problem.

4.5 Solving the dominant set selection problem: an outline

As we anticipated in the introduction above, we tackle the dominant set selection problem by splitting its resolution in three steps: (1) we extract preferences over single objects based on their features and on the preferences over the features; (2) we rank all possible sets of objects; and (3) we select the most preferred feasible set of objects. Figure 4.1 shows the general outline of the steps that we shall follow to solve the dominant set selection problem: *preference grounding*, *preference lifting*, and *feasibility check*. First, preference grounding is performed by grounding the preferences over objects' features to obtain a ranking over the objects in X. Second, preference lifting lifts this element ranking over the elements in X to a set ranking over $\mathcal{P}(X)$. Notice that preference lifting must ensure that the output ranking embodies dominance, thus meaning that a set dominates all its (strictly) less preferred sets in the ranking. Third, the feasibility check step finds the feasible set that is most preferred in the ranking over $\mathcal{P}(X)$. That set will be the set that is dominant over all other feasible sets and, thus, it will constitute the solution to our problem.

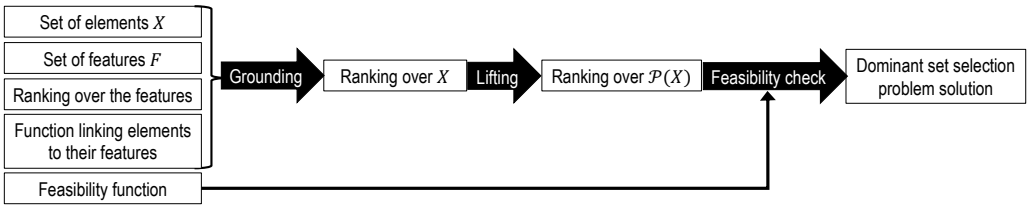


Figure 4.1: Outline of the steps to solve the dominant set selection problem.

The main difficulty when solving the dominant set selection problem lies on generating a ranking over all sets in $\mathcal{P}(X)$. Therefore, the next sections (4.6, 4.7 and 4.8) focus on that goal. Sections 4.6 and 4.7 introduce two key functions that will allow: (i) to transform a ranking over elements' features into a ranking over elements in X; and (ii) in turn this ranking over elements into a ranking over sets in $\mathcal{P}(X)$.

At this point, we warn the reader that Section 4.6 must be taken as background, since lex-cel was already introduced in [Bernardi et al., 2019], whereas sections 4.7 and 4.8 contain novel contributions.

4.6 The lex-cel ranking grounding function

The social ranking problem [Moretti and Öztürk, 2017] consists on transforming a ranking over $\mathcal{P}(X)$ into a ranking over the elements of X . Thus, a social ranking solution can be viewed as a function $srs : \mathcal{R}(\mathcal{P}(X)) \to \mathcal{R}(X)$, such that for a ranking $\succeq \in \mathcal{R}(\mathcal{P}(X))$, $srs(\succeq) = \succeq_e$ is a ranking of X. Informally, we say that a social ranking solution *grounds* the preferences over subsets to preferences over elements.

Several social ranking solutions have been proposed, such as: a grounding function based on the *ceteris paribus* majority principle [Haret et al., 2018]; a grounding function based on the notion of marginal contribution [Khani et al., 2019]; two rankings based on the analysis of majority graphs and minmax score [Allouche et al., 2020]; or the *lex-cel* ranking function [Bernardi et al., 2019], which is based on lexicographical preferences. Here, we adapt lex-cel to rank the elements in X based on their features in F.

In more detail, if we consider a ranking $\succeq$ over $\mathcal{P}(X)$, the transformation performed by lex-cel proceeds as follows. First, consider the quotient set $\mathcal{P}(X)/\sim$ (see Section 4.3) such that subsets related by indifference relations fall on the same equivalence class $\Sigma_i \in \mathcal{P}(X)/\sim$. Since the equivalence classes are not indifferent between them, we have a strict quotient order $\succ$ between them: $\Sigma_1 \succ \cdots \succ \Sigma_{|\mathcal{P}(X)/\sim|}$.

We now define a function $\mu : X \to \mathbb{N}^{|\mathcal{P}(X)/\sim|}$, which for an element $x \in X$ returns its profile vector, a natural vector whose dimension is the number of equivalence classes in the quotient set $|\mathcal{P}(X)/\sim|$. The i-th component of the profile vector for x stands for the number of times that x appears in the subsets of equivalence class Σ_i. Notice that equivalence class Σ_i is the class containing the i-th most preferred subsets of $\mathcal{P}(X)$ according to the preorder $\succeq$. For instance, if $\mu(x) = (c_1^x, \ldots, c_{|\mathcal{P}(X)/\sim|}^x)$, then c_i^x is the number of times that x appears in the subsets of equivalence class Σ_i. Formally, we define the profile vector for an element $x \in X$ as:

$$\mu(x) = (c_1^x, \ldots, c_{|\mathcal{P}(X)/\sim|}^x), \text{ where } c_i^x = |\{S \in \Sigma_i : x \in S\}| \tag{4.1}$$

Given any two elements $x, y \in X$, we can establish a preference between them by comparing their profile vectors with the lexicographical order of vectors. That is:

Definition 22. *We define the lexicographical order of vectors $\geq_L$ such that given two vectors $c = (c_1, \ldots, c_m), c' = (c'_1, \ldots, c'_m) \in \mathbb{N}^m$, we say that $c >_L c'$ iff $\exists i$, such that $c_1 = c'_1; \ldots; c_{i-1} = c'_{i-1}$ and $c_i > c'_i$. On the other hand, $c =_L c' \Leftrightarrow c = c'$.*

We then define the lexicographical-exellence (lex-cel) grounded ranking $le(\succeq) = \succeq_e$ between two elements by comparing their profile vectors. Given $x, y \in X$, we say that:

$$\begin{cases} x \succeq_e y \Leftrightarrow \mu(x) \geq_L \mu(y) \\ x \preceq_e y \Leftrightarrow \mu(x) \leq_L \mu(y) \\ x \sim_e y \Leftrightarrow \mu(x) = \mu(y) \end{cases} \tag{4.2}$$

In [Bernardi et al., 2019], the authors prove that grounding preferences with lex-cel satisfies properties that make the grounding fair. In particular, such properties are neutrality, coalitional anonymity, monotonicity and independence of the worst set. Next, we provide a short illustration of these four properties.

First, neutrality ensures that the ranking resulting from applying lex-cel does not depend on the elements' names/identities. Specifically, this property means that if we permute two elements x and y in a ranking $\succeq$ over $\mathcal{P}(X)$, the grounded ranking should obey to the same permutation. So, for instance, consider a ranking $\succeq$ over $\mathcal{P}(X)$, with $X = \{x, y, z\}$ and such that $\{x, y, z\} \succeq \{x\} \succeq \{y, z\} \succeq \{x, y\} \succeq \{y\} \succeq \{x, z\} \succeq \{z\}$. Suppose that the grounded ranking specifies the relation $x \succeq_e y$ on the ranking $\succeq$. Then, the grounded ranking should specify the relation $y \succeq'_e x$ on the ranking $\succeq'$ such that $\{x, y, z\} \succeq' \{y\} \succeq' \{x, z\} \succeq' \{x, y\} \succeq' \{x\} \succeq' \{y, z\} \succeq' \{z\}$, which is obtained from $\succeq$ by permuting x and y.

The coalitional anonymity property extends the anonymity principle to "non-informative" subsets of X: the relative ranking between two elements should only depend on the sequence in which they separately occur along the ranking over $\mathcal{P}(X)$. For instance, in the two rankings $\{x, y, z\} \succeq \{x\} \succeq \{y, z\} \succeq \{x, y\} \succeq \{y\} \succeq \{x, z\} \succeq \{z\}$ and $\{x, z\} \succeq' \{y, z\} \succeq' \{x, y, z\} \succeq' \{x, y\} \succeq' \{y\} \succeq' \{z\} \succeq' \{x\}$, if we focus on sets containing either x or y (but not both), from left to right: first, we have that element x occurs in the singleton set $\{x\}$ in $\succeq$ and in the set $\{x, z\}$ in $\succeq'$, then element y occurs in the set $\{y, z\}$ in both rankings $\succeq$ and $\succeq'$, y occurs in the set $\{y\}$ in both rankings $\succeq$ and $\succeq'$, and finally, x occurs in the subset $\{x, z\}$ in $\succeq$ and in $\{x\}$ in $\succeq'$. Therefore, since x and y occur according to the sequence x, y, y, x on both rankings $\succeq$ and $\succeq'$, a grounded ranking satisfying coalitional anonymity should specify the same relation between x and y on $\succeq$ and $\succeq'$ (i.e., $x \succeq_e y \Leftrightarrow x \succeq'_e y$).

As shown in [Bernardi et al., 2019], neutrality and coalitional anonymity together imply that if two elements x and y are such that $\mu(x) = \mu(y)$, then

they should be ranked indifferent in the grounded ranking.

A grounded ranking that satisfies monotonicity, breaks possible indifference relations in a consistent way. This means that if on a ranking $\succeq$ over $\mathcal{P}(X)$ a grounded ranking states that two elements x and y are indifferent (i.e, $x \sim_e y$), then, if we consider a new ranking $\succeq'$ obtained from $\succeq$ by improving the position of some subsets containing x but not y, we should have that the grounded ranking ranks x strictly better than y on $\succeq'$ (i.e, $x \succeq'_e y$ and $x \nsim'_e y$). For instance, suppose that on a ranking $\{x, y, z\} \succeq \{x\} \sim \{y, z\} \succeq \{x, y\} \succeq \{y\} \sim \{x, z\} \succeq \{z\}$ the grounded ranking is such that $x \sim_e y$. Now, if we improve the position of the subset $\{x, z\}$, so that we obtain the new ranking $\{x, y, z\} \succeq' \{x\} \sim' \{y, z\} \succeq' \{x, y\} \succeq' \{x, z\} \succeq' \{y\} \succeq' \{z\}$, according to monotonicity we have a grounded ranking such that $x \succeq'_e y$ and $x \nsim'_e y$.

Finally, the property of independence of the worst subsets is aimed at accounting higher ranked subsets over lower ranked ones. Thus, we say that a grounded ranking is independent of the worst subsets if, once the grounded ranking has stated that an element x is strictly better than y, any change in the relative ranking of subsets in the worst indifference class of the ranking over $\mathcal{P}(X)$ does not affect such an assertion. For instance, suppose that on the ranking $\{x, y, z\} \succeq \{x\} \succeq \{y, z\} \succeq \{x, y\} \succeq \{y\} \sim \{x, z\} \sim \{z\}$ the grounded ranking states $x \succeq_e y$ and $x \nsim_e y$, then it should state the same for $\{x, y, z\} \succeq' \{x\} \succeq' \{y, z\} \succeq' \{x, y\} \succeq' \{y\} \succeq' \{x, z\} \succeq' \{z\}$, which is obtained from $\succeq$ by just modifying the relation among elements of its last equivalence class $\{\{y\}, \{x, z\}, \{z\}\}$. So, giving more importance to occurrences in higher ranked subsets, this property actually rewards the 'excellence' of elements in a ranking over $\mathcal{P}(X)$.

In [Bernardi et al., 2019], the authors not only prove that lex-cel satisfies these (logically independent) axioms, but also that it is the only grounding function that satisfies them.

Even though lex-cel is formally defined in [Bernardi et al., 2019] as a function $le : \mathcal{R}(\mathcal{P}(X)) \to \mathcal{R}(X)$, here we adapt it to handle the input of the dominant set selection problem and thus perform the grounding process in figure 4.1. Therefore, we redefine lex-cel as a function $le : \mathcal{R}(F) \to \mathcal{R}(X)$. Then, given a ranking of features $f_1 \succeq_F \cdots \succeq_F f_{|F|}$, with quotient order $\Psi_1 \succ_F \cdots \succ_F \Psi_{|F/\sim_F|}$ over $F/\sim_F$, and an element $x \in X$, the function μ, would be defined as:

$$\mu(x) = (|\mathfrak{f}(x) \cap \Psi_1|, \ldots, |\mathfrak{f}(x) \cap \Psi_{|F/\sim_F|}|).$$

Example 13. *Following Example 12, note that we know that elements x_1 and x_2 have both the most preferred feature f_1, while x_3 and x_4 have the least preferred feature f_2. With this in mind, their μ vectors would be: $\mu(x_1) = (1,0)$, $\mu(x_2) = (1,0)$, $\mu(x_3) = (0,1)$, $\mu(x_4) = (0,1)$. Therefore, the grounded ranking over X would be $x_1 \sim_e x_2 \succeq_e x_3 \sim_e x_4$.*

4.7 The anti-lex-cel ranking lifting function

Thanks to lex-cel we can ground a ranking over features in F to a ranking over the elements in X. As shown in Figure 4.1, the next step is to lift this ranking over single elements to a ranking over sets of elements, namely over $\mathcal{P}(X)$. This procedure is similar to that of the *ranking sets of objects* problem surveyed in [Barberà et al., 2004]. The ranking sets of objects problem consists on building a ranking over sets from an ordering over individual elements. Some solutions for the ranking sets of objects problem are maxmin and minmax, as introduced in [Arlegi, 2003]. Maxmin assesses preferences over sets by comparing only their most preferred element except when these elements are the same, in which case it compares their least preferred elements. On the other hand, minmax does the inverse comparison. It assesses preferences over sets based only on how their least preferred elements are compared. If these elements are the same, the sets' most preferred elements are compared. Note that neither of these methods consider further elements than the most and least preferred ones. This makes them unsuitable for our purpose, since we want to take into account as many elements as possible. The leximin and leximax functions introduced in [Pattanaik and Peleg, 1984] represent alternative approaches. In summary, leximin and leximax are based on comparing lexicographically sets. In the case of leximin, preferences over sets depend on how their worst elements compare. If these elements are the same, their second worst elements are compared, and so on. If there is no difference, the larger set is preferred (the sets are indifferent if both have the same size). Conversely, leximax compares sets depending on how their best elements compare. If these elements are the same, their second best elements are compared, and so on. If there is no difference, the smaller set is preferred (the sets are indifferent if both have the same size). Unfortunately, we cannot use any of the solutions of the ranking sets of objects problem because they assume a total order of elements. Instead, we have a more general assumption, since we suppose a ranking on elements. Note that this is a crucial difference, since rankings allow for different ele-

ments to be indifferently preferred, whereas total orders are antisymmetric, meaning that an element cannot be equally preferred to another element.

Since, to the best of our knowledge, no lifting functions assuming element rankings exist, in this section we formalise a novel one, which we call anti-lex-cel. In Section, 4.7.1 we formally introduce anti-lex-cel. Thereafter, in Section 4.7.2 we provide an axiomatic characterisation of anti-lex-cel and we prove that it is the only lifting function satisfying such axioms. Finally, Section 4.7.3 draws the relationship between lex-cel and anti-lex-cel while Section 4.7.4 connects the results in this section with existing results in the literature.

4.7.1 Formal definition

Anti-lex-cel can be viewed as a function $ale : \mathcal{R}(X) \to \mathcal{R}(\mathcal{P}(X))$, such that for a ranking $\succeq_e \in \mathcal{R}(X)$, $ale(\succeq_e) = \succeq$ is a ranking over $\mathcal{P}(X)$. We formalise anti-lex-cel in a very similar way to lex-cel, but reversing the process.

To perform anti-lex-cel we start with a ranking $\succeq_e$ over the elements in X. First, we consider the quotient set $X/\sim_e$. Each equivalence class in $X/\sim_e$ contains a set of indifferently preferred elements. Equivalence classes in $X/\sim_e$ are ordered by the quotient order $\succ_e$. Hence, $\Xi_1 \succ_e \cdots \succ_e \Xi_r$, where $r = |X/\sim_e|$ and Ξ_i is the equivalence class containing the i-th most preferred elements. We define a function $\eta : \mathcal{P}(X) \to \mathbb{N}^r$ to count the appearances of the elements of a set in $\mathcal{P}(X)$ in each equivalence class. Thus, given a set $S \in \mathcal{P}(X)$, $\eta(S)$ is a vector of size r whose i-th component stands for the number of elements in S that are found in the equivalence class Ξ_i. Formally:

$$\eta(S) = (s_1, \ldots, s_r), \text{ where } s_i = |S \cap \Xi_i| \tag{4.3}$$

Note that, similarly to μ in Equation 4.1, $\eta(S)$ is a vector whose elements represent how preferred the elements in S are: the larger the first numbers of the vector, the more preferred the elements in S are (in terms of $\succeq_e$), and hence we can infer that the more preferred S is. This again means that ranking sets of elements is equivalent to lexicographically ordering their associated vectors as calculated by the η function. Thus, to compare two sets $S, S' \in \mathcal{P}(X)$, we compare lexicographically $\eta(S)$ and $\eta(S')$ (see Definition 22). With those considerations, we are now ready to tackle the formulation of the anti-lex-cel function ale. We define $\succeq$ as the ranking of sets in $\mathcal{P}(X)$ such that given two sets $S, S' \in \mathcal{P}(X)$, it orders them according to the

following rules:

$$\begin{cases} S \succeq S' \Leftrightarrow \eta(S) \geq_L \eta(S') \\ S \preceq S' \Leftrightarrow \eta(S) \leq_L \eta(S') \\ S \sim S' \Leftrightarrow \eta(S) = \eta(S') \end{cases} \tag{4.4}$$

After that, we are ready to formally define the anti-lexicographic-excellence ranking lifting function as follows:

Definition 23. *Given a set of elements X and a ranking $\succeq_e$ over the elements in X, the ranking lifting function $ale : \mathcal{R}(X) \to \mathcal{R}(\mathcal{P}(X))$ such that $ale(\succeq_e) = \succeq$ is called anti lexicographic excellence (anti-lex-cel).*

Example 14. *Consider the element ranking $x_1 \sim_e x_2 \succeq_e x_3 \sim_e x_4$ over X that we found in Example 13. We apply anti-lex-cel to this ranking by computing the η vector for the sets in $\mathcal{P}(X)$. Since the quotient order is $\Xi_1 \succ \Xi_2$, with $\Xi_1 = \{x_1, x_2\}$ and $\Xi_2 = \{x_3, x_4\}$, we have that, for instance, $\eta(\{x_1, x_2, x_3\}) = (2, 1)$ and $\eta(\{x_3, x_4\}) = (0, 2)$. Then, by comparing the η vectors of all sets we can build the following ranking over $\mathcal{P}(X)$:* $\{x_1, x_2, x_3, x_4\} \succeq \{x_1, x_2, x_3\} \sim \{x_1, x_2, x_4\} \succeq \{x_1, x_2\} \succeq \{x_1, x_3, x_4\} \sim \{x_2, x_3, x_4\} \succeq \{x_1, x_3\} \sim \{x_1, x_4\} \sim \{x_2, x_3\} \sim \{x_2, x_4\} \succeq \{x_1\} \sim \{x_2\} \succeq \{x_3, x_4\} \succeq \{x_3\} \sim \{x_4\}$.

4.7.2 Axiomatic characterisation

We now introduce four properties for a ranking lifting function $f : \mathcal{R}(X) \to \mathcal{R}(\mathcal{P}(X))$ and prove that they together axiomatically characterise *ale* and that *ale* is the unique lifting function that satisfies them.

The first axiom is a coherence property saying that the ranking of singleton sets should be "aligned" with $\succeq_e$, where $\succeq_e$ is a ranking of the elements of X.

Axiom 1 (Simple Dominance)**.** Given an element ranking $\succeq_e \in \mathcal{R}(X)$, a ranking lifting function f satisfies the *simple dominance* property iff

$$x \succeq_e y \text{ and } x \nsim_e y \Rightarrow \{x\} \succeq \{y\} \text{ and } \{x\} \nsim \{y\}$$

for all $x, y \in X$ and with $\succeq = f(\succeq_e)$.

The second axiom is an anonymity property: permuting the names of elements should not affect the ranking provided by a lifting function.

Axiom 2 (Neutrality)**.** Given an element ranking $\succeq_e \in \mathcal{R}(X)$, let π be a bijection on X and let $\succeq_e^{\pi} \in \mathcal{R}(X)$ be such that by

$$x \succeq_e x' \Leftrightarrow \pi(x) \succeq_e^{\pi} \pi(x')$$

for all $x, x' \in X$. A lifting function f satisfies the *neutrality* property iff

$$S \succeq S' \Leftrightarrow \pi(S) \succeq^{\pi} \pi(S')$$

for all $S, S' \in \mathcal{P}(X)$ and where $\pi(S)$ and $\pi(S')$ are the images of S and S' through π and where $\succeq = f(\succeq_e)$ and $\succeq^{\pi} = f(\succeq_e^{\pi})$.

The next axiom says that if a set S is (weakly) preferred to another one S', then adding new elements to the preferred one S makes this new set (strictly) preferred to S'.

Axiom 3 (Size Monotonicity)**.** Given an element ranking $\succeq_e \in \mathcal{R}(X)$, a ranking lifting function f satisfies the *size monotonicity* property iff

$$S \succeq S' \Rightarrow (S \cup \bar{S}) \succ S' \text{ and } (S \cup \bar{S}) \nsim S'$$

for all $S, S' \in \mathcal{P}(X)$ and $\bar{S} \subseteq (X \setminus S)$, $\bar{S} \neq \emptyset$, with $\succeq = f(\succeq_e)$.

The next axiom aims at rewarding the best elements preventing the overestimation of dominated ones and states that a strict preference between two sets S and S', i.e. $S \succeq S'$ and $S \nsim S'$, should not be affected by the addition of new single element that are strictly worse (with respect to the element ranking $\succeq_e$ of X) to those already contained in the preferred set S.

Axiom 4 (Independence of the Worst Elements)**.** Given an element ranking $\succeq_e \in \mathcal{R}(X)$, a ranking lifting function f satisfies the *independence of the worst elements* property iff

$$S \succeq S' \text{ and } S \nsim S' \Rightarrow S \succeq (S' \cup \bar{S}') \text{ and } S \nsim (S' \cup \bar{S}')$$

for all $S, S' \in \mathcal{P}(X)$ and $\bar{S}' \subseteq (X \setminus S')$, $\bar{S}' \neq \emptyset$, such that $x \succeq_e x'$ and $x \nsim_e x'$ for all $x \in S$ and $x' \in \bar{S}'$ and with $\succeq = f(\succeq_e)$.

The following proposition establishes that anti-lex-cel satisfies the four axioms above.

Proposition 2. *The anti-lex-cel lifting function ale satisfies axioms 1, 2, 3 and 4.*

Proof (Proposition 2)**.** *Let $\succeq_e$ be a ranking over the elements of X and $X/\sim_e$ its quotient order with equivalence classes $\Xi_1 \succ_e \ldots \succ_e \Xi_r$. Let $\succeq = ale(\succeq_e)$.*
Axiom 1:
Take $x, y \in X$ such that $x \succeq_e y$ and $x \nsim_e y$. Then, two elements $i, j \in \{1, \ldots, r\}$ exist such that $i < j$, $x \in \Xi_i$ and $y \in \Xi_j$. So, $\eta(\{x\}) >_L \eta(\{y\})$ and, by relation (4.4), we have $\{x\} \succeq \{y\}$ and $\{x\} \nsim \{y\}$, which proves that ale satisfies Axiom 1.
Axiom 2:
let π be a bijection on X and let $\succeq_e^{\pi} \in \mathcal{R}(X)$ be such that

$$x \succeq_e y \Leftrightarrow \pi(x) \succeq_e^{\pi} \pi(y)$$

for all $x, y \in X$. Let $X/\sim_e^{\pi}$ be the quotient order of $\succeq_e^{\pi}$ with equivalence classes $\Xi_1^{\pi} \succ_e^{\pi} \ldots \succ_e^{\pi} \Xi_r^{\pi}$. Notice that $\succeq_e$ and $\succeq_e^{\pi}$ have precisely the same number of equivalence classes. Moreover, for any $x \in X$ there exists $i \in \{1, \ldots, r\}$ such that

$$x \in \Xi_i \Leftrightarrow \pi(x) \in \Xi_i^{\pi}.$$

So, $\eta(S) = \eta(\pi(S))$ for any $S \in \mathcal{P}(X)$ and where $\pi(S)$ is the image of S through π. By relation (4.4), it follows that ale also satisfies Axiom 2.
Axiom 3:
Consider two sets $S, S' \in \mathcal{P}(X)$ such that $S \succeq S'$. Consider the case $S \sim S'$ (the case $S \succeq S'$ and $S \nsim S'$ is similar and left to the reader). By relation (4.4) we have that $\eta(S) = \eta(S')$. Now take another set $\bar{S} \subseteq (X \setminus S)$, $\bar{S} \neq \emptyset$. Consider the new set $S \cup \bar{S}$ which contains some elements not in S. Then, $\eta(S \cup \bar{S}) >_L \eta(S) = \eta(S')$ and, by relation (4.4), it follows that $(S \cup \bar{S}) \succ S'$ and $(S \cup \bar{S}) \nsim S'$ which proves that ale satisfies Axiom 3.
Axiom 4:
Consider two sets $S, S' \in \mathcal{P}(X)$ such that $S \succeq S'$ and $S \nsim S'$. By relation (4.4), it exists $i \in \{1, \ldots, r\}$ such that $\eta_k(S) = \eta_k(S)$ for all $k \in \{1, \ldots, i-1\}$ and $\eta_i(S) > \eta_i(S')$ (being $\eta_j(S)$, the j-th element of $\eta(S)$). Let $\bar{S}' \subseteq (X \setminus S')$, $\bar{S}' \neq \emptyset$, be such that $x \succeq_e x'$ and $x \nsim_e x'$ for all $x \in S$ and $x' \in \bar{S}'$. Since each element in S is strictly preferred to each element in $\bar{S}'$, then $\eta_k(S') = \eta_k(S' \cup \bar{S}')$ for all $k \in \{1, \ldots, i\}$ and, consequently, $\eta(S) >_L \eta(S' \cup \bar{S}')$, which finally proves the fact that ale also satisfies Axiom 4.

Having axiomatized anti-lex-cel, we can obtain a stronger result. Thus, the following theorem tells us that in fact anti-lex-cel is the only lifting function that satisfies these axioms.

Theorem 4. *Let $f : \mathcal{R}(X) \to \mathcal{R}(\mathcal{P}(X))$ be a ranking lifting function. Then f satisfies axioms 1, 2, 3 and 4 if and only if f is the anti-lex-cel lifting function ale.*

To prove the theorem we require the previous introduction of the following auxiliary lemmas.

Lemma 1. *Let $f : \mathcal{R}(X) \to \mathcal{R}(\mathcal{P}(X))$ be a ranking lifting function that satisfies Axiom 2. Let $S, S' \in \mathcal{P}(X)$ and $\succeq_e \in \mathcal{R}(X)$ be such that $\eta(S) = \eta(S')$. Then $S \sim S'$ (where $\sim$ is the symmetric part of relation $\succeq = f(\succeq_e)$).*

Proof (Lemma 1)**.** *Let $\succeq_e$ be a ranking over the elements of X and $X/\sim_e$ its quotient order with equivalence classes $\Xi_1 \succ_e \ldots \succ_e \Xi_r$. Since $\eta(S) = \eta(S') = (s_1, \ldots, s_r)$, with $s_i = |S \cap \Xi_i| = |S' \cap \Xi_i|$ for all $i \in \{1, \ldots, r\}$, we can define a bijection π on X such that the $\pi(S \cap \Xi_i) = S' \cap \Xi_i$ and $\pi(S' \cap \Xi_i) = S \cap \Xi_i$ for all $i \in \{1, \ldots, r\}$. So, $\pi(S) = S'$ and $\pi(S') = S$.*

Define a new ranking $\succeq_e^\pi \in \mathcal{R}(X)$ such that $x \succeq_e y \Leftrightarrow \pi(y) \succeq_e^\pi \pi(y)$ for all $x, y \in X$. Then, by Axiom 2, we have that

$$S \succeq S' \Leftrightarrow S' \succeq^\pi S.$$

On the other hand, $\succeq = \succeq^\pi$, and we may conclude that $S \succeq S' \Leftrightarrow S' \succeq S$, which precisely means that $S \sim S'$ for $\succeq$ is a total relation.

Lemma 2. *Given an element ranking $\succeq_e \in \mathcal{R}(X)$ and $f : \mathcal{R}(X) \to \mathcal{R}(\mathcal{P}(X))$ a ranking lifting function that satisfies axioms 1 and 4. Then, the resulting set ranking $f(\succeq_e) = \succeq$ is such that $\forall x \in X$, $\{x\} \succeq S$ and $\{x\} \nsim S$ for every $S \subseteq L(x)$, where $L(x) = \{x' \in X | x \succeq_e x'$ and $x \nsim_e x'\}$, is the set of elements strictly less preferred than x.*

Proof (Lemma 2)**.** *Given an element ranking $\succeq_e \in \mathcal{R}(X)$, let $\succeq = f(\succeq_e)$ and $\succeq^* = ale(\succeq_e)$. Let $x \in X$ and $L(x) = \{x' \in X | x \succeq_e x'$ and $x \nsim_e x'\}$. Take $y \in L(x)$. By Axiom 1, we have that $\{x\} \succeq \{y\}$ and $\{x\} \nsim \{y\}$. Now take $\bar{S}' \subseteq L(x) \setminus \{y\}$. By Axiom 4, with $\{x\}$ in the role of S, $\{y\}$ in the role of S', we have that $\{x\} \succeq \{y\} \cup \bar{S}'$ and $\{x\} \nsim \{y\} \cup \bar{S}'$. Since $\{y\} \cup \bar{S}'$ can be whatever subset of $L(x)$, we have proved the lemma.*

With these auxiliar lemmas we can prove Theorem 4 as follows:

Proof (Theorem 4)**.** *We know from Proposition 2 that the anti-lex-cel lifting function ale satisfies axioms 1, 2, 3 and 4.*

Conversely, suppose f satisfies axioms 1, 2, 3 and 4. Consider an element ranking $\succeq_e \in \mathcal{R}(X)$ having $X/\sim_e$ as quotient order with equivalence classes $\Xi_1 \succ_e \ldots \succ_e \Xi_r$. Let $\succeq = f(\succeq_e)$ and $\succeq^* = ale(\succeq_e)$. It has to be proved that

$$Q \succeq^* Q' \Leftrightarrow Q \succeq Q'$$

for all $Q, Q' \in \mathcal{P}(X)$.
We first prove the equivalence for the asymmetric parts, i.e.,

$$Q \succeq^* Q' \text{ and } Q \nsim^* Q' \Leftrightarrow Q \succeq Q' \text{ and } Q \nsim Q'$$

for all $Q, Q' \in \mathcal{P}(X)$.
($\Rightarrow$)
Let $Q \succeq^* Q'$ and $Q \nsim^* Q'$. By relation (4.4), it exists $i \in \{1, \ldots, r\}$ such that $\eta_k(Q) = \eta_k(Q')$ for all $k \in \{1, \ldots, i-1\}$ and $\eta_i(Q) > \eta_i(Q')$ (being $\eta_j(Q)$, the j-th element of $\eta(Q)$). We distinguish two cases:

- *(i): $\eta_k(Q') = 0$ for all $k \in \{1, \ldots, i\}$. Take an element $x \in \Xi_i \cap Q$ and an element $y \in Q' \cap \Xi_j$ where $j \in \{i+1, \ldots, r\}$ is the smallest index such that $Q' \cap \Xi_j \neq \emptyset$. Notice that $j > i$, so $x \succeq_e y$ and $x \nsim_e y$ for all $y \in Q'$. Since f satisfies both Axiom 1 and 4, by Lemma 2 we have that $\{x\} \succeq Q'$ and $\{x\} \nsim Q'$. Now, let $\bar{Q} = Q \setminus \{x\}$. Being $Q = \{x\} \cup \bar{Q}$, and applying Axiom 3 on f with $\{x\}$ in the role of S, Q' in the role of S' and $\bar{Q}$ in the role of $\bar{S}$, we finally have $Q \succeq Q'$ and $Q \nsim Q'$.*

- *(ii): $\eta_k(Q') \neq 0$ for some $k \in \{1, \ldots, i\}$. First, consider the two non-empty sets $T = \bigcup_{k \in \{1,\ldots,i\}} (Q \cap \Xi_k)$ and $T' = \bigcup_{k \in \{1,\ldots,i\}} (Q' \cap \Xi_k)$. Since $\eta_i(Q) = |Q \cap \Xi_i| > |Q' \cap \Xi_i| = \eta_i(Q')$ there must be at least $\eta_i(Q) - \eta_i(Q')$ elements in $Q \cap \Xi_i$ but not in $Q' \cap \Xi_i$. Let $I = \{x \in Q \cap \Xi_i \setminus Q'\}$ with $|I| = \eta_i(Q) - \eta_i(Q')$. Now, consider the two sets $T \setminus I$ and T'. By construction, $\eta(T \setminus I) = \eta(T')$. Then, since f satisfies Axiom 2, by Lemma 1 we have that $T \setminus I \sim T'$. We are now ready to apply Axiom 3 on f with $T \setminus I$ in the role of S and T' in the role of S' and $\bar{T} = Q \setminus (T \setminus I)$ in the role of $\bar{S}$. Then, being $Q = (T \setminus I) \cup \bar{T}$, we have $Q \succeq T'$ and $Q \nsim T'$. Finally, we use Axiom 4 on f with Q in the role of S, T' in the role of S' and $\bar{T}' = Q' \setminus T'$ in the role of $\bar{S}'$. So, we have that $Q \succeq T' \cup \bar{T}'$ and $Q \nsim T' \cup \bar{T}'$. Being $Q' = T' \cup \bar{T}'$, we conclude that $Q \succeq Q'$ and $Q \nsim Q'$.*

($\Leftarrow$)
Let $Q \succeq Q'$ and $Q \nsim Q'$. First, suppose that $Q \sim^* Q'$. Then, by relation

(4.4), we have that $\eta(Q) = \eta(Q')$. So, by Lemma 1, it must be $Q \sim Q'$, which yields a contradiction. On the other hand it cannot even be $Q' \succeq^* Q$ and $Q' \nsim^* Q$ (otherwise we would have a contradiction by the other implication proved earlier). So, by the fact that $\succeq^*$ is a total relation, it must be $Q \succeq^* Q'$ and $Q \nsim^* Q'$, which concludes the proof of the equivalence between the asymmetric parts of $\succeq$ and $\succeq^*$.
We now prove the equivalence for the symmetric parts, i.e.,

$$Q \sim^* Q' \Leftrightarrow Q \sim Q'$$

for all $Q, Q' \in \mathcal{P}(X)$.
($\Rightarrow$)
Let $Q \sim^* Q'$. Then, by relation (4.4), we have that $\eta(Q) = \eta(Q')$ and by Lemma 1 and the fact that $\succeq$ satisfies Axiom 2, it immediately follows that $Q \sim Q'$.
($\Leftarrow$)
Let $Q \sim Q'$. By the equivalence of the asymmetric part proved earlier, we cannot have $Q \succeq^* Q'$ and $Q \nsim^* Q'$, or, $Q' \succeq^* Q$ and $Q' \nsim^* Q$. So, since $\succeq^*$ is a total relation, it must be $Q \sim^* Q'$, which concludes the proof.

Finally, we look into another result. The following proposition proves that axioms 1, 2, 3 and 4 are logically independent, so they all are necessary to uniquely characterise *ale*.

Proposition 3. *Exists a function $f \neq ale$ satisfying any three of the axioms 1, 2, 3 and 4 and not fulfilling the remaining one.*

Proof (Proposition 3)**.** *A ranking lifting function that does not satisfy simple dominance. Consider a ranking lifting function $f^{sd} : \mathcal{R}(X) \to \mathcal{R}(\mathcal{P}(X))$ with $\succeq = f^{sd}(\succeq_e)$ such that $S \succeq T \Leftrightarrow \eta(S) \geq_L \eta(T)$ for all $S, T \in \mathcal{P}(X)$ with $|S| > 1$ or $|T| > 1$, and $\{x\} \sim \{y\}$ for all $x, y \in X$. Similar to ale, it is easy to verify along the lines of Proposition 2 that f^{sd} satisfies axioms 2, 3 and 4. But of course f^{sd} does not satisfies Axiom 1 (just take $\succeq_e$ such that $x \succeq_e y$ and $x \nsim_e y$).*

A ranking lifting function that does not satisfy neutrality. Let $z \in X$. Consider a ranking lifting function $f^n : \mathcal{R}(X) \to \mathcal{R}(\mathcal{P}(X))$ with $\succeq = f^n(\succeq_e)$ such that $S \succeq T$ and $S \nsim T$ for all $S, T \in \mathcal{P}(X)$ with $\eta(S) = \eta(T)$ and $z \in S \setminus T$, and $S \succeq T \Leftrightarrow \eta(S) \geq_L \eta(T)$ for all the remaining pairs of sets $S, T \in \mathcal{P}(X)$. It easy to verify along the lines of Proposition 2 that f^n satisfies axioms 1, 3 and 4. But of course f^n does not satisfies Axiom 2, for

f^n breaks some ties in ale if favour of the set containing element z.

A ranking lifting function that does not satisfy size monotonicity. Consider a ranking lifting function $f^{sm} : \mathcal{R}(X) \to \mathcal{R}(\mathcal{P}(X))$ with $\succeq = f^{sm}(\succeq_e)$ such that $S \succeq T \Leftrightarrow i_S \leq i_T$ for all $S, T \in \mathcal{P}(X)$, where i_S and i_T are, respectively, the smallest index such that $\eta_{i_S}(S) \neq 0$ and $\eta_{i_T}(T) \neq 0$. One can check that f^{sm} satisfies Axioms 1, 2 and 4, but it does not fulfil Axiom 3. To see that f^{sm} does not satisfy Axiom 3, consider $X = \{x_1, x_2, x_3\}$ with the element ranking $x_1 \succeq_e x_2 \succeq_e x_3$, which implies $\eta(\{x_1\}) = (1, 0, 0)$, $\eta(\{x_2\}) = (0, 1, 0)$ and $\eta(\{x_3\}) = (0, 0, 1)$. Then, $\{x_1, x_2\} \sim \{x_1, x_3\}$ (for $\eta_1(\{x_1, x_2\}) = \eta_1(\{x_1, x_3\}) = 1 \neq 0$), but it is not true that $\{x_1, x_2\} \cup \{x_3\} \succeq \{x_1, x_3\}$ and $\{x_1, x_2\} \cup \{x_3\} \nsim \{x_1, x_3\}$ (for $\eta_1(\{x_1, x_2, x_3\}) = \eta_1(\{x_1, x_3\}) = 1 \neq 0$).

A ranking lifting function that does not satisfy independence of the worst elements. Consider a ranking lifting function $f^{iwe} : \mathcal{R}(X) \to \mathcal{R}(\mathcal{P}(X))$ with $\succeq = f^{iwe}(\succeq_e)$ and such that $S \succeq T \Leftrightarrow |S| \geq |T|$ for all $S, T \in \mathcal{P}(X)$ with $|S| > 1$ or $|T| > 1$, and $\{x\} \succeq \{y\} \Leftrightarrow x \succeq_e y$ for all $x, y \in X$. One can check that f^{iwe} satisfies Axioms 1, 2 and 3, but it does not fulfil Axiom 4. To see that f^{iwe} does not satisfy Axiom 4, consider $X = \{x_1, x_2, x_3, x_4\}$ and the element ranking $x_1 \sim_e x_2 \succeq_e x_3 \sim_e x_4$ with quotient order $\Xi_1 = \{x_1, x_2\} \succ_e \Xi_2 = \{x_3, x_4\}$. We have that $\{x_1, x_2\} \succeq \{x_2\}$ and $\{x_1, x_2\} \nsim \{x_2\}$ (for $|\{x_1, x_2\}| = 2$ and $|\{x_2\}| = 1$) but it is not true that $\{x_1, x_2\} \succeq \{x_2\} \cup \{x_3, x_4\}$ and $\{x_1, x_2\} \nsim \{x_2\} \cup \{x_3, x_4\}$ (for $|\{x_2, x_3, x_4\}| = 3$).

4.7.3 On the relation between lex-cel and anti-lex-cel

As noticed above, the anti-lex-cel function is very similar to lex-cel, though it realises the reverse process (from ranking over elements to ranking over sets of elements). However, notice that, since $le : \mathcal{R}(\mathcal{P}(X)) \to \mathcal{R}(X)$ is not injective (it cannot be because $|\mathcal{R}(\mathcal{P}(X))| > |\mathcal{R}(X)|$), there is no inverse for lex-cel, and therefore, in general, anti-lex-cel is not the inverse of lex-cel. Nonetheless, in what follows we characterise the conditions under which anti-lex-cel becomes the inverse of a restriction of lex-cel.

Before establishing such formal result, we introduce an auxiliary result that will prove useful for that purpose. The following lemma states that we can build the profile vector for an element in a compositional, additive manner. More precisely, we can obtain the μ and η profile vectors by adding up the profile vectors restricted to each part in a partition of $\mathcal{P}(X)$. This

property will help us prove the relation between lex-cel and anti-lex-cel in the forthcoming Theorem 5.

Lemma 3. *Given $P_1, \ldots, P_k$ a partition of $\mathcal{P}(X)$:*
$\mu(x) = \mu|_{P_1}(x) + \cdots + \mu|_{P_k}(x), \forall x \in X$
$\eta(S) = \eta|_{P_1}(S) + \cdots + \eta|_{P_k}(S), \forall S \in \mathcal{P}(X)$
where

- $\mu|_{P_j}(x) = (c_1, \ldots, c_l)$, *with* $c_i = |\{S \in \Sigma_i \cap P_j : x \in S\}|$, *stands for the profile vector of element x restricted to partition P_j; and*
- $\eta|_{P_j}(S) = (s_1, \ldots, s_r)$, *with* $s_i = |(\{S\} \cap P_j) \cap \Xi_i|$, *stands for the profile vector of set S restricted to partition P_j.*

Proof (Lemma 3)**.** *The proof is straightforward considering that μ and η are vectors of cardinalities and cardinalities satisfy that if $S \subseteq X$, and $P_1, \ldots P_k \subseteq X$ is a partition of X, then $|S \cap X| = \sum_i |S \cap P_i|$.*

The following result tells us that given a ranking $\succeq_e$ over the elements of X, the composition of anti-lex-cel and lex-cel over it results in the very same ranking $\succeq_e$.

Theorem 5. *Given a ranking $\succeq_e \in \mathcal{R}(X)$, $le(ale(\succeq_e)) = \succeq_e$.*

Proof (Theorem 5)**.** *Suppose $ale(\succeq_e) = \succeq$ and $le(ale(\succeq_e)) = \succeq'_e$. First, note that if $\succeq_e$ is such that $\forall x, y \in X$, $x \sim_e y$, then when applying ale to $\succeq_e$ we would have that $\forall S \in \mathcal{P}(X)$, $\eta(S) = (|S|)$ which would mean that the preference of a set only depends on its cardinality (not on its elements), and when applying back le to the obtained set ranking we would have that $\forall x, y \in X, \mu(x) = \mu(y)$ as all elements in X appear in the same number of sets of a certain cardinality. Therefore, to prove the theorem we can consider that $x, y \in X$, such that $x \succeq_e y$ and $x \nsim_e y$ and prove that $x \preceq'_e y$ is not possible. Now consider $XYS = \{S \in \mathcal{P}(X), x, y \in S\}$, $XS = \{S \in \mathcal{P}(X), x \in S, y \notin S\}$, $YS = \{S \in \mathcal{P}(X), x \notin S, y \in S\}$ and $RS = \{S \in \mathcal{P}(X), x, y \notin S\}$, note that these subsets form a partition of $\mathcal{P}(X)$ ($\mathcal{P}(X) = XYS \cup XS \cup YS \cup RS$, and XYS, XS, YS, RS disjoint), thus when applying le ($le(\succeq) = \succeq_e$) we will have: $\mu(x) = \mu|_{XYS}(x) + \mu|_{XS}(x) + \mu|_{YS}(x) + \mu|_{RS}(x)$ and $\mu(y) = \mu|_{XYS}(y) + \mu|_{XS}(y) + \mu|_{YS}(y) + \mu|_{RS}(y)$, and $\mu|_{XYS}(x) = \mu|_{XYS}(y)$, $\mu|_{RS}(x) = \mu|_{RS}(y)$, and $\mu|_{YS}(x) = \mu|_{XS}(y) = (0, \ldots, 0)$. Since, $y \succeq'_e x$, we have that $\mu(y) \geq_L \mu(x)$ which considering the equalities above implies that $\mu|_{YS}(y) \geq_L \mu|_{XS}(x)$, since $|YS| = |XS|$, this would mean that $\exists S \in YS$,*

such that $\forall S' \in XS$, $S \succeq S'$ *or alternatively* $\eta(S) \geq_L \eta(S')$. *But this is not possible, consider* $S' = S \setminus \{y\} \cup \{x\}$, *this set contains* x, *therefore* $S' \in XS$ *and since* $x \succeq_e y$ *and* $x \nsim_e y$, $\eta(S') >_L \eta(S)$, *which proves the theorem.*

Based on the theorem above, we can establish that anti-lex-cel is the inverse of lex-cel for a restricted family of rankings $ILE = \{\succeq \in \mathcal{R}(\mathcal{P}(X)), \exists \succeq_e \in \mathcal{R}(X) ale(\succeq_e) = \succeq\}$.

Corollary 1. $le|_{ILE}$ *is the inverse of ale.*

Proof (Corollary 1)**.** *In Theorem 5 we have seen that* $le(ale(\succeq_e)) = \succeq_e$, *which means that* $le|_{ILE}(ale(\succeq_e)) = \succeq_e$ *as we are only restricting the domain of le, now due to this restriction le is injective and exhaustive, as is ale, so they are inverses.*

4.7.4 Related results from the literature

Next we investigate the relationship between anti-lex-cel and a related result in the literature. In [Bossert et al., 1994], Bossert et al. study a particular preorder on $\mathcal{P}(X)$ associated to a linear order on X. Therefore, analogously to anti-lex-cel, [Bossert et al., 1994] studies a lifting of preferences from the element level (the linear order) to the set level. Interestingly, in this section we show that when fed with a linear order on X (which is a particular type of ranking), the output of anti-lex-cel is precisely the preorder on $\mathcal{P}(X)$ studied in [Bossert et al., 1994]. Hence, this shows the generality of anti-lex-cel.

Given an element ranking $\succeq_e \in \mathcal{R}(X)$ that is also anti-symmetric (i.e., $\succeq_e$ is a linear order), in [Bossert et al., 1994] the authors have introduced the following properties for a preorder (a transitive and reflexive relation) $\succeq$ of $\mathcal{P}(X)$ associated to $\succeq_e$ (see also [Barberà et al., 2004] for a general review of the related literature):

- Simple dominance (SD): for any $x, y \in Y$, $x \succeq_e y$ and $x \nsim_e y \Rightarrow \{x\} \succeq \{y\}$ and $\{x\} \nsim \{y\}$;
- Simple Monotonicity (SM): for any $x, y \in X$ with $x \neq y$, $\{x, y\} \succeq \{x\}$ and $\{x, y\} \nsim \{x\}$;
- Independence (IND): for any $S, T \in \mathcal{P}(X)$, for each $x \in X \setminus (S \cup T)$, $S \succeq T \Leftrightarrow S \cup \{x\} \succeq T \cup \{x\}$;

- Robustness for strict Preferences (RP): for any $S, T \in \mathcal{P}(X)$, for each $x \in X \setminus (S \cup T)$

$$\left.\begin{array}{l} S \succeq T \text{ and } S \nsim T, \\ y \succeq_e x \text{ and } y \nsim_e x \ \forall y \in S, \\ z \succeq_e x \text{ and } z \nsim_e x \ \forall z \in T \end{array}\right\} \Rightarrow S \succeq T \cup \{x\} \text{ and } S \nsim T \cup \{x\}.$$

A particular preorder on $\mathcal{P}(X)$ associated to a linear order $\succeq_e$ on X has been studied in [Bossert et al., 1994]. To define it, we need some more notations. Without loss of generality, it is assumed that the elements of any set $S = \{x_1, \ldots, x_s\} \in \mathcal{P}(X)$ are ordered in decreasing preference according to $\succeq_e$, that is, $x_1 \succeq_e x_2 \succeq_e \ldots \succeq_e x_s$.

Let $u_{\succeq_e} : X \to \mathbb{R}_{>0}$ be a real-valued function such that for all $x, y \in X$, $u_{\succeq_e}(x) \geq u_{\succeq_e}(y) \Leftrightarrow x \succeq_e y$. For $S = \{x_1, \ldots, x_s\} \in \mathcal{P}(X)$, let $v(S)$ be an $|X|$-dimensional vector such that

$$v_{u_{\succeq_e}}(S) = (u_{\succeq_e}(x_1), \ldots, u_{\succeq_e}(x_s), 0, \ldots, 0),$$

so, the last $|X| - |S|$ components of the vector are completed with zeros.

The relation $\succeq_{u_{\succeq_e}}$ on $\mathcal{P}(X)$ is then defined as follows:

$$S \succeq_{u_{\succeq_e}} T \Leftrightarrow v_{u_{\succeq_e}}(S) \geq_L v_{u_{\succeq_e}}(T).$$

for all $S, T \in \mathcal{P}(X)$. The following result, which has been proved in [Bossert et al., 1994], states that $\succeq_{u_{\succeq_e}}$ is the unique preorder of $\mathcal{P}(X)$ that satisfies the four properties.

Theorem 6. *[Bossert et al., 1994] Let $\succeq$ be a preorder on $\mathcal{P}(X)$. $\succeq$ satisfies SD, SM, IND and RP iff $\succeq = \succeq_{u_{\succeq_e}}$.*

We now prove that the total preorder $\succeq = ale(\succeq_e)$ also satisfies properties SD, SM, IND and RP.

Proposition 4. *Given a linear order $\succeq_e \in \mathcal{R}(X)$, the ranking $\succeq = ale(\succeq_e)$ satisfies SD, SM, IND and RP.*

Proof (Proposition 4)**.** *From Proposition 2 we know that f satisfies axioms 1, 3 and 4.*

From Axiom 1 on ale, we directly have that $\succeq$ satisfies SD. Since $\succeq$ is total, we have $\{x\} \sim \{x\}$.

Then, the proof that $\succeq$ satisfies SM follows by Axiom 3 on ale with $\{x\}$ in the role of both S and S' and $\bar{S} = \{y\}$.

To prove IND, simply notice that for all $S, T \in \mathcal{P}(X)$, $\eta(S) \geq_L \eta(T) \Leftrightarrow \eta(S \cup \{x\}) \geq_L \eta(T \cup \{x\})$ *for all* $x \in X \setminus (S \cup T)$.

Finally, the proof that $\succeq$ *satisfies RP follows by Axiom 4 on ale with* T *in the role of* S' *and* $\bar{S}' = \{x\}$.

To end this section, the following corollary formally establishes the relationship between anti-lex-cel and the results by Bossert et al. in [Bossert et al., 1994].

Corollary 2. *Let* $\succeq_e$ *be a linear order on* X*, then* $\succeq = \succeq_{u_{\succeq_e}}$ *where* $\succeq = ale(\succeq_e)$.

Proof (Corollary 2)**.** *The proof follows directly from Theorem 6 and Proposition 4.*

4.8 Solving the dominant set selection problem

With both lex-cel and anti-lex-cel , we can now address solving the dominant set selection problem. As shown in Figure 4.1, we will build the solution through three steps. In particular we will transform the input of the DSSP into a ranking over X using lex-cel, then we use anti-lex-cel to obtain a ranking over $\mathcal{P}(X)$. Thanks to the properties of *ale*, we prove that this ranking embodies dominance as in Definition 21, meaning that a set is dominant over its least preferred sets in the ranking. With this ranking and the feasibility function, we can find the solution as the more preferred set in the ranking that is feasible.

We start with the elements in X, the set of features F, their ranking $\succeq_F$ and the feature function $\mathfrak{f}$ relating elements to their features. In Section 4.6, we have adapted the lex-cel function to ground the ranking $\succeq_F$ to a ranking over X such that $le(\succeq_F) = \succeq_e$. Then, we apply anti-lex-cel to this ranking, $ale(\succeq_e) = \succeq$, to obtain a ranking over $\mathcal{P}(X)$. Thus, by composing lex-cel and anti-lex-cel, we can define a function $dom : \mathcal{R}(F) \to \mathcal{R}(\mathcal{P}(X))$ that transforms a ranking over the features in F to a ranking over the sets in $\mathcal{P}(X)$ as $dom(\succeq_F) = ale(le(\succeq_F))$. We show that the resulting ranking from the *dom* function embodies dominance as stated by the following theorem.

Theorem 7. *Let* X *be a set of elements,* F *a set of features,* $\succeq_F$ *a ranking over* F*, and a function* $\mathfrak{f}$ *relating elements to their features. For any pair*

$S, S' \in \mathcal{P}(X)$, S is dominant over $S' \Leftrightarrow S \succ S'$ ($S \succeq S'$ and $S \nsim S'$), where $dom(\succeq_F) = ale(le(\succeq_F)) = \succeq$.

To show that $\succeq$ embodies dominance and therefore prove the theorem, first we need an auxiliary lemma showing that $\succeq_e$ embodies element dominance.

Lemma 4. *$\succeq_e = le(\succeq_F)$ embodies element dominance, that is $\forall x, y \in X$, x is dominant over $y \Leftrightarrow x \succ_e y$.*

Proof (Lemma 4)**.** *To start we have a set of features F and a ranking over them $\succeq_F$, which can be represented in general as: $f_1^1 \sim_F \cdots \sim_F f_1^s \succ_F \cdots \succ_F f_k^1 \sim_F \cdots \sim_F f_k^r$ meaning that in $F/\sim_F$ the quotient order is $\Psi_1 \succ_F \cdots \succ_F \Psi_k$, with $\Psi_i = \{f_i^1, \dots f_i^q\}$ (q depends on i, for Ψ_1, $q = s$ and for Ψ_k, $q = r$).*

With these considerations, suppose $x \succ_e y$, by the definition of lex-cel, this means that $\mu(x) >_L \mu(y)$, which by the definition of μ means that $\exists \Psi_i$, such that x has a larger number of features in Ψ_i than y: $|\mathfrak{f}(x) \cap \Psi_i| > |\mathfrak{f}(y) \cap \Psi_i|$, while $\forall \Psi_j \succ_F \Psi_i$, x and y have the same number of features in Ψ_j: $|\mathfrak{f}(x) \cap \Psi_j| = |\mathfrak{f}(y) \cap \Psi_j|$. This means that x is Ψ_i-dominant over y, while for all $\Psi_j \succ_F \Psi_i$, they are Ψ_j-indifferent, which is the definition of x being dominant over y.

Now having proved that $x \succ_e y$ implies x dominant over y, we tackle the other direction. Suppose x dominant over y, if $x \prec_e y$ we have seen that would imply y dominant over x which contradicts our assumption, therefore $x \succeq_e y$, but note that if $x \sim_e y$, then proceeding as above we obtain $|\mathfrak{f}(x) \cap \Psi_i| = |\mathfrak{f}(y) \cap \Psi_i| \forall \Psi_i$, which would mean that neither x dominates y nor vice versa, contradicting our initial assumption. Therefore the only possibility is that $x \succ_e y$, proving the lemma.

Having seen that $\succeq_e$ embodies element dominance, we can build upon this result to prove the theorem stating that $\succeq$ embodies dominance.

Proof (Theorem 7)**.** *First, suppose $S \succ S'$, then since $ale(le(\succeq_F)) = \succeq$ we know that $\eta(S) >_L \eta(S')$, that is $\exists \Xi_i \in X/\sim_e$, such that $|S \cap \Xi_i| > |S' \cap \Xi_i|$ and $\forall \Xi_j \in X/\sim_e, \Xi_j \succ_e \Xi_i$, $|S \cap \Xi_j| = |S' \cap \Xi_j|$, note though that these equivalence classes are ordered with $\succ_e$ which we have seen in the lemma that embodies element dominance, therefore for $\Xi \succ_e \Xi'$, all elements of Ξ are dominant over all elements of Ξ', while the elements in the equivalence class are indifferent between them. With this consideration and the previous findings $|S \cap \Xi_i| > |S' \cap \Xi_i|$ and $\forall \Xi_j \in X/\sim_e, \Xi_j \succ_e \Xi_i$, $|S \cap \Xi_j| = |S' \cap \Xi_j|$*

means that S contains more elements in Ξ_i than S', while containing the same number of elements for more preferred equivalence classes. Therefore, considering $S = \{s_1, \ldots, s_{|S|}\}$, $S' = \{s'_1, \ldots, s'_{|S'|}\}$ and the permutation σ explained in Section 4.4 (and used in the definition of dominance), we have that each $s_{\sigma(1)}, \ldots, s_{\sigma(r)}$ is indifferent with its counterpart $s'_{\sigma(1)}, \ldots, s'_{\sigma(r)}$ for $r = \sum_{j<i} |S \cap \Xi_j| + |S' \cap \Xi_i|$, but $s_{\sigma(r+1)}$ dominates $s'_{\sigma(r+1)}$, because either $s_{\sigma(r+1)} \in \Xi_i$ and $s'_{\sigma(r+1)} \in \Xi_l$ with $l > i$ or $r+1 > |S'|$, which is the definition of S being dominant over S'.

Now having proved that $S \succ S'$ implies S dominant over S', we tackle the other implication. Suppose S dominant over S', if $S \prec S'$ we have seen that would imply S' dominant over S which contradicts our assumption, therefore $S \succeq S'$, but note that if $S \sim S'$, then $|S \cap \Xi| = |S' \cap \Xi| \forall \Xi \in X/\sim_e$ and therefore $s_{\sigma(i)}$ and $s'_{\sigma(i)}$ are in the same equivalence class $\forall i$, meaning that $s_{\sigma(i)}$ and $s'_{\sigma(i)}$ are indifferent $\forall i$, which means that S and S' are indifferent contradicting our assumption. Therefore the only possibility is that $S \succ S'$, which proves the theorem.

Corollary 3. *Consider a dominant set selection problem with a set of elements X, a set of features F, a ranking $\succeq_F$ over F, and a function $\mathfrak{f}$ relating elements to their features. Consider a set $S_{pref} \in \mathcal{P}(X), \phi(S_{pref}) = \top$, such that $\forall S' \in \mathcal{P}(X)$ with $S' \succeq S_{pref}$ and $S' \nsim S_{pref} \Rightarrow \phi(S') = \bot$. Then, S_{pref} is a solution to the dominant set selection problem.*

Proof (Corollary 3)**.** *This result follows directly from Theorem 7.*

With this result, note that to find the solution to the dominant set selection problem, the only step left to do after building $\succeq$ is to check for feasibility from the most preferred set in $\succeq$ to the least preferred set in $\succeq$ until we find the most preferred one that is feasible.

Nonetheless, note that building the ranking $dom(\succeq_F) = \succeq$ to solve the dominant set selection problem turns out to be rather costly. It requires to compute the η profile vector in Eq. 4.3 for every subset in $\mathcal{P}(X)$, with cost $O(2^{|X|})$, to subsequently order them following Eq. 4.4, which requires $O(2^{|X|} \cdot log(2^{|X|}))$ in the average case ($O(2^{2|X|})$ in the worst case). Therefore, finding the solution has worse than exponential complexity on the number of elements of X, hence hindering applicability.

With the intent of solving the DSSP through optimisation techniques, we show an alternative way of comparing sets of $\mathcal{P}(X)$ avoiding the cost of explicitly building $\succeq$. In particular, we propose a function, the so-called *preference function* $\mathfrak{p} : \mathcal{P}(X) \to \mathbb{N}$, which embodies the preferences in the $\succeq$

ranking while not needing to build it. Given a set $S \in \mathcal{P}(X)$, the larger its value by the preference function, the more preferred it is in $\succeq$. Importantly, we prove that this function adheres to $\succeq$, meaning that for all pairs of sets, the ranking between each pair is maintained by the function's output.

Let $S \in \mathcal{P}(X)$ be a set of elements. We define its profile vector as $\eta(S) = (c_1^S, \ldots, c_r^S)$, where $c_i^S = |S \cap \Xi_i|$ and $r = |X/\sim_e|$. From that, we compute the preference value of S as follows:

$$\mathfrak{p}(S) = \sum_{i=1}^{r} |S \cap \Xi_i|(\sum_{j=i+1}^{r} \mathfrak{p}(\Xi_j) + 1), \text{ where } \mathfrak{p}(\Xi_r) = |\Xi_r|. \tag{4.5}$$

Recall that Ξ_r is the equivalence class containing the least preferred elements of X. Notice that by applying equation 4.5, we can compute the preference of each equivalence class Ξ_i: $\mathfrak{p}(\Xi_i) = |\Xi_i|(\sum_{j=i+1}^{r} \mathfrak{p}(\Xi_j) + 1)$. Hence, the preference of the classes in the quotient order $\Xi_1 \succ_e \cdots \succ_e \Xi_r$ can be recursively computed starting from Ξ_r. Note also that $\mathfrak{p}(S) \geq 0$, and $\mathfrak{p}(S) \in \mathbb{N}$ for any S.

The $\mathfrak{p}$ preference function embodies the ranking $\succeq$ over sets in $\mathcal{P}(X)$, as we now prove through the following theorem.

Theorem 8. *Given two sets $S, S' \in \mathcal{P}(X)$, $S \succeq S' \Leftrightarrow \mathfrak{p}(S) \geq \mathfrak{p}(S')$.*

In order to prove the theorem, we firstly require some lemmas regarding the properties of $\mathfrak{p}$.

Lemma 5. $\mathfrak{p}(S) = \sum_{w=1}^{r} \mathfrak{p}(S \cap \Xi_w)$

Proof (Lemma 5)**.** *By applying equation 4.5 we obtain the preference of an equivalence class Ξ_w as $\mathfrak{p}(S \cap \Xi_w) = \sum_{i=1}^{r} |S \cap \Xi_w \cap \Xi_i|(\sum_{j=i+1}^{r} \mathfrak{p}(\Xi_j) + 1) = |S \cap \Xi_w|(\sum_{j=w+1}^{r} \mathfrak{p}(\Xi_j) + 1)$, since all equivalence classes are disjoint, meaning that $|S \cap \Xi_w \cap \Xi_i| = |\emptyset| = 0$ when $i \neq w$ and $|S \cap \Xi_w \cap \Xi_i| = |S \cap \Xi_w|$, when $i = w$. Now $\sum_{w=1}^{r} \mathfrak{p}(S \cap \Xi_w) = \sum_{w=1}^{r} \sum_{i=1}^{r} |S \cap \Xi_w \cap \Xi_i|(\sum_{j=i+1}^{r} \mathfrak{p}(\Xi_j) + 1)$ $= \sum_{w=1}^{r} |S \cap \Xi_w|(\sum_{j=w+1}^{r} \mathfrak{p}(\Xi_j) + 1) = \mathfrak{p}(S)$.*

Lemma 6. $\forall w, \mathfrak{p}(\Xi_w) \geq \mathfrak{p}(S \cap \Xi_w)$

Proof (Lemma 6)**.** *Since all equivalence classes are disjoint, from equation 4.5 we have that $\mathfrak{p}(\Xi_w) = |\Xi_w|(\sum_{j=w+1}^{r} \mathfrak{p}(\Xi_j) + 1)$ and $\mathfrak{p}(S \cap \Xi_w) = |S \cap \Xi_w|(\sum_{j=w+1}^{r} \mathfrak{p}(\Xi_j) + 1)$. Since $|\Xi_w| \geq |S \cap \Xi_w|$, then $\mathfrak{p}(\Xi_w) \geq \mathfrak{p}(S \cap \Xi_w)$.*

With these lemmas in mind, we now prove Theorem 8. In other words, we prove that $\mathfrak{p}$ embodies the $\succeq$ ranking.

Proof (Theorem 8)**.** *We divide the proof into three steps. First we prove two implications, and we subsequently show that these implications suffice to prove the theorem.*

$S \succ S' \Rightarrow \mathfrak{p}(S) > \mathfrak{p}(S')$: Say that $S \succ S'$. From Equation 4.4, we have that $S \succ S' \Leftrightarrow \eta(S) >_L \eta(S')$. By using the definition of η in Equation 4.3 we can write $\eta(S) >_L \eta(S')$ as $(c_1^S, \dots c_r^S) >_L (c_1^{S'}, \dots c_r^{S'})$ (where $c_i^S = |S \cap \Xi_i|$ and $c_i^{S'} = |S' \cap \Xi_i|$ $\forall i$). Now, by using the formalisation of the lexicographical order (see Definition 22), we have that $(c_1^S, \dots c_r^S) >_L (c_1^{S'}, \dots c_r^{S'})$, which implies that $\exists k \in \{1, \dots, r\}$, s.t. $\forall t < k, c_t^S = c_t^{S'}$ and $c_k^S > c_k^{S'}$. In other words, $\exists k \in \{1, \dots, r\}$ s.t. $|S \cap \Xi_k| > |S' \cap \Xi_k|$ and $\forall t < k, |S \cap \Xi_t| = |S' \cap \Xi_t|$ and therefore $\mathfrak{p}(S \cap \Xi_t) = \mathfrak{p}(S' \cap \Xi_t)$.

Next we prove that $\mathfrak{p}(S) > \mathfrak{p}(S')$. First, note that by considering Lemma 5, we have that $\mathfrak{p}(S) = \sum_{i=1}^{k-1} \mathfrak{p}(S \cap \Xi_i) + \sum_{i=k}^{r} \mathfrak{p}(S \cap \Xi_i) \geq \sum_{i=1}^{k-1} \mathfrak{p}(S \cap \Xi_i) + \mathfrak{p}(S \cap \Xi_k)$ and applying Lemma 5 and Lemma 6 we have that $\mathfrak{p}(S') = \sum_{i=1}^{k-1} \mathfrak{p}(S' \cap \Xi_i) + \sum_{i=k}^{r} \mathfrak{p}(S' \cap \Xi_i) \leq \sum_{i=1}^{k-1} \mathfrak{p}(S' \cap \Xi_i) + \mathfrak{p}(S' \cap \Xi_k) + \sum_{i=k+1}^{r} \mathfrak{p}(\Xi_i)$. Therefore, to prove that $\mathfrak{p}(S) > \mathfrak{p}(S')$ it suffices to prove that $\sum_{i=1}^{k-1} \mathfrak{p}(S \cap \Xi_i) + \mathfrak{p}(S \cap \Xi_k) > \sum_{i=1}^{k-1} \mathfrak{p}(S' \cap \Xi_i) + \mathfrak{p}(S' \cap \Xi_k) + \sum_{i=k+1}^{r} \mathfrak{p}(\Xi_i)$. This is equivalent to show that $\mathfrak{p}(S \cap \Xi_k) - \mathfrak{p}(S' \cap \Xi_k) - \sum_{i=k+1}^{r} \mathfrak{p}(\Xi_i) > 0$.

Now, using Equation 4.5, $\mathfrak{p}(S \cap \Xi_k) - \mathfrak{p}(S' \cap \Xi_k) - \sum_{i=k+1}^{r} \mathfrak{p}(\Xi_i) = |S \cap \Xi_k|(\sum_{j=k+1}^{r} \mathfrak{p}(\Xi_j) + 1) - |S' \cap \Xi_k|(\sum_{j=k+1}^{r} \mathfrak{p}(\Xi_j) + 1) - \sum_{i=k+1}^{r} \mathfrak{p}(\Xi_i) = (|S \cap \Xi_k| - |S' \cap \Xi_k|)(\sum_{j=k+1}^{r} \mathfrak{p}(\Xi_j) + 1) - \sum_{i=k+1}^{r} \mathfrak{p}(\Xi_i)$.

As shown above, we know that $|S \cap \Xi_k| > |S' \cap \Xi_k|$. From that, and since these sets' cardinalities are natural numbers, we obtain the following lower bound: $|S \cap \Xi_k| - |S' \cap \Xi_k| \geq 1$. Therefore, $(|S \cap \Xi_k| - |S' \cap \Xi_k|)(\sum_{j=k+1}^{r} \mathfrak{p}(\Xi_j) + 1) - \sum_{i=k+1}^{r} \mathfrak{p}(\Xi_i)) \geq \sum_{j=k+1}^{r} \mathfrak{p}(\Xi_j) + 1 - \sum_{i=k+1}^{r} \mathfrak{p}(\Xi_i)) = 1 > 0$.

Recall that we assumed that $S \succ S'$. Since we have managed to prove that $S \succ S'$ implies that $\mathfrak{p}(S \cap \Xi_k) - \mathfrak{p}(S' \cap \Xi_k) - \sum_{i=k+1}^{r} \mathfrak{p}(\Xi_i) > 0$, which in turn implies that $\mathfrak{p}(S) > \mathfrak{p}(S')$, then it is clear that $S \succ S' \Rightarrow \mathfrak{p}(S) > \mathfrak{p}(S')$.

$S \succ S' \Leftarrow \mathfrak{p}(S) > \mathfrak{p}(S')$: Suppose that $\mathfrak{p}(S) > \mathfrak{p}(S')$. If $S \prec S'$, then we have already shown above that $\mathfrak{p}(S) < \mathfrak{p}(S')$, which contradicts our initial assumption. If $S \sim S'$, then $\eta(S) = \eta(S')$, which means that $(c_1^S, \dots, c_r^S) = (c_1^{S'}, \dots, c_r^{S'})$, and therefore $\forall i$ $c_i^S = c_i^{S'}$. This means that $\forall i$ $|S \cap \Xi_i| = |S' \cap \Xi_i|$, which implies that $\mathfrak{p}(S) = \sum_{i=1}^{r} |S \cap \Xi_i|(\sum_{j=i+1}^{r} \mathfrak{p}(\Xi_j) + 1) = \sum_{i=1}^{r} |S' \cap \Xi_i|(\sum_{j=i+1}^{r} \mathfrak{p}(\Xi_j) + 1) = \mathfrak{p}(S')$. The fact that $\mathfrak{p}(S) = \mathfrak{p}(S')$ also contradicts our initial assumption $\mathfrak{p}(S) > \mathfrak{p}(S')$. Thus, we conclude that $\mathfrak{p}(S) > \mathfrak{p}(S') \Rightarrow S \succ S'$.

$S \succ S' \Leftrightarrow \mathfrak{p}(S) > \mathfrak{p}(S')$ suffices to prove the theorem: Note that we have proved that $S \succ S' \Leftrightarrow \mathfrak{p}(S) > \mathfrak{p}(S')$, then it trivially follows that $S \prec S' \Leftrightarrow \mathfrak{p}(S) < \mathfrak{p}(S')$. And these two cases imply that $S \sim S' \Leftrightarrow \mathfrak{p}(S) = \mathfrak{p}(S')$. Finally, $S \succ S' \Leftrightarrow \mathfrak{p}(S) > \mathfrak{p}(S')$ and $S \sim S' \Leftrightarrow \mathfrak{p}(S) = \mathfrak{p}(S')$ imply that $S \succeq S' \Leftrightarrow \mathfrak{p}(S) \geq \mathfrak{p}(S')$, which ends the proof of the theorem.

The preference function $\mathfrak{p}$ together with the results in Theorems 7 and 8 are key to cast the dominant set selection problem as the optimisation problem expressed by the following corollary.

Corollary 4. *Consider a dominant set selection problem with a set of elements X, a set of features F, a ranking $\succeq_F$ over F, and a function $\mathfrak{f}$ relating elements to their features. A feasible set $S_{max} \in \mathcal{P}(X)$ with maximum preference $\mathfrak{p}$ (see Eq. 4.5):*

$$S_{max} = \underset{S \in \mathcal{P}(X),\ \phi(S)=\top}{\arg\max} \mathfrak{p}(S) \tag{4.6}$$

is a solution to the dominant set selection problem.

Proof (Corollary 4)**.** *This result follows directly from Theorems 7 and 8 and Corollary 3.*

Building the whole ranking or computing the preference of all possible subsets is computationally costly. Nonetheless, in those cases in which the feasibility function can be translated into linear or quadratic constraints we can profit from the preference function $\mathfrak{p}$ to encode the DSSP into a binary integer program (BIP) and solve it with state of the art solvers. Thus, hereafter we will assume that we can translate the feasibility function into a set of linear or quadratic constraints C. The first step to encode the dominant set selection problem is to build the objective function of the BIP. The challenge here is to compactly represent the sets of $\mathcal{P}(X)$. Notice that for $X = \{x_1, x_2, x_3\}$, the set $S = \{x_1, x_2\}$ can be represented as $\{x_1, x_2, \neg x_3\}$, or as the binary vector $(1, 1, 0)$. In general, any $S \in \mathcal{P}(X)$ can be encoded as a vector $(d_1, \ldots, d_{|X|})$, where $d_i \in \{0, 1\}$ is the decision variable for element $x_i \in X$: if $d_i = 1$ means that x_i is in S, while $d_i = 0$ means x_i is not in S. Using the $(d_1, \ldots, d_{|X|})$ encoding for sets and following equation 4.5, in general we can obtain the preference of any set as $\sum_{i=1}^{r}(\sum_{x_w \in \Xi_i} d_w)(\sum_{j=i+1}^{r} \mathfrak{p}(\Xi_j) + 1)$, making use of the fact that $|S \cap \Xi_i| = \sum_{x_w \in \Xi_i} d_w$. Therefore, solving Problem 2 amounts to find-

ing the assignment of variables $(d_1, \ldots, d_{|X|})$ representing a feasible set with maximum preference. For that, we propose to solve the following BIP:

$$\max \sum_{i=1}^{r} \Big(\sum_{x_w \in \Xi_i} d_w \Big) \Big(\sum_{j=i+1}^{r} \mathfrak{p}(\Xi_j) + 1 \Big) \tag{4.7}$$

We require that the selected set satisfies the constraints in C. Thus, we consider these constraints in the encoding.

Observe that our BIP employs $|X|$ binary decision variables ($d_i \in \{0, 1\}$) and avoids the expensive, explicit computation of the ranking. Instead, it only requires to compute the preference of the equivalence classes Ξ_i ($\mathfrak{p}(\Xi_i)$). Since our objective function is always linear, if the constraints in C are either linear or quadratic, we can resort to off-the-shelf integer programming solvers like CPLEX or Gurobi. If the constraints in C are linear we would have to solve a typical BIP, whereas if they are quadratic we would have to solve a Binary Integer Quadratically Constrained Program. Appendix C details the algorithms to build the BIP and provides a link to an implementation.

In the next section we show the BIP that results when encoding the value-aligned norm selection problem introduced in Section 4.2. There we show an example of objective function together with a collection of linear constraints.

4.9 Application: Value-aligned norm selection

Now we count on the tools to solve the dominant set selection problem. Hereafter we will revisit the value-alignment problem described in Section 4.2 to exploit such tools for solving it, hence helping the decision maker.

Recall that in our value-alignment problem the decision maker is presented with a set of candidate norms N and a value system, which includes moral values and their preferences. Each of the norms in N is linked to some values, meaning that each norm promotes the values it is linked to. The problem for the decision maker is to select the subset of norms in N that better aligns with the values.

Prior to casting the problem faced by the decision maker as a particular type of dominant set selection problem, we must formally characterise: (i) the elements (norms); and (2) the features and preferences over features (value system).

First we start with the elements. For that, we base the definition of norms and their relationships in a simplification of the normative domain used in Chapter 3, which we call norm net. Hereafter we consider that norms are meant to regulate a multi-agent system composed of a set of agents Ag, with a finite set of actions $\mathcal{A}$ available to them.

Furthermore, we consider a propositional language $\mathcal{L}$ (with propositions in $\mathcal{P}$ and the logical operator "and"); and a set of states St. Like [Morales et al., 2015b; Morales et al., 2015a], we consider a state transition function that changes the state of the world when agents perform actions.

Definition 24 (Norm)**.** *A norm is a pair $\langle \varphi, \theta(a) \rangle$, where $\varphi \subseteq \mathcal{P}$ is a precondition in the language $\mathcal{L}$ (a subset of predicates with the logical operator "and"), $a \in \mathcal{A}$ is the regulated action, and $\theta \in \{Obl, Per, Prh\}$ is a deontic operator.*

Example 15. *Say that a country has to decide the norms to apply to its airport borders. The following norms are considered: n_1 permits to cross the border, n_2 prohibits to scan the baggage, n_3 obliges to show a passport, and n_4 obliges to scan the baggage. These norms can be represented formally as follows: n_1 as $\langle \emptyset$,Per(cross)$\rangle$; n_2 as $\langle \emptyset$,Prh(scan-bag)$\rangle$; n_3 as $\langle \emptyset$,Obl(show-passport)$\rangle$; and n_4 as $\langle \emptyset$,Obl(scan-bag)$\rangle$. Since we will use these norms in following examples, for the sake of readability we will note them omitting their precondition.*

Given a set of norms N, relationships between norms may hold. Thus, we identify norm exclusivity and generalisation as norm relations. Such relationships are relations over norms, henceforth noted as R_x and R_g respectively. Two norms n, n' are *mutually exclusive*, noted as $(n, n') \in R_x$, when they cannot be enacted at once; and they have a *direct generalisation* relation, noted as $(n, n') \in R_g$, when n is more general than n' and there is no other $n_{mid} \in N$, such that n is more general than n_{mid} being n_{mid} more general than n'. We note $A(n)/S(n)$ the ancestors/successors of n.

By putting together norms and their relations, we fully characterise the normative dimension of our decision space.

Definition 25. *A norm net is a structure $\langle N, R \rangle$, where N is a set of norms and $R = \{R_x, R_g\}$ is the set of exclusive, and generalisation relations.*

Henceforth we shall refer to any subset $\Omega \subseteq N$ as a *norm system*. We are interested in a particular type of norm systems: those that contain neither conflicting nor redundant norms. Thus, we characterise norm systems that avoid both conflicts and redundancy as *sound* norm systems.

Definition 26. *Given a norm net $\langle N, R\rangle$, a norm system $\Omega \subseteq N$ is sound iff it is both conflict-free and non-redundant, that is a norm system $\Omega \subseteq N$ is sound if for each $n_i, n_j \in \Omega$, $(n_i, n_j) \notin R_x$; $n_j \notin A(n_i)$; and $\forall n \in N$, such that $|\bar{S}(n)| > 1$, then $\bar{S}(n) \nsubseteq \Omega$, where $\bar{S}(n)$ are the direct successors of n ($\bar{S}(n) = \{n' \in N, (n, n') \in R_g\}$).*

At this point, notice that sound norm systems represent *feasible* norm systems. Therefore, when casting our value-alignment problem as a dominant set selection problem, checking for feasibility would consist in checking for soundness.

Example 16. *Consider the norms in Example 15, note that we cannot jointly allow to cross the border freely while obliging to show a passport, therefore n_1 and n_3 are incompatible norms. On the other hand, we cannot both oblige to scan a bag and prohibit it, making norms n_2 and n_4 incompatible as well. Thus, the norm net for the norms in Example 15 is the one in Figure 4.2*

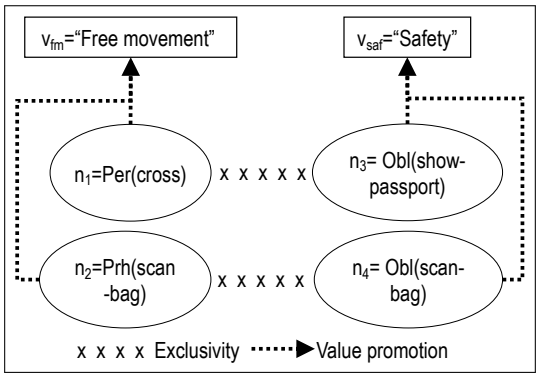


Figure 4.2: Example of candidate norms for border control along with their relations and their promotion of the free movement and safety values.

The features in this particular instance of a dominant set selection problem are values. Ethical reasoning typically involves a *value system*, that contains a set of moral values, which are principles that the society deems valuable. As noted in [Bench-Capon and Atkinson, 2009], within a value system, some values are preferred to others, and such preferences over moral values influence decision making. Therefore, the preferences over the moral values of a value system, together with the values themselves, have been identified as a core component for ethical reasoning in [Bench-Capon and Atkinson, 2009; Luo et al., 2017; Serramia et al., 2018a]. Formally,

Definition 27. *A value system is a pair $\langle V, \succeq_v \rangle$, where V stands for a non-empty set of values, and $\succeq_v$ is a ranking over the moral values in V.*

The definition of value system contains a ranking over moral values, and hence this is the ranking over features.

As required by the dominant set selection problem, we define a function linking the norms (elements) to their values (features). Thus:

Definition 28. *Given a norm net $\langle N, R \rangle$ and a value system $\langle V, \succeq_v \rangle$, we call value promotion function the function $\mathfrak{f} : N \to \mathcal{P}(V)$ that for each norm returns the set of values the norm promotes $\mathfrak{f}(n)$.*

In norm selection, a norm that does not promote any value and a value that is not promoted by any norm are irrelevant. Henceforth, we suppose that all norms promote at least one value ($\forall n \in N, \mathfrak{f}(n) \neq \emptyset$), and that all values are promoted by at least one norm ($\forall v \in V$, $\exists n \in N$, s.t. $v \in \mathfrak{f}(n)$).

Example 17. *Following Example 16, we observe that n_1 and n_2 promote free movement of people/goods ($\mathfrak{f}(n_1) = \mathfrak{f}(n_2) = \{v_{fm}\}$), whereas the rest of norms promote safety ($\mathfrak{f}(n_3) = \mathfrak{f}(n_4) = \{v_{saf}\}$), as depicted in Figure 4.2.*

With the definitions of the various structures of norms and values we can now define the problem faced by the decision maker that we want to solve.

Problem 3. *Given a norm net $\langle N, R \rangle$, a value system $\langle V, \succeq_v \rangle$ and a value promotion function $\mathfrak{f}$, we call value-aligned norm selection (VANS) problem, the problem of finding the set of norms $S \in \mathcal{P}(N)$, such that S is a sound norm system and any other norm system S', that dominates S is not sound.*

The value-aligned norm selection problem is a particular instance of the dominant set selection problem.

To solve the value-aligned norm selection problem, we proceed as detailed in Section 4.8. First, we apply lex-cel to the value ranking (as we have done in Example 13).

Once we obtain the ranking over N, we would just apply *ale* to obtain the ranking over all possible norm systems (as done in Example 14).

With the ranking over all norm systems in $\mathcal{P}(N)$, it remains to check for feasibility (in this case by checking for soundness).

Example 18. *An example value-aligned norm selection problem would be that where N are the norms in Example 15, with the relations R in Example 16 to assess feasibility and the value promotion function $\mathfrak{f}$ defined in Example*

17, supposing the value system $V = \{v_{fm}, v_{saf}\}$, with the value ranking (feature ranking) $v_{fm} \succeq_v v_{saf}$. Note that this structure is completely equivalent to the dominant set selection problem formulated in Example 12, therefore in this case we already know the $\mathcal{P}(N)$ ranking as we have found it in Example 14. Thus, the solution to this value-aligned norm selection problem is $\{n_1, n_2\}$ because it is the first sound (feasible) norm system in the ranking. Note that all norm systems with 3 or 4 norms contain a pair of exclusive norms and out of all the norm systems with 2 norms, $\{n_1, n_2\}$ is the most preferred one. In conclusion, we provided some norms to regulate an airport and by preferring freedom of movement of people/goods over security we selected the norms allowing to cross the border freely to both people and their belongings.

Nonetheless, as explained previously, the exhaustive approach followed above is computationally expensive. Instead, we can solve a VANS problem as an optimisation problem by encoding it into a BIP, as explained in Section 4.8. Building the objective function for this encoding is straightforward from Eq. 4.7. We must simply consider that there is one decision d_i variable for each norm $n_i \in N$. Moreover, we must add the following constraints to ensure that the resulting solution is feasible (the resulting norm system is sound):

- Mutually exclusive (incompatible) norms cannot be selected at once:

$$d_i + d_j \leq 1 \text{ for each } (n_i, n_j) \in R_x \tag{4.8}$$

- A norm cannot be simultaneously selected with any of its ancestors:

$$d_i + d_k \leq 1 \text{ for each } n_k \in A(n_i) \;\; 1 \leq i \leq |N| \tag{4.9}$$

- If a norm has more than one direct successor (we note $\bar{S}(n) = \{n' \in N, (n, n') \in R_g\}$), these direct successors cannot be simultaneously selected:

$$\text{If } |\bar{S}(n)| > 1 \text{ then } \sum_{n_j \in \bar{S}(n)} d_j < |\bar{S}(n)| \quad \text{for each } n \in N \tag{4.10}$$

Algorithms to build the BIP for a VANS problem and a link to an implementation can be found in appendix D.

Example 19. *Example 18, details a value-aligned norm selection problem and how to build a norm system ranking for it. Next we provide the BIP encoding for this example problem. First, we will build our objective function. Since the element ranking is* $n_1 \sim_e n_2 \succeq_e n_3 \sim_e n_4$ *(see Example 13), the quotient order is* $\Xi_1 \succ_e \Xi_2$*, we first compute* $\mathfrak{p}(\Xi_2) = |\Xi_2| = 2$*, because* $\Xi_2 = \{n_3, n_4\}$ *(we can also compute* $\mathfrak{p}(\Xi_1) = |\Xi_1| \cdot (\mathfrak{p}(\Xi_2) + 1) = 6$*, because* $\Xi_1 = \{n_1, n_2\}$*, though we do not need this number). Therefore, the objective function (following Equation 4.7) which we want to maximise is* $3d_1 + 3d_2 + d_3 + d_4$*. Since the norms of our running example have some relations between them, as shown in Figure 4.2, we consider the following constraints regarding exclusive norms:* $d_1 + d_3 \leq 1$*,* $d_2 + d_4 \leq 1$*. With this encoding the solution to the BIP is* $\{n_1, n_2\}$ *(the same we found in Example 18 using* $\succeq$*).*

Different value rankings may vary the selection of the value-aligned norm system. In previous examples we solved the problem of norm selection in an airport depicted in Figure 4.2 by considering the value ranking (feature ranking) $v_{fm} \succeq_v v_{saf}$. Subsequent Examples 20 and 21 explore how the solution changes for alternative value rankings ($v_{fm} \preceq_v v_{saf}$ and $v_{fm} \sim_v v_{saf}$).

Example 20. *Supposing* $v_{fm} \preceq_v v_{saf}$*, when grounding these preferences with lex-cel we obtain the norm ranking* $n_3 \sim_e n_4 \succeq_e n_1 \sim_e n_2$*. Then, lifting this ranking with anti-lex-cel we obtain:* $\{n_1, n_2, n_3, n_4\} \succeq \{n_1, n_3, n_4\} \sim \{n_2, n_3, n_4\} \succeq \{n_3, n_4\} \succeq \{n_1, n_2, n_3\} \sim \{n_1, n_2, n_4\} \succeq \{n_1, n_3\} \sim \{n_1, n_4\} \sim \{n_2, n_3\} \sim \{n_2, n_4\} \succeq \{n_3\} \sim \{n_4\} \succeq \{n_1, n_2\} \succeq \{n_1\} \sim \{n_2\}$*. In this case the solution is* $\{n_3, n_4\}$*, as it is the first sound (feasible) norm system in the ranking (on one hand, all norm systems containing 4 or 3 norms include a pair of exclusive norms and, on the other hand, out of all the norm systems with 2 norms it is the most preferred one).*

Example 21. *If* $v_{fm} \sim_v v_{saf}$*, when grounding these preferences with lex-cel we obtain the norm ranking* $n_1 \sim_e n_2 \sim_e n_3 \sim_e n_4$*. And lifting this ranking with anti-lex-cel we obtain:* $\{n_1, n_2, n_3, n_4\} \succeq \{n_1, n_2, n_3\} \sim \{n_1, n_2, n_4\} \sim \{n_1, n_3, n_4\} \sim \{n_2, n_3, n_4\} \succeq \{n_1, n_2\} \sim \{n_1, n_3\} \sim \{n_1, n_4\} \sim \{n_2, n_3\} \sim \{n_2, n_4\} \sim \{n_3, n_4\} \succeq \{n_1\} \sim \{n_2\} \sim \{n_3\} \sim \{n_4\}$*. In this case there are multiple solutions, namely:* $\{n_1, n_2\}$*,* $\{n_1, n_4\}$*,* $\{n_2, n_3\}$*, and* $\{n_3, n_4\}$*, as these are the most preferred sound (feasible) norm systems in the ranking (all norm systems containing 4 or 3 norms include a pair of exclusive norms, and the rest of norm systems with 2 norms are not sound). Notice that, when considering the values indifferently preferred, the possible solutions*

contain both permissive and restrictive norms for travellers. This was not the case in previous examples, as most permissive norms where selected when preferring freedom of movement over security and most restrictive norms when preferring security over freedom of movement.

4.10 Conclusions and limitations

This chapter describes tools to help decision makers select the most preferred set from a range of available options. We refer to this problem as the *dominant set selection problem* (DSSP). We solve it by considering both qualitative preference information over the features characterising the options and feasibility constraints.

Specifically, we propose to: first, ground the feature preferences to preferences over single objects; second, lift the object ranking into a ranking of all possible sets of options; and third, select the most preferred and feasible set of options. This requires the combination of existing results in the social choice literature, *lex-cel* [Bernardi et al., 2019], with our novel *anti-lex-cel*. To the best of our knowledge, the composition of ranking functions to obtain another ranking function (in our case we have composed *le* and *ale* to obtain *dom*) has not been previously explored and may pose an easier approach in those cases in which defining a ranking function directly is no easy task.

Moreover, we show how to encode the dominant set selection problem as a binary integer program (BIP) so that it can be solved with the aid of off-the-shelf solvers. We formally prove that solving the optimisation problem defined by the BIP encoding of the DSSP, we obtain a solution to the DSSP. Note that, for our set ranking we have found an encoding that allows to find the most preferred feasible set in the ranking while avoiding the computational cost of building it. We deem this strategy as promising when dealing with similar qualitative problems.

Thus, the overall contributions of this chapter are two-fold. Firstly, the formalisation of a novel qualitative decision-making problem: the dominant set selection problem. Secondly, the resolution of this problem, which requires the combination of methods from the literature as well as our novel anti-lex-cel method. Regarding anti-lex-cel, we also characterise it axiomatically, prove its uniqueness and show that it generalises former results in the literature. Furthermore, we use binary integer programming to encode (and solve) the problem. Finally, we illustrate our overall approach by means of the so-called value-alignment norm-selection problem.

This mostly amounts to the four contributions C8 - C11 from Section 1.3, which in turn answer the corresponding research questions in Section 1.2:

- Question Q8: How can we represent qualitatively the relations between norms and values? Through binary relations representing promotion or no promotion (contribution C8).
- Question Q9: How do we solve the value-aligned norm selection problem qualitatively? We obtain a ranking over norm systems by composing lex-cel and anti-lex-cel to then select the most preferred sound norm system in the ranking (C9).
- Question Q10: Are qualitative approaches computationally feasible? Yes, we provide a BIP encoding which allows to solve the problem avoiding the computational cost of building the norm system ranking (C10).
- Question Q11: How can we generalise the qualitative approach to value-aligned norm selection to use it in other multi-criteria decision making problems? This chapter has introduced and solved the dominant set selection problem, a general family of problems of which the VANS problem is a particular instance (C11).

However, the work in this chapter counts on some limitations that are worth discussing. First, the qualitative approach described in this chapter does not allow to express different degrees of relation between an element and features (or in general, criteria). Besides that, although we can express positive relationships between elements and criteria, we cannot express *negative* relationships. Thus, in terms of value-aligned norm selection, in this chapter we have assumed that norms relate to moral values through a binary promotion or no promotion relation. In general though, norms may promote or demote moral values in different degrees (as it is the case in the quantitative approach described in Chapter 3). Nonetheless, since lex-cel is not able to handle more expressive element-criterion relations, it is necessary to introduce a new grounding function that can consider them. We tackle all these limitations in the following Chapter 5.

Chapter 5

Graded qualitative value-aligned norm selection

5.1 Introduction

In terms of expressiveness, the qualitative approach of Chapter 4 has two important shortcomings: i) it cannot express that the relations between elements and criteria might have different degrees; and, ii) it cannot express negative relations between elements and criteria. Note that here the decision maker considers criteria, which are more general than the notion of feature employed in Chapter 4. Thus, in this chapter we overcome such shortcomings by providing the means to express both positive and negative relations with grades. Furthermore, we provide the means for decision makers to make qualitative decisions about multiple options while considering such graded relations between elements and criteria together with preferences over decision criteria. This major contribution is founded on the development of a novel ranking function that, unlike lex-cel, can handle graded relations between elements and criteria to ground the ranking over criteria to a ranking over elements.

As explained in Chapter 2, the social choice literature has studied several ranking functions, such as transforming a ranking over elements to a ranking over sets of elements (see [Barberà et al., 2004; Arlegi, 2003; Pattanaik and Peleg, 1984]) or transforming a ranking over sets of elements to a ranking over the elements themselves (see [Haret et al., 2018; Khani et al., 2019; Bernardi et al., 2019; Allouche et al., 2020]). In Chapter 4, we employed the lex-cel social ranking [Bernardi et al., 2019] to ground a ranking over criteria to a ranking over elements while considering binary relations between

elements and criteria. Unfortunately, none of the available ranking functions in the literature can handle graded relations between elements and criteria. Thus, this calls for the formalisation of a novel ranking function that is able to handle them as we do in this chapter. In particular, we introduce the so-called *multi-criteria based ranking* (MC ranking), a method to rank individual elements based on: i) how they relate to a collection of criteria; and ii) the preferences over these criteria. We introduce a particular MC ranking which we call MC lex-cel. This ranking mimics lex-cel, but it is able to handle graded relations between elements and criteria. Furthermore, it satisfies the dominance property.

Importantly, and as mentioned above, thanks to MC lex-cel we can tackle the solving of DSSPs with graded relations between elements and criteria following the general scheme proposed in Chapter 4 (which is graphically represented in Figure 4.1). In particular, we can transform the ranking over criteria to a ranking over sets of elements by composing an MC ranking (e.g. MC lex-cel) and a lifting function (e.g. anti-lex-cel). In the case of a VANS problem, whose norms and values have graded relations, we can obtain a ranking over all norm systems through the composition of MC lex-cel and anti-lex-cel.

The contributions of this chapter are:

- A formal definition of a new type of rankings: *multi-criteria based rankings* (MC rankings).

- A formal definition of dominance for MC rankings. This definition requires a non-straightforward adaptation from the desirable dominance property in social choice.

- A definition and study of the so-called *multi-criteria lex-cel*, a function to create MC rankings embodying dominance.

- A formal analysis showing the generality of our contributions with respect to recent results in the literature. Interestingly, *MC rankings* generalise social rankings [Moretti and Öztürk, 2017], while *multi-criteria lex-cel* generalises the lex-cel ranking function introduced in [Bernardi et al., 2019].

- A case study posing an ethical decision making problem that illustrates the use of MC rankings.

This chapter is organised as follows. We first introduce necessary background on order theory in Section 5.2, Section 5.3 formalises labels and label systems. Section 5.4 formalises MC rankings as well as the dominance property, and Section 5.5 introduces MC lex-cel. Section 5.6 studies the relation of our MC rankings and MC lex-cel with the social choice literature. Section 5.7 analyses a case study in ethical decision making, and Section 5.8 discusses the conclusions of this chapter. Recall that, for the ease of readability, we have included a List of Notation and Symbols.

5.2 Background: Recap of some concepts from Chapter 4

This section recaps some concepts from Chapter 4 needed for the work in this chapter. Note that the notation here may be different from that of Chapter 4, hence the reader might find useful the List of Notation and Symbols.

Let X be a set of objects. A binary relation $\succeq$ on X is said to be: *reflexive*, if for each $x \in X$, $x \succeq x$; *transitive*, if for each $x, y, z \in X$, ($x \succeq y$ and $y \succeq z$) $\Rightarrow x \succeq z$; *total*, if for each $x, y \in X$, $x \succeq y$ or $y \succeq x$; *antisymmetric*, if for each $x, y \in X$, $x \succeq y$ and $y \succeq x \Rightarrow x = y$. We can define preferences among the elements of X by means of binary relations. Moreover, we can categorise the type of preferences depending on the properties they hold as follows.

Definition 29 (Preorder, ranking, linear order and partial order)**.** *A preorder (or quasi-ordering) is a binary relation $\succeq$ that is reflexive and transitive. A preorder that is also total is called total preorder or ranking. A total preorder that is also antisymmetric is called a linear order. A preorder that is antisymmetric but not total is called a partial order.*

We build a lexicographical order for two tuples by comparing them element-wise from left to right. While the elements in both tuples are the same, we move to the next position on the tuples. We traverse the tuples until two elements differ (one is more preferred than the other). The more preferred tuple is the one with the more preferred element. If all elements are the same, the tuples are deemed equal. Formally:

Definition 30. *Given two tuples t, t', with $t = (t_1, \dots t_q)$ and $t' = (t'_1, \dots t'_q)$, we define the lexicographical order of tuples $\geq_{lex}$ as: $t \geq_{lex} t' \Leftrightarrow$ if either $t = t'$ or $\exists i \in \{1, \dots q\}$ s.t. $t_i > t'_i$ and $\forall j < i, t_j = t'_j$ (note that $t =_{lex} t' \Leftrightarrow t = t'$).*

The lexicographical order for tuples is used in the definition of the lex-cel ranking [Bernardi et al., 2019]. Let X be a set of elements, and $\succeq_S$ a ranking over the power set $\mathcal{P}(X)$, then lex-cel builds an element ranking $\succeq_e$ by means of assigning a tuple to each element (noted $\theta(x)$). To build this tuple, consider the quotient set $\mathcal{P}(X)/\sim_S$ with quotient order $\Sigma_1 \succ_S \Sigma_2 \succ_S \cdots \succ_S \Sigma_q$. Then, $\theta(x)$ is defined as:

$$\theta(x) = (x_1, \ldots, x_q) \text{ where } x_i = |S \in \Sigma_i : x \in S| \tag{5.1}$$

Lex-cel ranks elements in X by comparing lexicographically their corresponding θ tuples: $x \succeq_e y \Leftrightarrow \theta(x) \geq_{lex} \theta(y)$.

5.3 Relating elements to criteria

As explained in the introduction, in this chapter we consider that elements relate to criteria with different degrees. Not only that, but these relations can be positive (when an element aligns with a criterion), neutral, or negative (when an element is detrimental to a criterion). To specify these relations, with different degrees, we will use labels. Next, we introduce the notions of: label system (the object that defines labels for relating elements and criteria and their semantics), and labelling (a function to relate elements to criteria through labels).

A label system contains a set of labels and an order over them to establish their grading. Its labels must contain a neutral label between positive and negative labels. Positive labels are those more preferred than the neutral label, whereas negative labels are those less preferred than the neutral label. In terms of label grading, the more preferred a positive label, the higher the alignment degree between an element and a criterion it is meant to represent. Conversely, the less preferred a negative label, the higher the detrimental degree.

Definition 31 (Label system)**.** *A label system is a pair $\langle L, >_L \rangle$, where L is a set of labels, $>_L$ is a linear order over L. A label system includes a neutral label*[1] *$l_0 \in L$. Labels more preferred than l_0 are positive labels, whereas those less preferred than l_0 are negative labels.*

Note that a label system does not need to have a negative label for each positive label. In fact, it might only have positive labels. However,

[1] l_0 is unique because $>_L$ is a linear order. Thus, if there were two neutral labels, one would be necessarily preferred over the other.

a label system with more labels of one type than of another one hinders comparing labels. For example, given $l_2 >_L l_1 >_L l_0 >_L l_{-1}$, it is unclear whether the positive counterpart of l_{-1} is l_2 because both labels are the most extreme ones, or if it is l_1, because they are equally separated from l_0. To avoid these uncertainties, we focus on a particular type of label systems: the so-called symmetric label systems, for which each positive label has a negative counterpart. To ease their definition, we first introduce two auxiliary functions, namely the sign and strength of a label, which also provide a useful notation for the forthcoming sections.

Given a label system, the sign function signals if a label is positive (1), negative (-1), or the neutral label (0).

$$sgn(l) = \begin{cases} 1 & \text{if } l >_L l_0 \\ 0 & \text{if } l = l_0 \\ -1 & \text{if } l_0 >_L l \end{cases} \tag{5.2}$$

The strength function characterises the label's degree of preference in the label system order. In particular, we consider that, given a label l, the more labels between l and l_0 in the label order, the larger its strength. Formally:

$$stg(l) = \begin{cases} |\{l' \in L, l \geq_L l' >_L l_0\}| & \text{if } l >_L l_0 \\ 0 & \text{if } l = l_0 \\ |\{l' \in L, l_0 >_L l' \geq_L l\}| & \text{if } l_0 >_L l \end{cases} \tag{5.3}$$

Definition 32 (Symmetric label system)**.** *A label system $\langle L, >_L \rangle$ is symmetric if $\forall l \in L$, $\exists l' \in L$, such that $sgn(l) = -sgn(l')$ and $stg(l) = stg(l')$.*

Symmetric label systems have the same amount of positive and negative labels. Note that, without loss of generality, any label system can be transformed into a symmetric label system by simply adding "dummy" labels. Hereafter, we only consider symmetric label systems. Also, we can uniquely note each label in the label system as $l_{sgn(l)\cdot stg(l)}$ (for example, we note l_{-2} the label of sign -1 and strength 2).

Example 22. *An example symmetric label system is $L = \{l_2, l_1, l_0, l_{-1}, l_{-2}\}$ with order $l_2 >_L l_1 >_L l_0 >_L l_{-1} >_L l_{-2}$. Note for example, that l_{-2} is the label of sign $sgn(l_{-2}) = -1$ and strength $str(l_{-2}) = 2$.*

Using a label system, a decision maker can relate an element with a criterion by means of a labelling function.

Definition 33. *Given a set of elements X, a set of criteria C, and a label system $\langle L, >_L \rangle$, a labelling is a function $\lambda : X \times C \to L$ that assigns a label in L to each pair of element in X and criterion in C, hence establishing the relation between the element and the criterion. We note as $\mathcal{L}(X, C)$ the set of all possible labellings over X and C.*

If $\lambda(x, c) = l$, we say that element x is related to criterion c with degree l. From equation 5.3, we also say that the strength of the relation is $stg(l)$.

Example 23. *Consider the set of elements $X = \{x_1, \ldots, x_5\}$ and a set of criteria $C = \{c_1, \ldots, c_4\}$ and the label system of Example 22. An example labelling would be:*

λ	x_1	x_2	x_3	x_4	x_5
c_1	l_{-2}	l_1	l_1	l_0	l_0
c_2	l_2	l_1	l_0	l_0	l_0
c_3	l_0	l_{-1}	l_0	l_{-1}	l_2
c_4	l_0	l_{-1}	l_0	l_{-2}	l_0

5.4 Multi-criteria based rankings

As mentioned previously, we assume that the decision maker establishes a set of criteria and knows the preferences over them. We have learnt in Section 5.3 how to relate elements with criteria. Our goal is to build a ranking over the single elements in X from: (i) the relationships between elements and criteria; and (ii) the preferences over criteria. We will call such ranking a *multi-criteria based ranking* (MC ranking). In this section, we formally define it, as well as the fundamental notion of dominance for MC rankings.

An MC ranking considers a set of elements X, a set of criteria C, a ranking $\succeq_C$ over the criteria, and a labelling λ relating elements to criteria, and provides a ranking $\succeq$ over the single elements in X. Formally:

Definition 34. *Given a set of elements X, a set of criteria C and a set of labellings $\mathcal{L}(X, C)$, an MC ranking is a function $mcr : \mathcal{L}(X, C) \times \mathcal{R}(C) \to \mathcal{R}(X)$ that associates to any labelling $\lambda \in \mathcal{L}(X, C)$ relating elements with criteria and any ranking $\succeq_C \in \mathcal{R}(C)$ over criteria, another ranking $mcr(\lambda, \succeq_C) \in \mathcal{R}(X)$ over the elements of X.*

MC rankings call for the introduction of a novel notion of dominance between elements in X, as it is common in the literature (e.g. [Barberà et al.,

2004] [Moretti and Öztürk, 2017]). Such notion of dominance must ensure that the ranking on elements is strictly based on the ranking over criteria. However, defining dominance for MC rankings is intricate due to the richness of our labelling approach. Informally, our notion of dominance must require that an MC ranking function ranks the elements in X taking into account: the element-criterion relations, their associated labels, and the criteria preferences. Thus, the more preferred a criterion with which an element relates positively, the more preferred the element. Conversely, the more preferred the criterion with which an element relates negatively, the less preferred the element. The higher the degree of the labels on these positive/negative relations the more/less preferred the element. Furthermore, the larger the number of positive relations and the lower the number of negative relations for an element, the more preferred the element in the ranking.

Our notion of dominance between two elements is founded on the dominance within each equivalence class of criteria resulting from the ranking $\succeq_C$ over criteria. Thus, consider the quotient set of criteria $C/\sim_C$ with equivalence classes $\kappa_1, \ldots, \kappa_r$, and quotient order $\succ_C$. Notice that the criteria within each equivalence class $\kappa \in C/\sim_C$ are equally preferred. Given an equivalence class κ, our first aim is to establish whether an element $x \in X$ is κ-dominant (dominant within the scope of equivalence class κ) over another element $y \in X$. An element will be κ-dominant over another if it relates more strongly (and positively) with the criteria in κ than another one.

To define κ-dominance, we resort to an auxiliary function, the so-called net alignment function. Given an element x and a strength s, the net alignment function aggregates the positive and negative relations of x with the criteria in κ with strength s. Thus, the larger the net alignment, the more positive relations of strength s relating x and κ, the lower the net alignment the more negative relations of strength s relating x and κ. Formally, the net alignment function (noted na) is defined as the difference between the number of criteria positively and negatively related with x with strength s:

Definition 35. *Consider a criteria equivalence class $\kappa \in C/\sim_C$ and a relation strength $s \neq 0$. We define the* net alignment *of strength s of element x with class κ as:*

$$na(x, \kappa, s) = |\{c \in \kappa : \lambda(x, c) = l_s\}| - |\{c \in \kappa : \lambda(x, c) = l_{-s}\}| \qquad (5.4)$$

Let $s_{max} = \max_{l \in L} stg(l)$ be the maximum strength of the labels in the label system. Then, κ-dominance is defined as:

Definition 36. *Given two elements $x, y \in X$, a set of criteria C, a ranking over these criteria $\succeq_C$, a symmetric label system $\langle L, >_L \rangle$ and a criteria equivalence class $\kappa \in C/\sim_C$, we say that x is κ-dominant over y if $\exists s \in \{1, \ldots, s_{max}\}$, s.t. $na(x, \kappa, s) > na(y, \kappa, s)$ and $\forall s' > s$, we have $na(x, \kappa, s') = na(y, \kappa, s')$. If $\forall s \in \{1, \ldots, s_{max}\}$, $na(x, \kappa, s) = na(y, \kappa, s)$, we say x and y are κ-indifferent.*

Example 24. *Following Example 23, consider the criteria preferences $c_1 \succeq_C c_2 \sim_C c_3 \sim_C c_4$. The quotient set is $C/\sim_C = \{\kappa_1, \kappa_2\}$, with $\kappa_1 = \{c_1\}$ and $\kappa_2 = \{c_2, c_3, c_4\}$, and quotient order $\kappa_1 \succ_C \kappa_2$. Note that $na(x_1, \kappa_1, 2) = -1$, while for the rest of elements in X, their net alignment of strength 2 with κ_1 is 0, which is greater than -1. Thus, we say that x_2, x_3, x_4, and x_5 are κ_1-dominant over x_1.*

Using the concept of κ-dominance, we define dominance considering all equivalence classes in $C/\sim_C$ (and their quotient order $\succ_C$). We say that x is dominant over y if for a given criteria equivalence class x is κ-dominant over y, while for more preferred equivalence classes they are κ-indifferent.

Definition 37. *Given two elements $x, y \in X$ with criteria in C and a ranking over criteria $\succeq_C$, we say that x is dominant over y if there is a criteria equivalence class $\kappa \in C/\sim_C$, such that: (i) x is κ-dominant over y; and (ii) $\forall \kappa' \in C/\sim_C$, such that $\kappa' \succ_C \kappa$, x and y are κ'-indifferent. If neither element dominates the other (they are κ-indifferent $\forall \kappa \in C/\sim_C$), we say that they are indifferent.*

Example 25. *From κ-dominance in Example 23, we conclude that, x_2, x_3, x_4 and x_5 are dominant over x_1,because they are κ_1-dominant and κ_1 is the most preferred class.*

5.5 Multi-criteria lex-cel

Next, we introduce multi-criteria lex-cel (MC lex-cel), an MC ranking function. For each element in X, MC lex-cel builds a tuple, the so-called *multi-criteria profile* (MC profile), which summarises the relations between the element and the criteria. Then, MC lex-cel ranks the elements in X by comparing their MC profiles lexicographically. In Section 5.5.1, we describe how to build MC profiles. Section 5.5.2 defines MC lex-cel and proves that it embodies the dominance property in Definition 37.

5.5.1 Building MC profiles for elements

We will build the MC profile of an element $x \in X$ as a tuple $\mu(x)$ that is meant to summarise the relations of that particular element with all the criteria at hand.

Overall, we build an MC profile for an element through a nested process: (1) we start considering criteria preferences, from more preferred to less preferred; (2) thereafter, we delve into each equivalence class to consider the strengths of relations, from stronger to weaker.

Formally, we build an MC profile by considering the quotient set $C/\sim_C$, where $\kappa_1, \ldots, \kappa_q \in C/\sim_C$ are criteria equivalence classes with quotient order $\kappa_1 \succ_C \cdots \succ_C \kappa_q$. Each κ_i contains the i-th most preferred criteria.

We compose the MC profile $\mu(x)$ of an element x, from its equivalence class profiles $\mu(x, \kappa_1), \ldots, \mu(x, \kappa_q)$. An equivalence class profile $\mu(x, \kappa_i)$ summarises the relations between x and the equivalence class κ_i. We want to ensure that criteria preferences are satisfied according to $\succ_C$. Thus, we compose the MC profile $\mu(x)$ by considering that the relationships with more preferred criteria are positioned further on the left[2] of $\mu(x)$ as follows:

$$\mu(x) = (\mu(x, \kappa_1), \ldots, \mu(x, \kappa_q)) \tag{5.5}$$

Within an equivalence class κ, all criteria are indifferently preferred. Thus, what distinguishes the relations between x and κ here is their strength and sign. Recall that for each strength s, the net alignment function na aggregates the number of positive relations of strength s with the number of negative relations of strength s. Hence, we build the equivalence class profile of x for class κ out of the net alignments between x and κ for all non-zero[3] strengths, namely from $na(x, \kappa, 1), \ldots, na(x, \kappa, s_{max})$. Since we prefer strong relations over weak ones, the net alignments representing greater strengths, are positioned further to the left[2]. Therefore, the equivalence class profile is a tuple containing the net alignments of x and κ arranged from left to right in descending order of strength:

$$\mu(x, \kappa) = (na(x, \kappa, s_{max}), \ldots, na(x, \kappa, 1)), \tag{5.6}$$

[2]Recall that the MC lex-cel function in Section 5.5.2 applies a lexicographical order over $\mu(x)$, and thus left indicates greater preference.

[3]A strength zero relation (labelled l_0) represents that the element is neutral to the criterion. In other words, the element does not affect the criterion (the element neither aligns with nor is detrimental to the criterion). Hence, we should not take into account these relations in the MC profile.

where, as for Definition 36, $s_{max} = \max_{l \in L} stg(l)$ is the maximum strength of the label system. For the sake of understanding, next we illustrate how to build the MC profiles for the elements of our running example.

Example 26. *Following our running example, note that the criteria preferences $c_1 \succeq_C c_2 \sim_C c_3 \sim_C c_4$ imply that $C/\sim_C = \{\kappa_1, \kappa_2\}$, with $\kappa_1 = \{c_1\}$, $\kappa_2 = \{c_2, c_3, c_4\}$, and $\kappa_1 \succ_C \kappa_2$. Thus, $\forall x \in X$, $\mu(x) = (\mu(x, \kappa_1), \mu(x, \kappa_2))$. Now, the label system that we have considered contains labels of strength 2, 1 (and 0). Hence, since the maximum strength is 2, $\mu(x, \kappa) = (na(x, \kappa, 2), na(x, \kappa, 1))$ for each element x. In particular, regarding x_1, we have that $na(x_1, \kappa_1, 2) = -1$, because there is one label l_{-2} relating x_1 to κ_1, whereas $na(x_1, \kappa_1, 1) = 0$, because there are no labels of strength 1 relating x_1 to κ_1. By applying equation 5.6 above, we have that $\mu(x_1, \kappa_1) = (-1, 0)$. On the other hand, we have that $na(x_1, \kappa_2, 2) = 1$, because there is one label l_2 relating x_1 to κ_2, while $na(x_1, \kappa_2, 1) = 0$, because there are no labels of strength 1 relating x_1 to κ_2. Again, by means of equation 5.6, we have that $\mu(x_1, \kappa_2) = (1, 0)$. With these two equivalence class profiles, we can now apply equation 5.5 to build the MC profile of x_1 as $\mu(x_1) = ((-1, 0), (1, 0))$. By following an analogous procedure, we obtain the MC profiles for the rest of elements of X:*

$\mu(x_2) = ((0, 1), (0, -1))$ $\quad$ $\mu(x_3) = ((0, 1), (0, 0))$
$\mu(x_4) = ((0, 0), (-1, -1))$ $\quad$ $\mu(x_5) = ((0, 0), (1, 0))$

5.5.2 The multi-criteria lex-cel ranking function

Since the MC profile of an element $x \in X$ encodes its alignment with the criteria in C, we propose to compare elements in X by comparing their MC profiles by means of the lexicographical order. This is precisely what our multi-criteria lex-cel function captures as follows:

$$x \succeq y \Leftrightarrow \mu(x) \geq_{lex} \mu(y).$$

Definition 38. *Given a set of elements X, a set of criteria C and a set of labellings $\mathcal{L}(X, C)$, the multi-criteria lex-cel (MC lex-cel) function $mclex : \mathcal{L}(X, C) \times \mathcal{R}(C) \to \mathcal{R}(X)$ associates to any labelling $\lambda \in \mathcal{L}(X, C)$ and any ranking $\succeq_C \in \mathcal{R}(C)$, another ranking $\succeq = mclex(\lambda, \succeq_C) \in \mathcal{R}(X)$ such that for any two elements $x, y \in X$:*

$$x \succeq y \Leftrightarrow \mu(x) \geq_{lex} \mu(y), \tag{5.7}$$

where $>_{lex}$ the lexicographical order in Definition 30.

Notice that $\mu(x) >_{lex} \mu(y) \Leftrightarrow \exists \kappa \in C/\sim_C$, such that $\forall \kappa' \succ_c \kappa\ \mu(x,\kappa') = \mu(y,\kappa')$ and $\mu(x,\kappa) >_{lex} \mu(y,\kappa)$.

Example 27. *After applying MC lex-cel to the MC profiles obtained in Example 26, we obtain the following element ranking:* $x_3 \succ x_2 \succ x_5 \succ x_4 \succ x_1$.

Our purpose now is to prove that MC lex-cel embodies dominance according to Definition 37. Before that, we need an intermediary result showing that the lexicographical ordering of criteria profile captures κ-dominance within criteria equivalence classes.

Lemma 7. *Consider two elements* $x, y \in X$*, and a criteria equivalence class* $\kappa \in C/\sim_C$*, then* $\mu(x,\kappa) >_{lex} \mu(y,\kappa) \Leftrightarrow x$ *κ-dominant over* y*. Otherwise, we have that* $\mu(x,\kappa) = \mu(y,\kappa) \Leftrightarrow x$ *and* y *are κ-indifferent.*

Proof (Lemma 7)**.** *Suppose that* $\mu(x,\kappa) >_{lex} \mu(y,\kappa)$*. Since* $\mu(x,\kappa) = (na(x,\kappa,s_{max}),\ldots,na(x,\kappa,1))$ *and* $\mu(y,\kappa) = (na(y,\kappa,s_{max}),\ldots,na(y,\kappa,1))$*, being* $\mu(x,\kappa)$ *lexicographically greater than* $\mu(y,\kappa)$ *means that* $\exists s \in \{1,\ldots,s_{max}\}$*, such that* $na(x,\kappa,s) > na(y,\kappa,s)$*, and* $\forall s' > s$*,* $na(x,\kappa,s') = na(y,\kappa,s')$*. Notice that this is precisely the definition of* x *κ-dominant over* y *(Definition 36). Now, if* $\mu(x,\kappa) = \mu(y,\kappa)$*, then* $\forall s \in \{1,\ldots,s_{max}\}$*,* $na(x,\kappa,s) = na(y,\kappa,s)$*, which is the definition of* x *and* y *being κ-indifferent. Consider now the other direction of the implication, and then suppose that* x *is κ-dominant over* y*. In this case, neither* $\mu(y,\kappa) >_{lex} \mu(x,\kappa)$ *nor* $\mu(x,\kappa) = \mu(y,\kappa)$ *can be true because, by the already proved implication, it would contradict our assumption. Therefore, the only possibility is that* $\mu(x,\kappa) >_{lex} \mu(y,\kappa)$*. The same reasoning applies if we suppose* y *is κ-dominant over* x*, or* x *and* y *are κ-indifferent.*

With the help of Lemma 7, we are ready to prove that multi-criteria lex-cel embodies dominance.

Theorem 9. *MC lex-cel embodies dominance, that is, if* $mclex(\succeq_C) = \succeq$*, then for* $x, y \in X$*, we have that* $x \succ y \Leftrightarrow x$ *is dominant over* y*.*

Proof (Theorem 9)**.** *Suppose that* $x \succ y$*. Since* $\succ$ *has been obtained through MC lex-cel, we know that* $\mu(x) >_{lex} \mu(y)$*. This means that* $\exists \kappa \in C/\sim_C$*, such that* $\mu(x,\kappa) >_{lex} \mu(y,\kappa)$ *and* $\forall \kappa' \succ_C \kappa$*,* $\mu(x,\kappa') = \mu(y,\kappa')$*. Thanks to Lemma 7, we have seen that this means that* x *is κ-dominant over* y *and* $\forall \kappa' \succ_C \kappa$*,* x *and* y *are κ'-indifferent, which is the definition of dominance of* x *over* y*. Similarly, if* $\mu(x) = \mu(y)$*, then* $\forall \kappa$*,* $\mu(x,\kappa) = \mu(y,\kappa)$*, and thus*

x and y are κ-indifferent, meaning that they are indifferent. As to the other direction of the proof, say that x is dominant over y. If $\mu(x) <_{lex} \mu(y)$, it would imply that y is dominant over x, which contradicts our assumption. Similarly, if $\mu(x) = \mu(y)$, x and y should be indifferent, again contradicting our assumption. Therefore, the only possibility is that $\mu(x) >_{lex} \mu(y)$. The same reasoning applies if we suppose that y is dominant over x or x and y are indifferent.

5.6 MC ranking and social ranking

In this section we explore the relation between our MC ranking and the social ranking introduced in [Moretti and Öztürk, 2017]. We show that any social ranking can be encoded as an MC ranking, but that is not true the other way around. Therefore, the MC ranking is more general. Furthermore, we also show that our MC lex-cel generalises the lex-cel social ranking solution introduced in [Bernardi et al., 2019].

The social ranking [Moretti and Öztürk, 2017] considers a set of elements X, and a ranking over coalitions of these elements, namely a ranking over $\mathcal{P}(X)$. The purpose of a social ranking is to transform or ground this power set ranking into a ranking over X. Formally:

Definition 39. *A social ranking is a function $sr : \mathcal{R}(\mathcal{P}(X)) \to \mathcal{R}(X)$, which transforms a ranking over $\mathcal{P}(X)$ into a ranking over the elements of X.*

The goal of a social ranking and of an MC ranking is the same: to obtain a ranking over X. Nonetheless, both rankings start from different points. While a social ranking considers a ranking over the power set of X, an MC ranking considers criteria, a ranking over criteria and a labelling relating elements to criteria. Note though that it is possible to define a function that transforms a social ranking into an MC ranking. Since the input of *sr* is in $\mathcal{R}(\mathcal{P}(X))$ and the input of *mcr* is in $\mathcal{R}(C) \times \mathcal{L}(X, C)$, we propose a function $t : \mathcal{R}(\mathcal{P}(X)) \to \mathcal{R}(C) \times \mathcal{L}(X, C)$ to transform a social ranking's input into an MC ranking's input. Therefore, supposing X a set of elements and $\succeq_P$ a ranking over $\mathcal{P}(X)$, this transformation function is such that $t(\succeq_P) = (\lambda, \succeq_C)$. We build function t as follows:

1. We transform the sets in $\mathcal{P}(X)$ into criteria: $C = \{c_S, \forall S \in \mathcal{P}(X)\}$.
2. We obtain the ranking over criteria as a direct translation of the ranking over sets: $c_S \succeq_C c_{S'} \Leftrightarrow S \succeq_P S'$.

3. Finally, to define a labelling function, notice that a social ranking does not consider gradings. However, we can consider one label to indicate that an element aligns with criterion c_S (the element appears in set S), and another label to indicate that the element is neutral to this criterion (the element does not appear in S). For that, we define labels l_1 and l_0 respectively (along with the unused l_{-1} to make the label system symmetric). Hence, we define the label system $LS = \langle L, \geq_L \rangle$, with $L = \{l_1, l_0, l_{-1}\}$, and order $l_1 \geq_L l_0 \geq_L l_{-1}$. Then, we build a labelling λ that specifies whether an element x is related to c_S with label l_1 if $x \in S$, or with label l_0 if $x \notin S$:

$$\lambda(x, c_S) = \begin{cases} l_1, & \text{if } x \in S \\ l_0, & \text{if } x \notin S \end{cases} \tag{5.8}$$

The t function allows to transform any social ranking input into an MC ranking input. In fact, in what follows we prove that MC rankings generalise social rankings. Before that, we need an auxiliary result regarding the properties of function t as shown by the following lemma.

Lemma 8. *The t function is injective, but not exhaustive.*

Proof (Lemma 8)**.** *Suppose that t is not injective. Thus, for a given power set $\mathcal{P}(X)$, there are two different rankings $\succeq, \succeq' \in \mathcal{R}(\mathcal{P}(X))$, such that $t(\succeq) = t(\succeq')$. Since $\succeq, \succeq'$ are different rankings, $\exists Y, Z \in \mathcal{P}(X)$, such that $Z \succeq Y$, while $Z \not\succeq' Y$. Note though that in these cases when applying t, we would have that $c_Z \succeq_C c_Y$ and $c_Z \not\succeq'_C c_Y$, which contradicts the assumption that $t(\succeq) = t(\succeq')$. Thus, t is injective. In terms of exhaustivity, t is not exhaustive because labellings using labels other than l_1 and l_0 can never be the image of a social ranking.*

Thanks to lemma 8, we prove our first general result.

Theorem 10. *MC rankings generalise social rankings. That is, given a set of elements X, a power set $\mathcal{P}(X)$, a ranking over the power set $\succeq_P$, and a social ranking $sr : \mathcal{R}(\mathcal{P}(X)) \to \mathcal{R}(X)$, there exists an MC ranking mcr, such that $sr(\succeq_P) = mcr(t(\succeq_P))$, but the reverse does not hold in general.*

Proof (Theorem 10)**.** *To prove the theorem we have to find a mcr function such that $sr(\succeq_P) = mcr(t(\succeq_P))$. Consider $mcr = sr \circ t^{-1}$. In this case, we would have that $mcr(t(\succeq_P)) = sr(t^{-1}(t(\succeq_P))) = sr(\succeq_P)$. In the previous lemma we have seen that t is injective but not exhaustive in general, meaning*

that in general it is not invertible. Note though that t is invertible when restricted to the domain $t(\mathcal{R}(X))$. In this case, since we start in this domain, t^{-1} exists, meaning that $mcr = sr \circ t^{-1}$ is a valid function, which proves the theorem.

This last theorem proves that all social rankings can be cast as an equivalent MC ranking. Also, since t is not exhaustive there are many MC rankings that cannot be cast as social rankings, meaning that the MC ranking is more general. From this last result, an interesting question we have to address is the relation between MC lex-cel and lex-cel (see Section 5.2). The next theorem shows that MC lex-cel generalises lex-cel.

Theorem 11. *MC lex-cel generalises lex-cel, that is, given a set X and a ranking $\succeq_P$ over $\mathcal{P}(X)$, $mclex(t(\succeq_P)) = lex(\succeq_P)$.*

Proof (Theorem 11)**.** *Suppose that $lex(\succeq_P) = \succeq_e$ and $mclex(t(\succeq_P)) = \succeq'_e$. We will see that given $x, y \in X$, $x \succeq_e y \Leftrightarrow x \succeq'_e y$. We start with $x \succeq_e y \Rightarrow x \succeq'_e y$. First, suppose that $x \succ_e y$ ($x \succeq_e y$ and $x \nsim_e y$). Then, from the definition of lex-cel, we would have that $\theta(x) >_{lex} \theta(y)$. Now, suppose that $\mathcal{P}(x)/\sim_P = \{\Sigma_1, \dots \Sigma_k\}$ with quotient order $\Sigma_1 \succ_P \cdots \succ_P \Sigma_k$. Then, $\theta(x) = (|S \in \Sigma_1 : x \in S|, \dots, |S \in \Sigma_k : x \in S|)$ and $\theta(y) = (|S \in \Sigma_1 : y \in S|, \dots, |S \in \Sigma_k : y \in S|)$. Hence, $\theta(x) >_{lex} \theta(y)$ means that $\exists\ \Sigma_i$ such that $|S \in \Sigma_i : x \in S| > |S \in \Sigma_i : y \in S|$, and $\forall\ \Sigma_j \succ_P \Sigma_i$, $|S \in \Sigma_i : x \in S| = |S \in \Sigma_i : y \in S|$. By applying t to $\succeq_P$, we obtain that $\succeq_C$, such that any sets $S, S' \in \mathcal{P}(X)$ are transformed into criteria $c_S, c_{S'} \in C$, and $\succeq_P$ is transformed into $\succeq_C$ following $S \succeq_P S' \Leftrightarrow c_S \succeq_C c_{S'}$. Hence, the image for t of each equivalence class $\Sigma_i \in \mathcal{P}(x)/\sim_P$ is a criterion equivalence class $\kappa_i \in C/\sim_C$, and the quotient order then satisfies that $\Sigma_i \succ_P \Sigma_j \Leftrightarrow \kappa_i \succ_C \kappa_j$. Recall that the labelling obtained by t is built following $\lambda(x, c_S) = l_1 \Leftrightarrow x \in S$. Thus, $|S \in \Sigma_i : x \in S| > |S \in \Sigma_i : y \in S|$ implies that $|c \in \kappa_i : \lambda(x, c) = l_1| > |c \in \kappa_i : \lambda(y, c) = l_1|$. Similarly, $\forall\ \Sigma_j \succ_P \Sigma_i$, $|S \in \Sigma_i : x \in S| = |S \in \Sigma_i : y \in S|$ implies that $\forall \kappa_j \succ_C \kappa_i$, $|c \in \kappa_j : \lambda(x, c) = l_1| = |c \in \kappa_j : \lambda(y, c) = l_1|$. Note that, in this case, since the label system only contains l_1, l_0, and l_{-1}, we have that $s_{max} = 1$, and hence $\forall \kappa \in C/\sim_C$, $\mu(x, \kappa) = (na(x, \kappa, 1))$. Moreover, from the definition of t, the labelling does not assign l_{-1}. Therefore, we have that $na(x, \kappa, 1) = |c \in \kappa : \lambda(x, c) = l_1|$, and overall $\mu(x, \kappa) = (|c \in \kappa : \lambda(x, c) = l_1|)$. Now, we have that $|c \in \kappa_i : \lambda(x, c) = l_1| > |c \in \kappa_i : \lambda(y, c) = l_1|$, implying that $\mu(x, \kappa_i) >_{lex} \mu(y, \kappa_i)$, and $\forall \kappa_j \succ_C \kappa_i$, $|c \in \kappa_j : \lambda(x, c) = l_1| = |c \in \kappa_j : \lambda(y, c) = l_1|$, which implies that $\mu(x, \kappa_j) = \mu(y, \kappa_j)$. This is precisely the*

definition of $\mu(x) \succ_{lex} \mu(y)$*, which means that* $x \succ'_e y$*. Similarly, if* $x \sim_e y$*,* $\theta(x) = \theta(y)$*. Therefore,* $\forall i$*,* $|S \in \Sigma_i : x \in S| = |S \in \Sigma_i : x \in S|$*. If we apply* t*, this means that* $\forall i$*,* $|c \in \kappa_i : \lambda(x, c) = l_1| = |c \in \kappa_i : \lambda(y, c) = l_1|$*, and then* $\mu(x) = \mu(y)$*, hence following that* $x \sim'_e y$*.*

When it comes to the reverse implication, $x \succeq'_e y \Rightarrow x \succeq_e y$*, suppose that* $x \succ'_e y$*. In this case,* $x \preceq_e y$ *cannot happen because we have seen above that it would imply that* $x \preceq'_e y$*, which is not true. Then, the only possibility is that* $x \succ_e y$*. We can follow the same reasoning to prove that* $x \prec'_e y \Rightarrow x \prec_e y$ *and* $x \sim'_e y \Rightarrow x \sim_e y$*.*

5.7 Case study: value-aligned norm selection

The purpose of this section is to illustrate how MC-lexcel can be used to perform value-aligned norm selection. Next, Section 5.7.1 introduces our decision making problem. Thereafter, Section 5.7.2 discusses how to exploit MC-lexcel to computationally solve the decision problem. Finally, in Section 5.7.3 we discuss a case study in a healthcare context, concerned with selecting norms related to hospital admission. Furthermore, we compare the qualitative solving method detailed in Section 5.7.2 with previous methods.

5.7.1 The generalised value-aligned norm selection problem

Next we introduce our new formalisation of the value-aligned norm selection problem considering graded promotion and demotion relationships between norms and values, which we call generalised value-aligned norm selection problem. Thus, we first introduce the formal objects required for the problem, namely: norms, value system, and relationships between norms and values.

We define the core notion of our problem, the norm, as a simplification of the one in [López y López et al., 2002]. We start by considering a MAS (multi-agent system) with a set of agents Ag that can perform actions in a finite set $\mathcal{A}$. Furthermore, we consider a propositional language $\mathcal{PL}$ (with propositions in $\mathcal{P}$ and the logical operator "and"), a set of states S, and a state transition function that changes the state of the world when agents perform actions (following the multi-agent system model introduced in [Morales et al., 2015a; Morales et al., 2015b]). Then, a norm is composed of a precondition $\varphi \subseteq \mathcal{P}$ (with an "and" semantic between propositions), an

action in $\mathcal{A}$, and a deontic operator θ to establish Obligations (Obl), Permissions (Per), and Prohibitions (Prh). With these definitions in place, we define a norm as:

Definition 40 (Norm)**.** *A norm is a pair $\langle \varphi, \theta(a) \rangle$, where φ is a precondition in the language $\mathcal{PL}$; $a \in \mathcal{A}$ is the regulated action, and $\theta \in \{Obl, Per, Prh\}$ is a deontic operator.*

Example 28. *Within a healthcare context, we may have a norm permitting hospital admission of incoming patients: $\langle$patient_in, Per(admit)$\rangle$.*

Let N be a set of candidate norms, norms in N might have relationships between themselves [Serramia et al., 2018b]. We consider two types of such norm relations, namely norm exclusivity and norm generalisation and note them as R_x, and R_g respectively (we assume the decision maker has enough domain knowledge to detect and provide these norm relations). On the one hand, we say n, n' are *exclusive norms*, noted as $(n, n') \in R_x$, when we cannot enact both of them at once. On the other hand, we say they have a direct generalisation relation, noted $(n, n') \in R_g$, meaning n is more general than n'. With regards to generalisation relations, we note by $S(n)$ and $A(n)$, the successors and ancestors of n respectively. Formally:

Definition 41. *Given a norm $n \in N$, its ancestors are the norms that (directly or indirectly) generalise it: $A(n) = \{n' \in N : \exists n_1, \ldots, n_k$, and $(n', n_1), \ldots, (n_k, n) \in R_g\}$. Conversely, successors are the norms that are (directly or indirectly) generalised by n: $S(n) = \{n' \in N : \exists n_1, \ldots, n_k$, and $(n, n_1), \ldots, (n_k, n') \in R_g\}$.*

Norms and their relations form a structure called norm net.

Definition 42. *Let N be a set of norms and $R = \{R_x, R_g\}$ the set of norm relations (exclusivity and generalisation), we call* norm net *the tuple $\langle N, R \rangle$.*

Definition 43. *We call* norm system *to any subset $\Omega \subseteq N$.*

Not all norm systems are of our interest, note that norm systems may have conflicts (if they contain exclusive norms) or redundancy (if they contain norms related through generalisation). Thus, we focus on *sound* norm systems, i.e. those that are conflict-free and non-redundant [Serramia et al., 2018b].

Definition 44. *Let $\langle N, R \rangle$ be a norm net, then we sat a norm system $\Omega \subseteq N$ is sound iff it is:*

- *Conflict-free:* $\forall n_i, n_j \in \Omega$, $(n_i, n_j) \notin R_x$
- *Non-redundant:* $\forall n, n' \in \Omega$, $n \notin A(n')$; *and* $\forall n$, *with* $|\bar{S}(n)| > 1$, *then* $\bar{S}(n) \not\subseteq \Omega$.

Where $\bar{S}(n) = \{n' \in N, (n, n') \in R_g\}$ *stands for the set of direct successors.*

As decision criteria, we consider the *value system*, a structure formed by moral values and their preferences. Therefore, we define the value system as follows.

Definition 45. *Let* V *be a non-empty set of moral values, and* $\succeq_v$ *a ranking over* V*, we call* value system *the tuple* $\langle V, \succeq_v \rangle$.

Now we are ready to formalise how norms and values relate. For that, we can leverage on the notion of label system $\langle L, >_l, \lambda \rangle$, introduced by Definition 31, with each label corresponding to either a certain degree of promotion or demotion, and with function $\lambda : N \times V \to L$ assigning a label to each norm-value pair. We impose a neutral label l_0 in L to express that a norm and a value are unrelated. This label also sets the boundary between promoting and demoting labels: labels more preferred than l_0 are promoting labels, while those less preferred than l_0 are demoting labels. Notice that the sign function in equation 5.2 signals if a label represents promotion (1), demotion (-1), or if it is neutral (0). Moreover, the strength function in equation 5.3 characterises the degree of promotion/demotion of labels. Thus, given a label $l \in L$, the more labels between l and l_0, the larger its promotion/demotion degree (i.e., the stronger l is).

Thanks to the objects formally introduced so far, we are ready to introduce our decision-making problem, the so-called *generalised value-aligned norm selection problem* (GVANS)[4]. The input of the GVANS problem is: (i) a norm net $\langle N, R \rangle$; (ii) a value system $\langle V, \succeq_v \rangle$; and (iii) a symmetric label system $\langle L, >_l, \lambda \rangle$ that sets the relation between norms and values. Solving a GVANS problem consists in composing the sound norm system which best aligns with the value system, taking into account the degree of promotion/demotion of norm-value relations as expressed by the label system.

[4]The GVANS problem is a generalisation of the VANS problem introduced in Chapter 4, which disregarded demotions and promotion relations with different degrees.

5.7.2 Solving the GVANS

When deciding on the most value-aligned norm system, we follow the proposition: the more preferred the values promoted by a norm system, the more preferred the norm system, or, in other words, the more *value-aligned*. To obtain the most value-aligned norm system (i.e., to solve a GVANS problem) we will proceed in two steps.

First, we can exploit MC-lexcel to obtain a ranking over individual norms from a ranking over values in a value system. This is straightforward if we consider that the values in V act like criteria (i.e. $C = V$), and value preferences are cast over the elements of the decision (i.e. the norms in N). Importantly, our aim is to use the norm ranking to later select the set of norms that best aligns with the value system. Since norms can both promote and demote values, there might be norms that overall demote more preferred values than those that they promote. We call these norms *non-beneficial* norms. In contrast, *beneficial* norms are those that promote more preferred values than those that they demote. A simple informal way to differentiate beneficial and non-beneficial norms is to compare them with respect to a neutral norm n_0. We define n_0 as an artificial norm that is neutral with regards to all moral values in the value system. Thus, informally:

Definition 46. *A beneficial norm is a norm that is more preferred than n_0. Norms less preferred than or indifferently preferred to n_0 are non-beneficial norms. We note by $N_{ben} \subseteq N$ the subset of beneficial norms in N.*

When selecting a set of norms, we want to select only beneficial norms and avoid non-beneficial norms. In other words, the solution to the GVANS problem is a set of norms in N_{ben}. With MC-lexcel we can obtain a ranking that allows us to compare norms, but we must also know which norms are beneficial and which are not. In line with previous Definition 46, we exploit MC profiles to differentiate them. Thus, in the case of MC profiles:

Definition 47. *We say that a norm $n \in N$ is beneficial if $\mu(n) >_{lex} \mu(n_0)$. On the other hand, a norm is non-beneficial if $\mu(n_0) \geq_{lex} \mu(n)$. Thus, in this case, $N_{ben} = \{n \in N : \mu(n) >_{lex} \mu(n_0)\}$.*

Indeed, since we build the ranking from the MC profiles of norms, a norm that is less preferred than n_0 will be a norm whose MC-profile is worse than that of a totally neutral norm. This is the case when an MC profile contains more demotion labels than promotion labels, or contains demotion labels associated to more preferred values. Thus, by applying MC-lexcel

considering $N \cup \{n_0\}$, we obtain a ranking $mclex(\succeq_v) = \succeq_n$ in which not only we can compare norms, but also in which n_0 partitions norms between beneficial (when $n \succ_n n_0$) and non-beneficial (when $n_0 \succeq_n n$) norms.

The next step is to use the norm ranking to compose the desired set of value-aligned norms. Since only beneficial norms should be taken into account to compose the norm set, we discard the non-beneficial norms hereafter. Hence, we now consider the ranking only over beneficial norms $\succeq_n^{ben}$ obtained from the MC ranking over all norms. We formalise this using the following restriction:

Definition 48. *The restriction function ben is a function* $ben : \mathcal{R}(N) \rightarrow \mathcal{R}(N_{ben})$*, such that* $\forall \succeq_n \in \mathcal{R}(N)$ *and* $\forall n_1, n_2 \in N_{ben}$*,* $ben(\succeq_n) = \succeq_n^{ben}$ *is such that* $\succeq_n^{ben} = \{(n_1, n_2) \in \succeq_n : n_1, n_2 \in N_{ben}\}$*.*

Our final step consists in transforming the ranking over beneficial norms into a ranking over norm systems. For that, we resort to the anti-lex-cel operator introduced in Chapter 4. Let N_{ben} be a set of beneficial norms, and $\succeq_n^{ben}$ a ranking over these norms, the anti-lex-cel function $ale : \mathcal{R}(N_{ben}) \rightarrow \mathcal{R}(\mathcal{P}(N_{ben}))$ is a lifting function which generates a ranking over subsets of beneficial norms, namely over the norm systems in $\mathcal{P}(N_{ben})$. Therefore, the composition of MC-lexcel, the restriction to beneficial norms, and anti-lex-cel, transforms preferences over values in a value system to preferences over beneficial norm systems. We formally define this composition as follows:

Definition 49. *We call* $nsr : \mathcal{R}(V) \rightarrow \mathcal{R}(\mathcal{P}(N_{ben}))$ *(nsr for norm system ranking) the function* $nsr = ale \circ ben \circ mclex$*. Thus, for a value ranking* $\succeq_v \in \mathcal{R}(V)$*,* $nsr(\succeq_v) = ale(ben(mclex(\succeq_v))) = \succeq$ *is a ranking over norm systems (introduced in Definition 43) composed of beneficial norms.*

The solution to the GVANS problem at hand will be the most preferred sound norm system in the obtained norm system ranking. Unfortunately, although we have managed to formally solve our problem, the cost of building a whole ranking over norm systems (elements in $\mathcal{P}(N_{ben})$) turns out to be rather costly. As discussed in Chapter 4 , building the ranking using anti-lex-cel takes $O(2^{2^{|N|}})$ in the worst case (and when all norms are beneficial). Nonetheless, in Chapter 4 we show that it is possible to avoid to explicitly build a whole ranking over norm systems by encoding it as a BIP (Binary Integer Program). We can employ the very same approach here: firstly, we apply MC-lexcel to obtain a norm ranking; secondly we restrict the norm ranking to beneficial norms; then use this beneficial norm ranking to encode

the GVANS problem as a BIP; and finally, we solve it with the aid of standard BIP solvers (e.g. CPLEX [IBM, 1988] or Gurobi[Gurobi Optimization, 2010]). We henceforth refer to this method as the **qualitative approach with graded value promotion and demotion**. Next section illustrates and compares it to previous approaches.

5.7.3 Comparing solving methods

Following Example 28 on healthcare, here we introduce a simple example that illustrates the **qualitative approach with graded value promotion and demotion** described in Section 5.7.2. Furthermore, we use it to compare the approach of this chapter to those of previous chapters. On the one hand, Chapter 3 proposes a numerical approach that first assigns a numerical value alignment to each norm, and then selects norms by maximising their cumulative value alignment. Here, we show that considering these numerical evaluations of norm value alignment may introduce biases that the qualitative approach of this chapter avoids. On the other hand, although the work presented in Chapter 4 is also qualitative, it has limited expressiveness, since it does not allow for demotion, nor for different degrees of promotion/demotion. Overall, we show that, for specific cases, these other methods in the literature fail to produce a norm system that is most aligned with the given moral values.

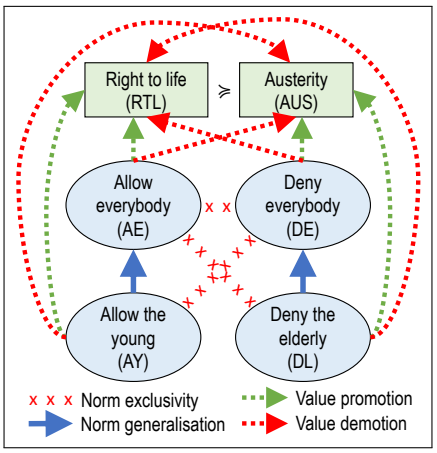


Figure 5.1: Representation of the norms, norm relations, values, and value promotion/demotion in our healthcare case study.

As previously mentioned, our case study focuses on selecting norms related to hospital admission. In particular, as Figure 5.1 shows, we consider

four norms:

- AE = ⟨patient_in, Per(admit)⟩: Allow admission to Everybody;
- DE = ⟨patient_in, Prh(admit)⟩: Deny admission to Everybody;
- AY = ⟨young_patient_in, Per(admit)⟩: Allow admission to the Young;
- DL = ⟨elder_patient_in, Prh(admit)⟩: Deny admission to the eLderly.

Thus, N={AE, DE, AY, DL}. Furthermore, as shown in Figure 5.1, since admission cannot be allowed and denied simultaneously, some of these norms are exclusive: (AE, DE)$\in R_x$, (AY, DE)$\in R_x$, (AE, DL)$\in R_x$. Moreover, regulating the admission to everybody includes the young and the elders, and hence, AE generalises AY and DE generalises DL.

As for values, in this setting, we consider two moral values V ={RTL, AUS}: "Right To Life/medical care" (RTL) and "Austerity" (AUS), and a preference of RTL $\succeq_v$ AUS. Figure 5.1 depicts that norms allowing admission promote RTL and demote AUS whereas the norms denying it behave conversely. However, to express promotion/demotion degrees we consider a label system $\langle L, >_l, \lambda\rangle$ with labels: high promotion (HP), promotion (P), neutral (l_0), demotion (D) and high demotion (HD) ($L = \{HP, P, l_0, D, HD\}$) and linear order $HP >_l P >_l l_0 >_l D >_l HD$. Note that, e.g., $stg(HP) = 2$ and $sgn(HP) = 1$, whereas $stg(HD) = 2$, and $sgn(HD) = -1$. The λ columns in Table 5.1 detail the λ function that completes our label system. Overall, general norms that apply to everybody are strongly related to the values. This is also the case for elders, since they are most likely to require admission. Alternatively, we consider the relationship with youngsters to be less strong, since they are less likely to require admission.

From here, we apply our *qualitative approach with graded value promotion and demotion* to compute the norm ranking as AE $\succeq_n$ AY $\succeq_n n_0 \succeq_n$ DL $\succeq_n$ DE. This is because $\mu(AE) = (HP, HD)$, $\mu(AY) = (P, D)$, $\mu(DL) = (HD, HP)$, and $\mu(DE) = (HD, HP)$. Note that DE and DL are non-beneficial norms because they are less preferred than n_0 (due to their demotion of the most preferred value). Therefore, by restricting the ranking to beneficial norms, we have that AE $\succeq_n^{ben}$ AY. Next, we obtain the norm system ranking {AE, AY} $\succeq$ {AE} $\succeq$ {AY}. However, {AE, AY} is not sound (see Definition 44) because AE generalises AY and, hence, the method will choose {AE} as the most value-aligned norm system to be enacted. Indeed, considering that AE is the most general norm with the highest promotion of RTL –the most preferred value–, {AE} stands for the expected solution.

	AE			DE			AY			DL		
	λ	Num	B	λ	Num	B	λ	Num	B	λ	Num	B
RTL	HP	1	1	HD	-1	0	P	0.7	1	HD	-0.8	0
AUS	HD	-1	0	HP	1	1	D	-0.2	0	HP	0.8	1

Table 5.1: Value-norm relationships for the three methods: λ (the method used in this chapter); Num (Numerical approach in Chapter 3); and B (Binary approach in Chapter 4).

Alternatively, when considering the **quantitative approach** in Chapter 3, the task of assigning numerical norm promotions turns out to be more difficult. Num columns in Table 5.1 detail the numeric norm promotions we use in this comparison. Extreme grades now become 1 and -1 respectively. AY promotes RTL with 0.7 and demotes AUS with -0.2 because the young are just a portion of the incomers and the cost of their medical care is (relatively) low. DL demotes RTL with -0.8 and promotes AUS with 0.8 since most people at risk of dying are elders and they usually require most expensive medical care.

Subsequently, the procedure in Chapter 3 computes the value alignment of the available sound norm systems –note in this case the values' relevance are $r(RTL) = 2$ and $r(AUS) = 1$– as: $va(\{\text{AE}\}) = 1$, $va(\{\text{DE}\}) = -1$, $va(\{\text{AY}\}) = 1.2$, $va(\{\text{DL}\}) = -0,8$, and $va(\{\text{AY,DL}\}) = 0,4$. Hence, the sound norm system with highest utility is {AY}. This means that the quantitative utilitarian method selects a norm system that fails to regulate admissions of elder people. This is so because AE strongly demotes the AUS value, and this diminishes its numerical value alignment.

If we now consider the **binary qualitative approach** from Chapter 4, B columns in Table 5.1 are limited to represent promotion (1) and no promotion (0). This method produces a norm ranking of AE $\sim_n$ AY $\succeq_n$ DE $\sim_n$ DL. Notice that AE $\sim_n$AY because both norms promote RTL and the method is not expressive enough to capture different grades of promotion, even though admitting everybody (AE) is clearly a better norm (i.e., it is far more inclusive) than just admitting the young (AY). Consequently, this norm ranking leads to the following ranking of sound norm systems: {AY, DL} $\succeq$ {AE} $\sim$ {AY} $\succeq$ {DE} $\sim$ {DL}, where {AY, DL} supports both RTL and AUS values. Hence, the binary qualitative method selects to enact {AY, DL}, which fails to be aligned with the value system because denying admission to elders (DL) demotes the most preferred value of right to life (RTL). The reason for considering such an undesirable norm is a direct

consequence of its failure to capture demotion. In fact, it only selects norms based on their merits without considering their detrimental effects.

In conclusion, albeit its simplicity, this example illustrates how the method of this chapter overcomes the shortages of the methods of Chapter 3 and Chapter 4 in producing a norm system that is most aligned with the value system at hand. In fact, [Serramia et al., 2020] already reported a flaw in [Serramia et al., 2018b] that causes that a number of norms slightly promoting least preferred values can end up having more utility –and thus being chosen– than a single really useful norm if they are exclusive. Indeed, although most preferred values should prevail, the quantitative method also fails to capture the absolute preferences of the value system.

The advantages of the method in this chapter are two-fold. First, its graded qualitative labels for promotion and demotion are much simpler to define –and less prone to biases– than numerical degrees and also provide far more expressiveness than just binary promotion. Second, its ranking method captures the preferences of the value system into the selection of the norm system to enact satisfying the dominance property, which means that the resulting ranking is excellence-rewarding.

5.8 Conclusions

In this chapter we have tried to make headway in supporting decision makers that are challenged with comparing, and ultimately ranking, elements with regards to how such elements satisfy multiple criteria and how such criteria are preferred by them. This calls for a new decision making framework, which we have formally introduced here. Our framework is based on a novel method for ranking single elements.

Ranking functions have been widely used to transform rankings. For instance, the social ranking function transforms a ranking over sets of elements into a ranking over the elements themselves. The contributions of this chapter advance the state of the art with a novel family of ranking functions –multi-criteria (MC) rankings– and a function of this family –MC lex-cel– so to transform complex preference (criteria) information into a neat and clear ranking of individual elements. Furthermore, we have positioned our findings with respect to the current literature, by showing that our MC ranking generalises the social ranking and MC lex-cel generalises the lex-cel social ranking function and embodies dominance.

Importantly, MC rankings can be used to solve DSSP with graded rela-

tions between elements and criteria. Indeed, by composing an MC ranking with a lifting function, we can transform preferences over criteria to preferences over sets of elements. Once this ranking is obtained, the DSSP can be solved following the same procedure explained in Chapter 4. We have pictured this application of MC rankings to support a decision maker to tackle an ethical decision making problem. Specifically, we define the Generalised Value-Aligned Norm Selection (GVANS) problem and solve it with a qualitative approach with graded value promotion and demotion. Overall, this method overcomes the shortages of previous methods in producing a norm system that is most aligned with the value system at hand.

Hence this chapter has addressed the remaining part of contributions C8 - C11 from Section 1.3, which in turn answer the corresponding research questions in Section 1.2:

- Question Q8: How can we represent qualitatively the relations between norms and values? Through graded positive/negative relations which we capture through labels (remaining part of contribution C8).

- Question Q9: How do we solve the value-aligned norm selection problem qualitatively? We obtain a ranking over norm systems by composing MC lex-cel and anti-lex-cel to then select the most preferred sound norm system in the ranking (remaining part of contribution C9).

- Question Q10: Are qualitative approaches computationally feasible? Yes, once we know the ranking over norms, we can apply the BIP encoding of Chapter 4 which allows to solve the problem avoiding the computational cost of building the norm system ranking (C10).

- Question Q11: How can we generalise the qualitative approach to value-aligned norm selection to use it in other multi-criteria decision making problems? In this chapter we have re-formalised the dominance property to encompass labelled relations between elements and criteria. Thus, considering this new formalisation, the dominant set selection problem definition of Chapter 4 remains (C11).

Chapter 6

Conclusions and future work

To finish the thesis, we outline some conclusions and future work. Firstly, Section 6.1 draws conclusions with regards to the various contributions of the thesis. Then, section 6.2 discusses the lessons learned along the work of this thesis. Finally, Section 6.3 provides paths for future research.

6.1 Conclusions

With the advent and adoption of intelligent systems, the ethical implications of their actions have become increasingly concerning. The value alignment problem [Russell, 2019] addresses this concern. In this thesis we have formalised a particular instance of the value alignment problem, namely the value-aligned norm selection problem (VANS). Moreover, we provide different methods to compose norm systems whose norms align with the value system of the society. In that regard we provide the following conclusions, following the two main blocks of contributions presented in Section 1.3.

6.1.1 Formalisation of the value-aligned norm selection problem

With the increased awareness of ethical issues in the AI community, moral values have started to be studied within the AI literature. In particular, the introduction of value systems [Bench-Capon and Atkinson, 2009; Luo et al., 2017; Serramia et al., 2018a] have allowed to implement moral value criteria in decision making. Nonetheless, previous definitions of the value system lacked the nuances discussed in the ethics literature, as well as a clear

reasoning on their definition. This is particularly necessary for value-aligned decision making problems dealing with actions or norms, since the relation between these two elements and values has been thoroughly studied. Thus, we propose a formalisation of value system that considers such relations.

Figure 6.1 provides a graphical summary of the study on norms, values, and norm value alignment of Section 3.2. On the one hand, moral values judge how good or bad are actions to perform or to not perform. Exploiting this judgement of action performance and non-performance we have formalised moral values by providing them with such semantic through the so-called judgement functions. This has allowed us to better formalise the notion of value system as a set of moral values (composed of their action judgement function), as well as preferences over these values in the form of a ranking (which we have deemed the most appropriate way to represent value preferences).

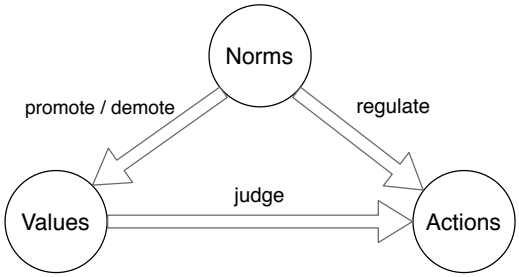


Figure 6.1: Relation between actions, norms and moral values.

Furthermore, since values judge actions that are regulated by norms, we can also derive the promotion/demotion relation between values and norms. Thus, we have formalised promotion functions relating a norm to a value taking into account how the value judges the action regulated by the norm.

Besides being useful for tackling value-aligned norm selection, it is important to remark that our formalisation of value system can be useful for other applications dealing with actions, norms and values, which can benefit from the relations in Figure 6.1.

Focusing on the value-aligned norm selection problem, the aim is to select the set of norms that best aligns with a value system. Note though, that norms can be interrelated: they can be mutually exclusive (e.g., a norm prohibiting an action is exclusive with a norm obliging it), or they can be redundant (e.g., a norm may generalise another one by having a broader scope). Thus, not all possible norm systems constitute a feasible solution as

we want to avoid exclusive and redundant norms inside a norm system. We say norm systems are sound when they are free of exclusive and redundant norms. In this manner, the value-aligned norm selection problem can be informally defined as the problem of finding the sound set of norms that best aligns with the value system. Its formal definition though is not as simple as it depends on how we represent relations between norms and values.

On the one hand, in Chapter 3 we consider that the decision maker can numerically asses the relation between a norm and a value. In other words, the decision maker is able to provide the judgement functions for the values. Thus, we formalise the VANS problem using our definition of promotion function (which depends on the value's judgement functions). Hence, the VANS problem is formalised as the problem of selecting the sound set of norms that maximise value alignment utility, which depends on how norms promote values, and the preferences over these values.

On the other hand, in Chapters 4 and 5 we have not defined the VANS problem directly, but through a more general problem: the dominant set selection problem. Informally, this problem consists on selecting a feasible set of elements that best aligns with a set of criteria considering how elements relate to criteria and the preferences over these criteria. Unlike in Chapter 3, in these chapters we assume that the decision maker is not able to numerically asses how elements (in the case of VANS, norms) relate to criteria (in the case of VANS, values). Chapter 4 supposes a set of elements that relate to a set of criteria through a binary relation, hence an element is either aligned or not with each criterion. On top of that, we consider a function that allows us to check if a solution is feasible. With that in mind, we formalised element dominance, and subsequently, set dominance. These properties have allowed us to define the dominant set selection problem (DSSP). The DSSP aims at finding the feasible set of elements that is not dominated by any other set (in other words, the set of elements that dominates or is indifferent to all other feasible sets). In this case, the VANS problem is a DSSP whose elements are norms, and whose criteria are values, and the feasibility of its solution depends on the soundness of norm systems.

Chapter 5 also provides a formalisation of the DSSP with a more expressive relation between elements and criteria. There, we capture the relations through a label system, these labels allow the decision maker to express different degrees of positive and negative relations between elements and criteria. The number of labels is flexible, and therefore, the decision maker can define labels depending on the granularity of their knowledge. This new labelled relations require to re-formalise the property of element dominance.

Nonetheless, the DSSP formalisation remains the same once considered such re-formalisation.

6.1.2 Solving the value-aligned norm selection problem

When it comes to solving the VANS problem, we have proposed several approaches. In Chapter 3 we have introduced a quantitative approach to solve the VANS problem. Conversely, Chapters 4 and 5 introduce qualitative approaches with different degrees of expressiveness. Here we detail the overall conclusions of each of these approaches.

Quantitative reasoning

The quantitative approach of Chapter 3 relies on value alignment utility functions based on the norm promotion function (which in turn is based on the action judgement functions). These functions allow for the problem to be encoded into a binary integer program. We show that this approach allows to solve even large VANS problems in affordable computational times, which vary depending on the structure of the problem.

Qualitative reasoning

Chapters 4 and 5 provide a solution to the general DSSP, their resolutions can be applied to the VANS problem because the DSSP is a generalisation of the VANS problem. To solve the DSSP, the approaches of both chapters aim at transforming the criteria preferences to element preferences, and in turn these element preferences to set preferences. In practice, this is done through the composition of a preference grounding function and a lifting function.

In Chapter 4, we adapted lex-cel to be used as our grounding function and used anti-lex-cel as our lifting function. By composing these two functions, we are able to transform the ranking over criteria to a ranking over sets of elements. In more detail, first lex-cel transforms the ranking over criteria to a ranking over elements considering how elements relate to the criteria, then anti-lex-cel lifts the ranking over elements to a ranking over sets of these elements. We proved that the most preferred feasible set in the obtained ranking is a solution to the DSSP problem. This resolution serves as a first approach to the composition of ranking functions to solve

such problems from a qualitative perspective. Unfortunately, building this ranking can be computationally costly, nonetheless we are able to encode this problem into a binary integer program (BIP). This allows us to avoid the computational cost of building the preferences over all sets. Importantly, we prove that the BIP produces an equivalent solution to the one we would obtain by building the preferences and selecting the most preferred feasible set. As argued before, we show how this general framework can be used for the particular problem of value-aligned norm selection. Nonetheless, it is important to remark that besides value-aligned norm selection, we argue that the dominant set selection problem characterises a family of problems with similar characteristics (the goal is to select a set of elements, the decisionmaker counts on some criteria and preferences over these criteria, and we have relations between the elements and the criteria). For example, problems such as grant allocation or personnel selection can be seen as specific instances of a DSSP. The resolution we provide in Chapter 4 can be applied to any problem that can be formulated as a DSSP.

Due to the simple binary element-criterion relation, in Chapter 4 we ground criteria preferences using ranking functions already available in the literature. Nonetheless, in Chapter 5 we consider relations between elements and criteria that are graded. Unfortunately, the literature has not discussed grounding functions that are able to cope with the above-mentioned relations between elements and criteria. Hence, in order to solve the DSSP in Chapter 5, we present a novel family of ranking functions called multi-criteria (MC) rankings. MC rankings are functions that transform preferences over criteria to preferences over elements considering the labelled relations between elements and criteria. We formalise a particular MC ranking, called MC lex-cel. MC lex-cel is a generalisation of lex-cel. We prove that MC lex-cel satisfies element dominance. Thus, we compose MC lex-cel with anti-lex-cel to solve the DSSP with labelled element-criterion relations. In particular, we have shown how to apply the composition of MC lex-cel and anti-lex-cel to solve the VANS problem with labelled relations between norms and values. Nonetheless, as previously mentioned, the framework and resolution provided in Chapter 5 can be useful for a wide range of problems.

6.2 Lessons learned

As previously discussed, Chapters 3, 4 and 5 propose different approaches to value-aligned norm selection, namely a quantitative, and two qualitative

approaches. A difference between these approaches is the different levels of knowledge required from the decision maker to define their input. Recall that, while our quantitative approach requires the decision maker to provide numerical assessments of how actions relate to values, our qualitative approaches do not require such detailed information. Nonetheless, we have also argued that our quantitative approach prefers quantity over quality of norms: the larger the norm system, the better. Indeed, as explained in Section 3.10, the quantitative approach might favour selecting a set of mediocre (weakly-aligned) norms instead of a single excellent (strongly-aligned) norm. We argued that this is due to utility additivity, the sum of lots of small utilities adds up and surpasses a single very high utility. We argued that when selecting value-aligned norms, our preference must be to select the best norms, no matter if this means selecting a smaller set of norms. Thus, we should favour quality, in terms of value alignment, over quantity. This is addressed through our qualitative approaches, which satisfy the dominance property. Nonetheless, note that this distinction between our quantitative and qualitative approaches lays on a deeper difference between them. On the one hand, the quantitative approach assigns utilities to norms individually. Thus, the utility of a norm is independent of the utilities of other norms. On the other hand, our qualitative approach builds preferences over norms. Thus, it evaluates the value-alignment of each norm not independently, but in comparison with the other candidate norms. The selected set of norms is derived from the individual utilities and from the preferences over norms respectively. This causes that our qualitative approaches satisfy both the dominance property while the quantitative approach is affected by the additivity of utility. In the quantitative approach, since all norm utilities are independent, we cannot demand any sense of dominance (in other words, that an excellent norm always has better utility than any sum of more mediocre norms). In the qualitative approaches though, since the preferences over sets of norms are built from the comparison of the individual norms, we can demand that the resulting ranking satisfies dominance.

In conclusion and in the larger picture, we have looked into two groups of approaches to decision making, evaluative approaches (those that evaluate elements individually), and comparative approaches (those that evaluate elements in comparison). Both of these approaches can be useful to select a single element. Nonetheless, in problems requiring to select more than one element, evaluative approaches can have unforeseen consequences. When the set of selected elements is required to satisfy some property (like dominance), it is necessary to solve the problem through a comparative approach. In

other words, comparative approaches allow the decision maker to have better control over the type of solution obtained, whereas evaluative approaches do not allow for this degree of control.

6.3 Future work

While this thesis provides a thorough and detailed study of value-aligned norm selection, it is by no means a complete study on the subject. In fact, by studying this problem we have found many paths for future research in this and other topics. On the one hand, a possible path is to research the assumptions made throughout this thesis (known moral value preferences, norm-value relations). On the other hand, the methods presented in Chapters 4 and 5 can be more broadly studied from the field of computational social choice. We provide more detail on these future research paths in the following subsections.

6.3.1 Enriching the expressiveness of actions, norms and relations

Actions and norms

Enriching the action and norm language (/expressivity) would improve the scope of applicability of our work. The way we have formalised actions does not consider any particular type of structure or semantics for actions. On the one hand, while we have assumed atomic actions, in reality actions might have a more complex structure. For example, some actions might not be discrete, and hence, we cannot represent actions such as, e.g., "move x meters" where x is a real number. Recently, [Montes and Sierra, 2021] have studied value-guided synthesis of parametric norms. Thus, a possible way to include non-atomic actions could be the integration of their findings to our norm selection optimisation. On the other hand, the definitions of norm and normative framework have limited expressiveness. This hinders the application of our work to complex scenarios requiring norms that include, for instance, conditional and temporal aspects (such as in [Garcia-Camino et al., 2005]), or punishments [Bou et al., 2006].

Norm relations

We have to point out that the different approaches presented in this thesis have been explained using different formalisations of the normative domain or norm net. Our first approach, the quantitative approach in Chapter 3, tried to fully formalise them and define properties that they should maintain. Quickly, we detected that requiring these properties does not help to ensure that the norm relations are well-defined, but instead it constrains the types of relations that can be considered. In the following approaches the qualitative ones in Chapters 4 and 5, we were less restrictive in their definition. These differing definitions do not impact the correctness of the approach in terms of value-alignment, they may only affect the constraints in the binary integer programs. Nonetheless, further studying norm relations and unifying their definitions in the normative domain or norm net remains an important task to research.

6.3.2 Building the value system

This thesis assumes that the decision maker knows the society's preferences over values. Nonetheless, how to obtain these preferences remains an open question. As for future research we propose two different strategies for obtaining value preferences and therefore building the value system of the society.

Collective agreement

The first approach to build the value system is relying on the society to agree over the preferences over moral value. For that end, we already count on useful tools such as participatory platforms and online debates. Participatory platforms have recently become popular tools for governments to know the opinions of their citizens. Participatory portals are designed to enable informed and reasoned decisions, where citizens can share their opinions with their governments. Indeed, we can find several e-participation and e-governance ICT systems such as Loomio [Loomio, 2012], or Consider.it [Consider.it, 2010]. From these, we highlight Consul [Consul, 2015], which has been adopted by 100 institutions in 33 different countries and has been used by 90 million citizens. Additionally, some governments provide their own participation portals. For instance, since its launching, the UK's portal [Petitions, 2015] has received more than 20000 petitions, some of them being extremely popular (at the time of writing this article, the proposal "Revoke

Article 50 and remain in the EU" received more than 6 million signatures). France [Parlement & Citoyens, 2013] and New Zealand [New Zealand's Ministry of Justice, 2015] are also making an attempt to close the gap between their parliaments and their people. Furthermore, these attempts are done at a local level, with city councils, such as Reykjavik [City of Reykjavík, 2012] and Barcelona [Decidim, 2016] being committed to enable participation, giving the citizens the chance to present and debate their ideas. These portals could serve as the platform where citizens debate and agree on the value system of the society. As for the debates themselves, [Klein, 2012] introduces an argumentation structure that can be used for large-scale deliberation. We have also discussed online debates in [Serramia et al., 2019a]. There, users are able to post arguments in favour or against a statement, as well as, vote other peoples' arguments. Then, we propose functions to assess the acceptance or rejection of the statement based on the arguments of the debate and their votes. Either of these debate structures could be exploited to argue on the value preferences of the society.

Learning of moral values

While ideal for building the true preferences of the society, the collective agreement approach relies on the active participation of the society's individuals. This can be problematic for at least two reasons. Obviously, the individuals of the society may not care to make this effort. But also, it may be the case that only particular groups of individuals are willing to participate in the process, meaning that others may end underrepresented. Conversely, we can opt for an automated approach that does not need for the active participation of the individuals of the society. This could consist for example in applying inverse reinforcement learning techniques [Abbeel and Ng, 2004] for building individual's value preferences by observing their behaviour. Recently there has been some work in this direction, [Liscio et al., 2021b; Liscio et al., 2021a] identify context-specific values through the analysis of opinion corpora (with a hybrid approach of human annotators and NLP techniques). While this approach allows to cover and represent all individuals, it also has its shortcomings. A particular concern can be the difference between an individual's true value preferences and their portrayed value preferences. In other words, there might be a difference between the values somebody holds and those displayed in their actions. For example, a person that has a precarious job may not agree in low remuneration but has to take that job or otherwise they could not sustain themselves.

Aggregating value systems

The learning of moral values would allow to learn the value system of individuals, but we could obtain the value system representing the whole society through aggregation. Aggregating rankings is not a novel problem [Brandenburg et al., 2013] [Meena and Bharadwaj, 2020]. For example, when considering several individual rankings, Kemeny's rule [Kemeny, 1959] sees these rankings into a geometrical space and aims at finding the ranking minimising the overall distance between the original rankings and itself. This same idea is explained more formally in [Brandt et al., 2016] through the use of Kendall's tau distance [Kendall, 1938]. Nonetheless, note that the problem of aggregating value systems cannot be reduced to just aggregating their respective value rankings. Different people might have different conceptions of what a particular value means. Thus, these different conceptions should be taken into account when aggregating value systems.

6.3.3 Tools for decision makers

Our experimental framework provides the foundations to develop an interactive tool to enable policy/decision makers and the community to collectively specify the norms and values (the input) of the VANS problem. In this regard, we could build explanations for why an individual norm or a set of norms did not get selected. For example, for the approach of Chapter 3, these explanations could be based on the utilities of a norm system and could suggest changes in the settings (norm-value promotions or value preferences) for which the norm/norm system would have been selected. Finding these alternative settings could be addressed using optimisation techniques. This interactivity could also be extended throughout the VANS problem resolution, thus treating it not as a single shot process, but implementing a design methodology where policy/decision makers and the community interact along the decision making process to fine-tune the problem's specification.

6.3.4 Deepening on multi-criteria rankings and the composition of ranking functions

In Chapter 5 we have introduced a novel family of ranking functions called MC rankings. Furthermore, in that chapter we also formalised MC lex-cel as a function of this family. On the one hand, while our qualitative approach to solve the VANS problem relies on MC lex-cel, in reality, other MC

rankings can be used. This inevitably means that the selected set of norms would have different properties, since another MC ranking may not satisfy dominance, for example. Studying other MC rankings remains a future path of research. On the other hand, in Chapter 5 the family of MC rankings is only briefly studied as the main goal is to apply it to the problem at hand. Nonetheless, we think it may be interesting to further study MC rankings from the computational social choice perspective. In particular, note that as explained in Section 5.6, MC rankings generalise social rankings by considering labelled relations with elements. This means that the traditional properties of ranking functions may not be directly translatable to MC rankings and may not suffice to fully study them. Note that new properties on the labelled criteria-element relation have to be introduced and formalised.

While we have applied MC lex-cel to the value-aligned norm selection problem, this function can be further studied from the social choice perspective. In this sense, it would be interesting to provide an axiomatisation of MC lex-cel, like it has been done for many social ranking solutions [Bernardi et al., 2019; Haret et al., 2018; Khani et al., 2019]. As explained in the previous point, since MC rankings use labels, their properties will be greater in number and complexity with regards to the classic social choice properties. In this sense, this task seems quite daunting for two reasons. Firstly, because properties for MC rankings are not yet known. And secondly, because a full axiomatisation for MC lex-cel may require a greater number of axioms as for what is normal in known axiomatisations, which also conveys difficulty in proving that its axioms are independent.

Finally, and in a similar vein to previous points, the qualitative approach introduced in Chapters 4 and 5 relies on the composition of a social ranking function (lex-cel) or an MC ranking (MC lex-cel) with a lifting function (anti-lex-cel), which opens two questions. Firstly, and as discussed before, we have chosen specific functions to perform each of the steps of this composition, but others may be possible. While we have chosen the functions based on their properties which we have deemed appropriate, other functions will imply other properties, which is a question worth studying. But perhaps more importantly, we see the composition of ranking functions as an important topic worth researching. Note that, while the literature has thoroughly studied many ranking functions individually, the compositions of these functions have been disregarded, to the best of our knowledge. Thus, studying formally the properties of these compositions remains an open question. As we have remarked, we think these compositions can be useful to solve DSSPs and other decision making problems.

List of Notation and Symbols

This list provides a relation of the notation of each chapter and their meaning. Note that different chapters might use the same symbols to represent different concepts.

Chapter 3

Ag	A set of agents.
A	A set of actions the agents can perform.
$\mathcal{P}$	A set of propositions.
$\mathcal{L}$	A propositional language with propositions in $\mathcal{P}$ and the logical operator "and".
S	A set of states.
$\varphi/\varphi'/\varphi_i$	A context, in other words, a subset of the propositions of the language, hence $\varphi \subseteq \mathcal{P}$.
(φ, a)	A contextualised action, in other words, a pair of a context and an action.
$\mathbb{A}$	A set of contextualised action, hence $\mathbb{A} \subseteq \mathcal{P} \times A$.
$a/a'/a_i$	An action. This notation is used in two cases, to represent an action $a \in A$ or a contextualised action $a \in \mathbb{A}$. Most of Chapter 3, works on contextualised actions, therefore they are referred simply as actions.
R_i	The set of action incompatibility relations. These are binary relations between actions: $(a, a') \in R_i \subseteq \mathbb{A} \times \mathbb{A}$, in this case a and a' are incompatible actions.
R_g	The set of action generalisation relations. These are binary relations between actions: $(a, a') \in R_g \subseteq \mathbb{A} \times \mathbb{A}$, in this case a generalises a'.
$A(a)$	The ancestors of action a, namely the actions in $\mathbb{A}$ that are more general than a.

$S(a)$	The successors of action a, namely the actions in $\mathbb{A}$ that are more specific than a.
R	The set of action relations, containing action incompatibility and action generalisation relations $R = \{R_i, R_g\}$.
$\langle \mathbb{A}, R \rangle$	An action domain.
D	An action domain.
θ	A deontic operator, this can be $\theta \in \{Obl, Per, Prh\}$ an obligation, a permission or a prohibition respectively.
$n/n'/n_i$	A norm of the form $n = \theta(a)$. Hence, a norm is formed of a deontic operator and a (contextualised) action.
N	A set of norms.
sgn	The sign of a norm, see Equation 3.1.
$\mathfrak{R}_i$	The set of norm incompatibility relations. These are binary relations between norms: $(n, n') \in \mathfrak{R}_i \subseteq N \times N$, in this case n and n' are incompatible actions.
$\mathfrak{R}_g$	The set of norm generalisation relations. These are binary relations between norms: $(n, n') \in \mathfrak{R}_g \subseteq N \times N$, in this case n generalises n'.
$A(n)$	The ancestors of norm n, namely the norms in N that are more general than n.
$S(n)$	The successors of norm n, namely the norms in N that are more specific than n.
$\mathfrak{R}$	The set of norm relations, containing norm incompatibility and norm generalisation relations $\mathfrak{R} = \{\mathfrak{R}_i, \mathfrak{R}_g\}$.
$\langle D, N, \mathfrak{R} \rangle$	A normative domain.
Ω	A norm system, namely a subset of norms $\Omega \subseteq N$.
$v/v'/v_i$	A value.
V	A set of values.
α_v^+/α_v^-	The action judgement functions of value v $\alpha_v^+, \alpha_v^- : \mathbb{A} \to [-1, 1]$, α_v^+ judges the performance of actions, while α_v^- judges the non-performance of an action.
$\succeq$	A ranking.
$\sim$	The symmetric part of ranking $\succeq$, hence $x \sim y \Leftrightarrow x \succeq y, y \succeq x$.

$\langle V, \succeq \rangle$	A value system.
VS	A value system.
π	A norm promotion function $\pi : V \times N \rightarrow [-1, 1]$.
$\pi^{Obl}/\pi^{Per}/\pi^{Prh}$	The obligation, permission and prohibition cases (respectively) of a norm promotion function.
ϵ	A number in $[0, 1]$.
π_{base}	The base norm promotion function, see Equation 3.3.
π_{sup}	The supererogatory norm promotion function, see Equation 3.4.
r	The relevance function that transforms value preferences into a numeric value relevance $r : V \rightarrow \mathbb{R}$.
η	A value equivalence class $\eta \in V/\sim$.
va	The function assessing the value alignment of a norm system Ω, see Equation 3.6. A norm system can be composed of a single norm, in that case instead of noting $va(\{n\})$, we use the notation $va(n)$.
x_i	A binary decision variable $x_i \in \{0, 1\}$, this marks if norm n_i is selected ($x_i = 0$) or not ($x_i = 0$).

Chapter 4

General notation

X	A set of elements.
$x/x'/x_i$	An element $x/x'/x_i \in X$.
$\mathcal{P}(X)$	The power set of X.
$S/S'/S_i$	A subset of elements of X, $S/S'/S_i \in \mathcal{P}(X)$.
$\succeq$ / $\succeq'$ /$\succeq_X$	A generic ranking, where $\succ/\succ'/\succ_X$ is its antisymmetric part and $\sim/\sim'/\sim_X$ its symmetric part. We also use this notation for the ranking over $\mathcal{P}(X)$ obtained through anti-lex-cel.
F	A set of features.
$\mathcal{R}(X)$	All possible rankings over a set.
$\succeq_F$	A ranking of features.
$\mathfrak{f}$	A function that receives and element in X and returns the set of its features.
ϕ	The feasibility function that receives a set in $S \in P(X)$ and returns $\top$ if the set is feasible and $\bot$ if it is not feasible.
$\Psi/\Psi'/\Psi_i$	A feature equivalence class in $F/\sim_F$.
σ	A permutation of indexes of the elements in a set S with respect to a property (dominance). Thus, $\sigma(i)$ is the index of the i-th best element with regards to the property.
srs	A social ranking solution, that is a function $srs : \mathcal{R}(\mathcal{P}(X)) \to \mathcal{R}(X)$.
le	The lex-cel (lexicographic excellence) function.
$\succeq_e$	A general grounded ranking, that is a ranking over X, obtained from a social ranking solution (a grounding) srs (in particular a ranking obtained through lex-cel).
μ	The function that builds the profile vector of lex-cel.
$\Sigma/\Sigma'/\Sigma_i$	An equivalence class of $\mathcal{P}(X)/\sim$.
c_i^x	The i-th element of $\mu(x)$. Thus, $\mu(x) = (c_1^x, \dots c_k^x)$.
$\geq_L$	The lexicographic order of vectors.
ale	The anti-lex-cel function.
$\Xi/\Xi'/\Xi_i$	An element equivalence class of $X/\sim_e$.
η	The vector used by anti-lex-cel to build the ranking over $\mathcal{P}(X)$.
c_i^S	The i-th element of $\eta(S)$. Thus, $\eta(S) = (c_1^S, \dots c_q^S)$.

dom	A function that transforms a ranking over the features in F to a ranking over the sets in $\mathcal{P}(X)$. This function is the composition of lex-cel and anti-lex-cel $dom = ale \circ le$.
$\mathfrak{p}$	The preference function that assigns a natural number to each set in $\mathcal{P}(X)$, with regards to the ranking obtained trough dom.
S_{pref}	The feasible set in $\mathcal{P}(X)$ that is most preferred with regards to the set ranking $\succeq$ obtained through dom.
S_{max}	The feasible set of maximum preference $\mathfrak{p}$.
d_i	The decision variable representing element $x_i \in X$.

Value-aligned norm selection

A	A set of actions.
Ag	A set of agents.
$\mathcal{P}$	A set of propositions.
$\mathcal{L}$	A propositional language with propositions in $\mathcal{P}$ and the logical operator "and".
S	A set of states.
φ	A precondition of the form $\varphi \subseteq \mathcal{P}$ (with an "and" semantic between propositions).
θ	A deontic operator (prh/per/obl).
a	An action $a \in A$.
$n/n'/n_i$	A norm.
N	A set of norms.
$v/v'/v_i$	A moral value.
V	A set of moral values.
$\succeq_v$	A ranking over moral values in V.
R	The set of norms relations, $R = \{R_x, R_g\}$.
R_x	The set of exclusive relations, $(n, n') \in R_x$ if n and n' are mutually exclusive norms.
R_g	The set of generalisation relations, $(n, n') \in R_x$ if n generalises n'.
$A(n)$	The ancestors of n, the norms that generalise n.
$S(n)$	The successors of n, the norms that are generalised by n.

$\bar{S}(n)$	The direct successors of n, the norms n' such that $(n, n') \in R_g$.
d_i	The decision variable representing norm $n_i \in N$.

Chapter 5

General notation

X	A set of elements.
$x/x'/x_i$	An element $x/x'/x_i \in X$.
$\mathcal{P}(X)$	The power set of X.
S	A set in $\mathcal{P}(X)$.
$\succeq/\succeq'/\succeq_X$	A generic ranking, where $\succ/\succ'/\succ_X$ is its antisymmetric part and $\sim/\sim'/\sim_X$ its symmetric part. In this chapter we use this notation for several rankings. In Section 5.5 we use it to note the ranking over X obtained through MC lex-cel ($mclex$). In Section 5.7 we use this to note the ranking over norm systems composed of beneficial norms obtained through the norm system ranking function (nsr).
$\mathcal{R}(X)$	All possible rankings over a set (in this case X).
$\geq_{lex}$	The lexicographical order of tuples, see Definition 30.
le	The lex-cel (lexicographic excellence) function.
$\succeq_S$	A ranking over the power set $\mathcal{P}(X)$.
$\succeq_e$	The ranking produced by applying lex-cel.
θ	The function that builds the profile vector of lex-cel.
$\Sigma/\Sigma'/\Sigma_i$	An equivalence class of $\mathcal{P}(X)/\sim_S$.
$l/l'/l_i$	A label.
l_0	The neutral label.
L	A set of labels.
$>_L$	A linear order over the labels in L.
$\langle L, >_L \rangle$	A label system. We note the labels of a label system as l_i, where i marks its position with respect to the neutral label l_0.
LS	A label system.
sgn	The sign function, see Equation 5.2.
stg	The strength function, see Equation 5.3.
$c/c'/c_i$	A criterion.
C	A set of criteria.
λ	A labelling, in other words, a function $\lambda : X \times C \to L$ that given an element and a criterion returns the label of their relation.

$\mathcal{L}(X, C)$	All possible labelling functions over X and C.
$\succeq_C$	A ranking over the criteria (C).
mcr	An MC ranking function, hence $mcr : \mathcal{L}(X, C) \times \mathcal{R}(C) \to \mathcal{R}(X)$.
$\kappa/\kappa'/\kappa_i$	A criteria equivalence class, hence $\kappa \in C/\sim_C$.
s	A label strength.
s_{max}	The maximum label strength of the label system.
na	The net alignment function, see Equation 5.4.
μ	MC profiles used for MC lex-cel, $\mu(x)$ is the MC profile of x.
$mclex$	The MC lex-cel function.
sr	A social ranking function $sr : \mathcal{R}(\mathcal{P}(X)) \to \mathcal{R}(X)$.
$\succeq_P$	A ranking over $\mathcal{P}(X)$.
t	The transformation function that transforms a social ranking's input into an MC ranking's input, $t : \mathcal{R}(\mathcal{P}(X)) \to \mathcal{R}(C) \times \mathcal{L}(X, C)$.
c_S	The criterion associated to $S \in \mathcal{P}(X)$ by the transformation function t.

Value-aligned norm selection

Ag	A set of agents.
$\mathcal{A}$	A set of actions.
a	An action $a \in \mathcal{A}$.
$\mathcal{P}$	A set of propositions.
$\mathcal{PL}$	A propositional language (with propositions in $\mathcal{P}$ and the logical operator "and").
S	A set of states.
θ	A deontic operator, this can be $Prh/Per/Obl$ prohibition, permission or obligation respectively.
φ	A precondition $\varphi \subseteq \mathcal{P}$ (with an "and" semantic between propositions).
$n/n'/n_i$	A norm of the form $\langle \varphi, \theta(a) \rangle$.
N	A set of norms.
R	The set of norms relations, $R = \{R_x, R_g\}$.
R_x	The set of exclusive relations, $(n, n') \in R_x$ if n and n' are mutually exclusive norms.
R_g	The set of generalisation relations, $(n, n') \in R_x$ if n generalises n'.

$A(n)$	The ancestors of n, the norms that generalise n.
$S(n)$	The successors of n, the norms that are generalised by n.
$\langle N, R\rangle$	A norm net, composed of a set of norms and their relations.
$\bar{S}(n)$	The direct successors of n, the norms n' such that $(n, n') \in R_g$.
Ω	A norm system $\Omega \subseteq N$.
$v/v'/v_i$	A moral value.
V	A set of moral values.
$\succeq_v$	A ranking over the moral values in V.
$\langle V, \succeq_v\rangle$	A value system.
n_0	The neutral norm, an artificial norm that is neutral to all values.
N_{ben}	The set of beneficial norms, the norms that are more preferred than n_0.
$\succeq_n$	The ranking over norms obtained through MC lex-cel.
ben	A restriction function $ben : \mathcal{R}(N) \to \mathcal{R}(N_{ben})$, restricting rankings over norms to rankings over beneficial norms (N_{ben}) only.
$\succeq_n^{ben}$	The restriction of $\succeq_n$ through ben.
ale	The anti-lex-cel function.
nsr	The norm system ranking function $nsr : \mathcal{R}(V) \to \mathcal{R}(\mathcal{P}(N_{ben}))$. This function is defined as the composition of $mclex$, ben and ale.
$\succeq/\succeq'$	The ranking over norm systems composed of beneficial norms. $\succ/\succ'$ represents its antisymmetric part and $\sim/\sim'$ its symmetric part.

Appendix A

Implementation to solve VANS problems

In order to smooth the path for the applicability of our proposal, we have implemented a tool to solve a VANS problem specified in a human-comprehensible format. We have made the code publicly available at [Serramia et al., 2021d]. The repository contains two main python programs. The first one, VANS.py, is devoted to encode a description of a VANS problem into a BIP. The second program, BIPSolver.py, invokes CPLEX with the resulting BIP to provide the final solution:

- Firstly, VANS.py takes as input a VANS problem specification (a text file) containing the actions, values, action judgement, norms, value preferences, exclusivity relations, and generalisation relations of the problem at hand. VANS.py then produces, as output, a BIP file containing the maximization formula in Equation 3.8, the constraints, and the binary decision variables in a solver-readable format. The BIP contains constraints for the exclusivity and generalisation relations following equations 3.9 and 3.10. VANS.py also automatically includes constraints for non-aligned norms (having negative or neutral value alignment) as in Equation 3.11.

- Secondly, BIPSolver.py takes as input the resulting BIP file and solves the problem by invoking CPLEX. Its output is thus a solution file containing the assignment of the binary decision variables, so that only those norms that are assigned a value of 1 will be selected.

Appendix B

VANS generation

Algorithm 1 details how to generate a VANS problem. It considers six parameter variables: num_norms, the number of norms; num_val, the number of values (in our experiments this has been fixed to 10% of the number of norms); rel_density, the probability of two norms having a relation; incomp_perc, the probability of a relation being an incompatibility relation (otherwise, it is a generalisation relation); val_perc, the probability of a norm promoting/demoting a value (otherwise, the norm is unrelated, in our experiments this has been fixed to 20%); and class_prob, an equivalence class probability. When generating a new value, class_prob is the probability of the value being more preferred than the last generated (otherwise, they are equally preferred). This is used to determine the value equivalence classes to compute relevance as in Eq. 3.5. In our experiments this probability was fixed to 20%. The algorithm uses the functions: random_bool(prob), which returns True with probability prob (otherwise, it returns False); and random(x, y), which returns a number in the interval $[x, y]$. Lines in the algorithm come along with comments to make pseudo-code self explanatory.

Algorithm 1 Generation of artificial VANS problem

```
1: procedure GEN_VANS(num_norms, num_val, rel_density, incomp_perc,
   val_perc, class_prob)
2:     for i ∈ [1, num_norms] do
3:         norms ← n_i                                  ▷ Generate the norms
4:     end for
5:     for i ∈ [1, num_norms − 1] do
6:         for j ∈ [i + 1, num_norms] do  ▷ For each pair of norms (n_1, n_j)
7:             if random_bool(rel_density) then     ▷ Assign a relation with
   probability rel_density
8:                 if random_bool(incomp_perc) then     ▷ The relation is of
   incompatibility with prob. incomp_perc
9:                     incomp_rel ← (n_i, n_j)
10:                else  ▷ Otherwise, the relation is of generalisation (hence,
   probability 1-incomp_perc)
11:                    gen_rel ← (n_i, n_j)
12:                end if
13:            end if
14:        end for
15:    end for
16:    class ← 1        ▷ The equivalence class of the value, the greater the
   number, the more preferred the class
17:    for i ∈ [1, num_val] do
18:        values ← v_i                                 ▷ Generate the value
19:        relevance[v_i] ← 2^class                ▷ Assign relevance to the value
20:        if random_bool(class_prob) then        ▷ Change class to a more
   preferred one with probability class_prob
21:            class ← class + 1
22:        end if
23:    end for
```

```
24:     for n ∈ norms do
25:         for v ∈ values do                ▷ For each pair of norm n and value v
26:             if random_bool(val_perc) then   ▷ n promotes/demotes v (is not neutral with v) with probability val_perc
27:                 promotion(n, v) ← random(−1, 1)      ▷ The promotion is randomly selected from [-1, 1]
28:             else                                ▷ Otherwise, n and v are neutral
29:                 promotion(n, v) ← 0                  ▷ Assign promotion 0
30:             end if
31:         end for
32:     end for
33: end procedure
```

Appendix C

DSSP algorithm and implementation

Algorithm 3 encodes a DSSP into a BIP, and also writes this encoding into a file that can be fed into a BIP solver. This algorithm receives as input:

- a non-empty list of elements X;
- a list F of feature equivalence classes in descending order of preference (each equivalence class being a list of its indifferently preferred features, hence the feature order $f1 \succ_F f2 \sim_F f3 \succ_F f4$ would be represented as as $F = [[f1], [f2, f3], [f4]]$);
- a mapping $\mathfrak{f}$ relating elements to their features; and
- a list of constraints C (each constraint $c \in C$ being a string).

The algorithm uses several auxiliary functions: sort(l, k) sorts a list l in ascending order using as key a function k; write(s, file) writes a string s in a separated line in the given file; str(num) converts a numeric value num into a string; and get(l, i), which returns the element in position i in a list l. Finally, given two strings s and s', we represent string concatenation as $s + s'$. Notice also that "+" represents a string solely composed by character '+'.

Algorithm 2 provides a function used in Algorithm 3 to compute profile vectors according to the definition in Section 4.6. Algorithm 2 receives as input an element $x \in X$, a list F of feature equivalence classes in descending order of preference, and a mapping $\mathfrak{f}$ relating elements to their features. From that, it builds the profile of the element, $\mu(x)$ (as a list instead of a

vector), following Eq. 4.1. The auxiliary function has_feature in Algorithm 2 checks if an element has a given feature or not.

There is a publicly-available implementation of our DSSP encoder and solver at https://gitlab.iiia.csic.es/marcserr/dssp.

Example 29 below shows an example of the BIP output by Algorithm 3.

Example 29. *Suppose that a school wants to award three scholarships to their best last year students. Eight candidate students are considered, for which we assign decision variables* $X = \{s1, s2, ..., s8\}$ *(where* $si = 1$ *means student si is awarded with a scholarship and* $si = 0$ *means that si is not). The features that the school considers are: academic excellence (ax), good behaviour (gb), having helped the staff (hl), and punctuality (p). Academic excellence is the most important feature to compete for a scholarship, followed by good behaviour and helping staff members, which are indifferently preferred. Finally, punctuality is the least preferred feature. Hence, the feature order is:* $ax \succ gb \sim hl \succ p$*, meaning that F= [[ax], [gb, hl], [p]]. The features of each student are:*

s1: ax, gb, hl	*s2: ax, gb, hl, p*	*s3: gb, hl, p*	*s4: ax, gb, p*
s5: gb, hl, p	*s6: hl, p*	*s7: ax, gb, p*	*s8: gb, hl*

Finally, in terms of constraints, since there are only three scholarships, we have to consider a constraint that enforces that we have to exactly select three students: s1+s2+s3+s4+s5+s6+s7+s8=3. Furthermore, suppose that s1-s4 belong to one group, whereas s5-s8 belong to another group, and the school wants to give at least one scholarship to each group. Thus, we have to consider constraints "s1+s2+s3+s4>=1" and "s5+s6+s7+s8>=1". Then, the resulting BIP after applying Algorithm 3 would be :

Maximize
1s6 + 2s8 + 4s3 + 4s5 + 12s4 + 12s7 + 36s1 + 72s2
Subject To
s1+s2+s3+s4+s5+s6+s7+s8=3
s1+s2+s3+s4>=1
s5+s6+s7+s8>=1
Binaries
s1
s2
s3
s4

s5
s6
s7
s8
End

The solution to the BIP above is {s1, s2, s7}.

Algorithm 2 Compute profile vector μ (see Section 4.6)

```
1: procedure MU(x, F, f)
2:     mu ← An empty list                          ▷ Initialise profile vector
3:     for c ∈ F do        ▷ From most preferred to least preferred feature
   equivalence class
4:         num ← 0                          ▷ Counter of features of x in class c
5:         for f ∈ c do              ▷ For all features in the equivalence class
6:             if has_feature(x, f, f) then
7:                 num ← num + 1                    ▷ Increase feature counter
8:             end if
9:         end for
10:        append(num, mu)      ▷ Append num at the end of profile vector
11:    end for
12:    return mu
13: end procedure
```

Algorithm 3 Encoding a DSSP into a BIP

1: **procedure** BUILD_BIP(X, F, ʃ, C)
2: FILE bip ▷ Create an empty BIP file
3: pref ← 1 ▷ Initialise preference variable
4: toadd ← 1 ▷ This variable accumulates preferences while "traversing" the same equivalence class, and it is added to pref when changing to another equivalence class
5: i ← 1 ▷ Initialise index
6: sort(X, MU(x, F, ʃ)) ▷ Lex-cel uses μ in Alg.2 to sort X from least to most preferred
7: write("Maximize", bip) ▷ Write objective function
8: objfunc ← str(pref)+get(X, i) ▷ Initialise objective function to "$1x_1$"
9: **while** $i \leq |X|$ **do**
10: $i \leftarrow i + 1$ ▷ Increase index i
11: **if** MU(get(X,i),F, ʃ) = MU(get(X,i-1),F, ʃ) **then** ▷ $x_i \sim_e x_{i-1}$ in the same class
12: objfunc ← objfunc + "+" + str(pref) + get(X, i) ▷ eg: objfunc="$1x_1 + 1x_2$"
13: toadd ← toadd + pref
14: **else** ▷ $x_i \succ_e x_{i-1}$ in a more preferred equivalence class
15: pref ← pref + toadd
16: objfunc ← objfunc + "+" + str(pref) + get(X, i) ▷ eg.: objfunc="$1x_1 + 2x_2$"
17: toadd ← pref
18: **end if**
19: **end while**
20: write(objfunc, bip)
21: write("Subject To", bip) ▷ Write constraints
22: **for** $c \in C$ **do**
23: write(c, bip)
24: **end for**
25: write("Binaries", bip) ▷ Write binary decision variables
26: **for** $x \in X$ **do**
27: write(x, bip)
28: **end for**
29: write("End", bip)
30: **end procedure**

Appendix D

VANS algorithm and implementation

Algorithm 4 encodes a VANS problem into a BIP. The input of the algorithm contains: a non-empty set of norms N; a list of value equivalence classes V (each equivalence class being a list of its indifferently preferred values); a mapping relating norms to their promoted values $\mathfrak{f}$; a list of mutually exclusive relations (the relations being binary tuples); and a generalisation graph (implementing generalisation relations, being the parents of a norm, more general norms). First, the algorithm builds the constraints on norms discussed in Section 4.9 using Algorithm 5. Thereafter, it feeds N, V, $\mathfrak{f}$, and the obtained constraints into Algorithm 3 in appendix C.

Algorithm 5 uses several auxiliary functions (some of them already used in appendix C): append(x, list) appends x at the end of the list; get(l, i) returns the element in position i of the list l; parents(n, G) returns a list of all parents (direct or not) of n in the graph G; direct_siblings(n, G) returns a list of the direct siblings of n in G. Finally, given two strings s and s', we represent string concatenation as s + s'. Notice also that "+" represents a string solely composed by character '+'.

Example 30 below shows the BIP obtained after applying Algorithm 4 to Example 18.

Example 30. *BIP file:*
Maximize
3n1 + 3n2 + 1n3 + 1n4
Subject To
n1 + n3 <= 1
n2 + n4 <= 1

Binaries
n1
n2
n3
n4
End

The solution to the BIP above is {*n1, n2*}.

There is a publicly-available implementation of our encoder for VANS problems at https://gitlab.iiia.csic.es/marcserr/vans-problem.

Algorithm 4 Encoding a VANS into a BIP

```
procedure VANS_BIP(N, V, f, R_x, G)
    C ← VANS_CONSTRAINTS(N, R_x, R_g)        ▷ See Algorithm 5
    BUILD_BIP(N, V, f, C)                    ▷ See Algorithm 3 in appendix C
end procedure
```

Algorithm 5 Building constraints for VANS from norm relations

```
procedure VANS_CONSTRAINTS(N, R_x, G)
    C ← An empty list                                   ▷ Initialise constraints
    for (n1, n2) ∈ R_x do                ▷ Loop over norm exclusivity relations
        excon ← n1 + " + " + n2 + " <= 1"  ▷ Exclusivity constraint as
in Eq. 4.8
        append(excon, C)
    end for
    for n ∈ N do
        for p ∈ parents(n, G) do        ▷ Look at all the parents of norm n
            gencon ← n + " + " + p + " <= 1"    ▷ Generate constraints
based on Eq. 4.9
            append(gencon, C)
        end for
        if |direct_siblings(n, G)| > 1 then
            i ← 1
            gencon ← get(direct_siblings(n, G), i)   ▷ Generate constraints
based on Eq. 4.10
            while i ≤ |direct_siblings(n, G)| do
                i ← i + 1
                gencon ← gencon + " + " + get(direct_siblings(n, G), i)
            end while
            gencon ← gencon + " <= " + str(|direct_siblings(n, G)| - 1)
            append(gencon, C)
        end if
    end for
    return C
end procedure
```

Bibliography

[Abbeel and Ng, 2004] Abbeel, P. and Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pages 1–, New York, NY, USA. ACM.

[Abel et al., 2016] Abel, D., MacGlashan, J., and Littman, M. L. (2016). Reinforcement learning as a framework for ethical decision making. In *AAAI Work.: AI, Ethics, and Society*, volume 92.

[Ågotnes et al., 2007] Ågotnes, T., Van Der Hoek, W., Sierra, C., and Wooldridge, M. (2007). On the logic of normative systems. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, IJCAI '07, pages 1175–1180.

[Ågotnes and Wooldridge, 2010] Ågotnes, T. and Wooldridge, M. (2010). Optimal Social Laws. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 667–674.

[Ajmeri, 2018] Ajmeri, N. (2018). *Engineering Multiagent Systems for Ethics and Privacy-Aware Social Computing.* PhD thesis, North Carolina State University.

[Allouche et al., 2020] Allouche, T., Escoffier, B., Moretti, S., and Öztürk, M. (2020). Social ranking manipulability for the cp-majority, banzhaf and lexicographic excellence solutions. In *29th International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence {IJCAI-PRICAI-20}*, pages 17–23.

[Anderson et al., 1996] Anderson, D., Cavalier, R., and Covey, P. (1996). *A Right to Die?: The Dax Cowart Case an Ethical Case Study on Cd-Rom.* Routledge, London.

[Andrighetto et al., 2013] Andrighetto, G., Governatori, G., Noriega, P., and van der Torre, L. W. (2013). *Normative multi-agent systems*, volume 4. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[Arcos et al., 2005] Arcos, J. L., Esteva, M., Noriega, P., Rodríguez-Aguilar, J. A., and Sierra, C. (2005). Engineering open environments with electronic institutions. *Engineering applications of artificial intelligence*, 18(2):191–204.

[Arlegi, 2003] Arlegi, R. (2003). A note on bossert, pattanaik and xu's "choice under complete uncertainty: axiomatic characterization of some decision rules". *Economic Theory*, 22(1):219–225.

[Arrow, 2012] Arrow, K. J. (1951-2012). *Social choice and individual values*, volume 12. Yale university press.

[Atkinson et al., 2006] Atkinson, K., Bench-Capon, T. J. M., and McBurney, P. (2006). Parmenides : Facilitating deliberation in democracies. *Artificial Intelligence and Law*, 14:261–275.

[Audi, 1999] Audi, R. (1999). *The Cambridge Dictionary of Philosophy*. Cambridge University Press.

[Axelrod, 1986] Axelrod, R. (1986). An evolutionary approach to norms. *American political science review*, 80(4):1095–1111.

[Azar, 2004] Azar, O. H. (2004). What sustains social norms and how they evolve?: The case of tipping. *Journal of Economic Behavior & Organization*, 54(1):49–64.

[Barberà et al., 2004] Barberà, S., Bossert, W., and Pattanaik, P. K. (2004). Ranking sets of objects. In *Handbook of utility theory*, pages 893–977. Springer.

[Beauchamp and Childress, 2009] Beauchamp, T. L. and Childress, J. F. (2009). *Principles of biomedical ethics*. Oxford University Press, New York.

[Bench-Capon, 2016] Bench-Capon, T. (2016). Value-based reasoning and norms. *Artificial Intelligence for Justice*, pages 9–17.

[Bench-Capon et al., 2013] Bench-Capon, T., Prakken, H., Wyner, A., and Atkinson, K. (2013). Argument schemes for reasoning with legal cases using values. In *Proceedings of the fourteenth international conference on artificial intelligence and law*, pages 13–22.

[Bench-Capon and Atkinson, 2009] Bench-Capon, T. J. M. and Atkinson, K. (2009). Abstract argumentation and values. In *Argumentation in Artificial Intelligence*, pages 45–64. Springer.

[Bernardi et al., 2019] Bernardi, G., Lucchetti, R., and Moretti, S. (2019). Ranking objects from a preference relation over their subsets. *Social Choice and Welfare*, 52(4):589–606.

[Boddington, 2017] Boddington, P. (2017). *Towards a code of ethics for artificial intelligence*. Springer.

[Boella and van der Torre, 2004] Boella, G. and van der Torre, L. (2004). Regulative and constitutive norms in normative multiagent systems. *Proceedings of KR'04*, pages 255–265.

[Boella et al., 2006] Boella, G., van der Torre, L., and Verhagen, H. (2006). Introduction to normative multiagent systems. *Computational & Mathematical Organization Theory*, 12(2-3):71–79.

[Bossert et al., 1994] Bossert, W., Pattanaik, P. K., and Xu, Y. (1994). Ranking opportunity sets: an axiomatic approach. *Journal of Economic theory*, 63(2):326–345.

[Bou et al., 2006] Bou, E., López-Sánchez, M., and Rodríguez-Aguilar, J. A. (2006). Norm adaptation of autonomic electronic institutions with multiple goals. *ITSSA*, 1(3):227–238.

[Bouyssou, 1986] Bouyssou, D. (1986). Some remarks on the notion of compensation in mcdm. *European Journal of Operational Research*, 26(1):150–160.

[Brandenburg et al., 2013] Brandenburg, F. J., Gleißner, A., and Hofmeier, A. (2013). Comparing and aggregating partial orders with kendall tau distances. *Discrete Mathematics, Algorithms and Applications*, 5(02):1360003.

[Brandt et al., 2016] Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D. (2016). *Handbook of computational social choice*. Cambridge University Press.

[Campos et al., 2013] Campos, J., López-Sánchez, M., Salamó, M., Avila, P., and Rodríguez-Aguilar, J. A. (2013). Robust regulation adaptation in multi-agent systems. *ACM Transactions on Autonomous and Adaptive Systems*, 8:1–27.

[Castelfranchi, 1999] Castelfranchi, C. (1999). Prescribed mental attitudes in goal-adoption and norm-adoption. *Artificial Intelligence and Law*, 7(1):37–50.

[Charisi et al., 2017] Charisi, V., Dennis, L., Fisher, M., Lieck, R., Matthias, A., Slavkovik, M., Sombetzki, J., Winfield, A. F. T., and Yampolskiy, R. (2017). Towards moral autonomous systems.

[Chisholm, 1963] Chisholm, R. M. (1963). Supererogation and offence: A conceptual scheme for ethics. *Ratio (Misc.)*, 5(1):1.

[City of Reykjavík, 2012] City of Reykjavík (2012). Participation portal. http://reykjavik.is/en/participation. Accessed 06/2021.

[Cointe et al., 2016] Cointe, N., Bonnet, G., and Boissier, O. (2016). Ethical judgment of agents' behaviors in multi-agent systems. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 1106–1114. International Foundation for Autonomous Agents and Multiagent Systems.

[Consider.it, 2010] Consider.it (2010). https://consider.it/. Accessed 06/2021.

[Consul, 2015] Consul (2015). http://consulproject.org/en/. Accessed 06/2021.

[Cooper, 1993] Cooper, D. (1993). *Value pluralism and ethical choice*. St. Martin Press, Inc.

[Cranefield et al., 2017] Cranefield, S., Winikoff, M., Dignum, V., and Dignum, F. (2017). No pizza for you: Value-based plan selection in bdi agents. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 178–184.

[Decidim, 2016] Decidim (2016). Decidim Barcelona. https://www.decidim.barcelona/. Accessed 06/2021.

[Di Tosto and Dignum, 2012] Di Tosto, G. and Dignum, F. (2012). Simulating social behaviour implementing agents endowed with values and drives. In *International Workshop on Multi-Agent Systems and Agent-Based Simulation*, pages 1–12. Springer.

[Dignum, 1999] Dignum, F. (1999). Autonomous agents with norms. *Artif. Intell. Law*, 7(1):69–79.

[Dignum, 2017] Dignum, V. (2017). Responsible autonomy. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, pages 4698–4704.

[Dybalova et al., 2013] Dybalova, D., Testerink, B., Dastani, M., and Logan, B. (2013). A framework for programming norm-aware multi-agent systems. In *International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems*, pages 364–380. Springer.

[Etzioni and Etzioni, 2016] Etzioni, A. and Etzioni, O. (2016). Ai assisted ethics. *Ethics and Information Technology*, 18(2):149–156.

[Fieser and Dowden, 2021] Fieser, J. and Dowden, B. (2021). Ethics. *The Internet Encyclopedia of Philosophy*. Accessed 06/2021.

[Fishburn, 1992] Fishburn, P. C. (1992). Signed orders and power set extensions. *Journal of Economic Theory*, 56(1):1–19.

[Fitoussi and Tennenholtz, 2000] Fitoussi, D. and Tennenholtz, M. (2000). Choosing social laws for multi-agent systems: Minimality and simplicity. *Artificial Intelligence*, 119(1-2):61–101.

[Floridi and Sanders, 2004] Floridi, L. and Sanders, J. (2004). On the morality of artificial agents. *Minds and Machines*, 14(3):349–379.

[Frankena, 1973] Frankena, W. K. (1973). *Ethics, 2nd edition*. Englewood Cliffs, N.J. : Prentice-Hall.

[Friedler et al., 2019] Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 329–338.

[Gale and Shapley, 1962] Gale, D. and Shapley, L. S. (1962). College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15.

[Garcia-Camino et al., 2005] Garcia-Camino, A., Noriega, P., and Rodriguez-Aguilar, J. A. (2005). Implementing norms in electronic institutions. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*, AAMAS '05, page 667–673, New York, NY, USA. Association for Computing Machinery.

[Garcia-Camino et al., 2009] Garcia-Camino, A., Rodriguez-Aguilar, J. A., Sierra, C., and Vasconcelos, W. (2009). Constraint rule-based programming of norms for electronic institutions. *Autonomous agents and multi-agent systems*, 18(1):186–217.

[Gert and Gert, 2020] Gert, B. and Gert, J. (2020). The Definition of Morality. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2020 edition.

[Greco et al., 2019] Greco, S., Ishizaka, A., Tasiou, M., and Torrisi, G. (2019). On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. *Social Indicators Research*, 141(1):61–94.

[Griffiths and Luck, 2010] Griffiths, N. and Luck, M. (2010). Norm Emergence in Tag-Based Cooperation. In *Proceedings of COIN*, pages 79–86.

[Grossi and Dignum, 2005] Grossi, D. and Dignum, F. (2005). From abstract to concrete norms in agent institutions. In *Proceedings of the Third international conference on Formal Approaches to Agent-Based Systems*, FAABS'04, pages 12–29, Berlin, Heidelberg. Springer-Verlag.

[Gurobi Optimization, 2010] Gurobi Optimization (2010). Gurobi. http://www.gurobi.com/. Accessed 06/2021.

[Haerpfer et al., 2020] Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., Lagos, M., Norris, P., Ponarin, E., and Puranen, B. (2020). World values survey.

[Hansson, 2001] Hansson, S. O. (2001). *The Structure of Values and Norms*. Cambridge Studies in Probability, Induction and Decision Theory. Cambridge University Press, Cambridge.

[Hansson, 2013] Hansson, S. O. (2013). Representing supererogation. *Journal of Logic and Computation*, 25(2):443–451.

[Hansson, 2018] Hansson, S. O. (2018). *Formal Investigations of Value*, pages 499–522. Springer International Publishing, Cham.

[Haret et al., 2018] Haret, A., Khani, H., Moretti, S., and Öztürk, M. (2018). Ceteris paribus majority for social ranking. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 303–309. International Joint Conferences on Artificial Intelligence Organization.

[Harris Jr et al., 2013] Harris Jr, C. E., Pritchard, M. S., Rabins, M. J., James, R., and Englehardt, E. (2013). *Engineering ethics: Concepts and cases*. Cengage Learning.

[Hartman, 1967] Hartman, R. S. (1967). *The structure of value: Foundations of scientific axiology*. Southern Illinois University Press.

[Heyd, 2019] Heyd, D. (2019). Supererogation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2019 edition.

[Horgan and Timmons, 2010] Horgan, T. and Timmons, M. (2010). Untying a knot from the inside out: Reflections on the "paradox" of supererogation. *Social Philosophy and Policy*, 27:29 – 63.

[Hubner et al., 2007] Hubner, J. F., Sichman, J. S., and Boissier, O. (2007). Developing organised multiagent systems using the moise+ model: programming issues at the system and agent levels. *International Journal of Agent-Oriented Software Engineering*, 1(3-4):370–395.

[IBM, 1988] IBM (1988). Cplex. https://www.ibm.com/analytics/data-science/prescriptive-analytics/cplex-optimizer. Accessed 06/2021.

[IEEE Standards Association, 2016] IEEE Standards Association (2016). The ieee global initiative for ethical considerations in artificial intelligence and autonomous systems. http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html. Accessed 06/2021.

[Janoff-Bulman et al., 2009] Janoff-Bulman, R., Sheikh, S., and Hepp, S. (2009). Proscriptive versus prescriptive morality: Two faces of moral regulation. *Journal of personality and social psychology*, 96(3):521.

[Karp, 1972] Karp, R. M. (1972). Reducibility among combinatorial problems. In *Complexity of computer computations*, pages 85–103. Springer.

[Kasenberg et al., 2018] Kasenberg, D., Arnold, T., and Scheutz, M. (2018). Norms, rewards, and the intentional stance: Comparing machine learning approaches to ethical training. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 184–190. ACM.

[Kemeny, 1959] Kemeny, J. G. (1959). Mathematics without numbers. *Daedalus*, 88(4):577–591.

[Kendall, 1938] Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

[Khani et al., 2019] Khani, H., Moretti, S., and Öztürk, M. (2019). An ordinal banzhaf index for social ranking. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 378–384. International Joint Conferences on Artificial Intelligence Organization.

[Klein, 2012] Klein, M. (2012). Enabling large-scale deliberation using attention-mediation metrics. *Comput. Supported Coop. Work*, 21(4–5):449–473.

[Kohler et al., 2014] Kohler, T., Steghoefer, J.-P., Busquets, D., and Pitt, J. (2014). The value of fairness: Trade-offs in repeated dynamic resource allocation. In *Self-Adaptive and Self-Organizing Systems (SASO), 2014 IEEE Eighth International Conference on*, pages 1–10. IEEE.

[Kollingbaum et al., 2006] Kollingbaum, M. J., Norman, T. J., Preece, A., and Sleeman, D. (2006). Norm conflicts and inconsistencies in virtual organisations. In *International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems (2006)*, pages 245–258. Springer.

[Kollingbaum et al., 2007] Kollingbaum, M. J., Norman, T. J., Preece, A., and Sleeman, D. (2007). *Norm Conflicts and Inconsistencies in Virtual Organisations*, pages 245–258. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Lieberman and Hillier, 2005] Lieberman, G. J. and Hillier, F. S. (2005). *Introduction to operations research*. McGraw-Hill, New York.

[Liscio et al., 2021a] Liscio, E., van der Meer, M., Jonker, C. M., and Murukannaiah, P. K. (2021a). *A Collaborative Platform for Identifying Context-Specific Values*, page 1773–1775. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC.

[Liscio et al., 2021b] Liscio, E., van der Meer, M., Siebert, L. C., Jonker, C. M., Mouter, N., and Murukannaiah, P. K. (2021b). *Axies: Identifying and Evaluating Context-Specific Values*, page 799–808. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC.

[Loomio, 2012] Loomio (2012). https://www.loomio.org/. Accessed 06/2021.

[Lopez et al., 2012] Lopez, S. J., Snyder, C., Handelsman, M. M., Knapp, S., and Gottlieb, M. C. (2012). Positive ethics: Themes and variations.

[Lopez-Sanchez et al., 2017] Lopez-Sanchez, M., Serramia, M., Rodriguez-Aguilar, J. A., Morales, J., and Wooldridge, M. (2017). Automating decision making to help establish norm-based regulations. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS'17)*, pages 1613–1615. International Foundation for Autonomous Agents and Multiagent Systems.

[López y López et al., 2002] López y López, F., Luck, M., and d'Inverno, M. (2002). Constraining autonomy through norms. In *AAMAS*, pages 674–681. ACM.

[Luo et al., 2017] Luo, J., Meyer, J.-J., and Knobbout, M. (2017). Reasoning about opportunistic propensity in multi-agent systems. In *AAMAS 2017 Workshops, Best Papers.*, pages 1–16.

[McLaren, 2006] McLaren, B. M. (2006). Computational models of ethical reasoning: Challenges, initial steps, and future directions. *IEEE intelligent systems*, 21(4):29–37.

[McLaren and Ashley, 1995] McLaren, B. M. and Ashley, K. D. (1995). Case-based comparative evaluation in truth-teller. In *the Proceedings From the Seventeenth Annual Conference of the Cognitive Science Society.*

[McNamara, 2011] McNamara, P. (2011). Praise, blame, obligation, and dwe: Toward a framework for classical supererogation and kin. *Journal of Applied Logic*, 9(2):153–170.

[Meena and Bharadwaj, 2020] Meena, R. and Bharadwaj, K. K. (2020). A genetic algorithm approach for group recommender system based on partial rankings. *Journal of Intelligent Systems*, 29(1):653–663.

[Mercuur et al., 2019] Mercuur, R., Dignum, V., Jonker, C., et al. (2019). The value of values and norms in social simulation. *Journal Artificial Societies and Social Simulation*, 22(1):1–9.

[Meyer and Wieringa, 1993] Meyer, J.-J. C. and Wieringa, R. J., editors (1993). *Deontic logic in computer science: normative system specification*. John Wiley and Sons Ltd., Chichester, UK.

[Modgil, 2006] Modgil, S. (2006). Value based argumentation in hierarchical argumentation frameworks. In *Proceedings of the 2006 Conference on Computational Models of Argument: Proceedings of COMMA 2006*, pages 297–308, Amsterdam, The Netherlands. IOS Press.

[Montague, 1989] Montague, P. (1989). Acts, agents, and supererogation. *American Philosophical Quarterly*, 26(2):101–111.

[Montes and Sierra, 2021] Montes, n. and Sierra, C. (2021). Value-guided synthesis of parametric normative systems. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2021)*, pages 907–915. International Foundation for Autonomous Agents and Multiagent Systems.

[Morales et al., 2015a] Morales, J., Lopez-Sanchez, M., Rodriguez-Aguilar, J. A., Vasconcelos, W., and Wooldridge, M. (2015a). On-line automated synthesis of compact normative systems. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 10(1):2:1–2:33.

[Morales et al., 2013] Morales, J., Lopez-Sanchez, M., Rodriguez-Aguilar, J. A., Wooldridge, M., and Vasconcelos, W. (2013). Automated synthesis of normative systems. In *AAMAS 2013*, pages 483–490.

[Morales et al., 2014] Morales, J., Lopez-Sanchez, M., Rodriguez-Aguilar, J. A., Wooldridge, M., and Vasconcelos, W. (2014). Minimality and simplicity in the on-line automated synthesis of normative systems. In *AAMAS 2014*, pages 109–116, Richland, SC. IFAAMAS.

[Morales et al., 2015b] Morales, J., Lopez-Sanchez, M., Rodriguez-Aguilar, J. A., Wooldridge, M., and Vasconcelos, W. (2015b). Synthesising liberal normative systems. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '15, pages 433–441.

[Morales et al., 2015c] Morales, J., Mendizábal, I., Sánchez-Pinsach, D., López-Sánchez, M., and Rodriguez-Aguilar, J. A. (2015c). Using iron to build frictionless on-line communities. *AI Communications*, 28(1):55–71.

[Morales et al., 2018] Morales, J., Wooldridge, M., Rodríguez-Aguilar, J. A., and López-Sánchez, M. (2018). Off-line synthesis of evolutionarily stable normative systems. *Autonomous Agents and Multi-Agent Systems*, 32(5):635–671.

[Moretti and Öztürk, 2017] Moretti, S. and Öztürk, M. (2017). Some axiomatic and algorithmic perspectives on the social ranking problem. In *International Conference on Algorithmic DecisionTheory*, pages 166–181. Springer.

[Murukannaiah et al., 2020] Murukannaiah, P. K., Ajmeri, N., Jonker, C. M., and Singh, M. P. (2020). New foundations of ethical multiagent systems. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, page 1706–1710, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

[New Zealand's Ministry of Justice, 2015] New Zealand's Ministry of Justice (2015). Consultations hub. https://consultations.justice.govt.nz/. Accessed 06/2021.

[Parlement & Citoyens, 2013] Parlement & Citoyens (2013). https://parlement-et-citoyens.fr/. Redirected to https://purpoz.com/projects.

[Pattanaik and Peleg, 1984] Pattanaik, P. K. and Peleg, B. (1984). An axiomatic characterization of the lexicographic maximin extension of an ordering over a set to the power set. *Social Choice and Welfare*, 1(2):113–122.

[Pereira-Moniz and Saptawijaya, 2016] Pereira-Moniz, L. and Saptawijaya, A. (2016). *Programming machine ethics*, volume 26. Springer.

[Perello-Moragues and Noriega, 2020] Perello-Moragues, A. and Noriega, P. (2020). *Using Agent-Based Simulation to Understand the Role of Values in Policy-Making*, pages 355–369. Springer, Cham.

[Petitions, 2015] Petitions (2015). UK Government and Parliament. https://petition.parliament.uk/. Accessed 06/2021.

[Petruzzi et al., 2015] Petruzzi, P. E., Busquets, D., and Pitt, J. (2015). A generic social capital framework for optimising self-organised collective action. In *Self-Adaptive and Self-Organizing Systems (SASO), 2015 IEEE 9th International Conference on*, pages 21–30. IEEE.

[Pitt et al., 2014] Pitt, J., Busquets, D., and Macbeth, S. (2014). Distributive justice for self-organised common-pool resource management. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 9(3):14.

[Riveret et al., 2014] Riveret, R., Artikis, A., Pitt, J., and Nepomuceno, E. G. (2014). Self-governance by transfiguration: From learning to prescription changes. In *Proceedings of the 2014 IEEE Eighth International Conference on Self-Adaptive and Self-Organizing Systems*, SASO '14, page 70–79, USA. IEEE Computer Society.

[Rodriguez-Soto et al., 2020] Rodriguez-Soto, M., Lopez-Sanchez, M., and Rodríguez-Aguilar, J. A. (2020). A structural solution to sequential moral dilemmas. In *Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Aystems (AAMAS 2020)*.

[Roobavannan et al., 2018] Roobavannan, M., Van Emmerik, T. H., Elshafei, Y., Kandasamy, J., Sanderson, M. R., Vigneswaran, S., Pande, S., and Sivapalan, M. (2018). Norms and values in sociohydrological models. *Hydrology and Earth System Sciences*, 22(2):1337–1349.

[Roth and Sotomayor, 1992] Roth, A. E. and Sotomayor, M. (1992). Two-sided matching. *Handbook of game theory with economic applications*, 1:485–541.

[Roy and Słowiński, 2013] Roy, B. and Słowiński, R. (2013). Questions guiding the choice of a multicriteria decision aiding method. *EURO Journal on Decision Processes*, 1(1):69–97.

[Russell, 2019] Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin.

[Santhanam, 2016] Santhanam, G. R. (2016). Qualitative optimization in software engineering: A short survey. *Journal of Systems and Software*, 111:149–156.

[Savarimuthu et al., 2013] Savarimuthu, B., Cranefield, S., Purvis, M., and Purvis, M. (2013). Identifying prohibition norms in agent societies. *Artificial Intelligence and Law*, 21(1):1–46.

[Savarimuthu et al., 2007] Savarimuthu, B. T. R., Purvis, M., Cranefield, S., and Purvis, M. (2007). Mechanisms for norm emergence in multiagent societies. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, AAMAS '07, pages 173:1–173:3, New York, NY, USA. ACM.

[Schwartz, 2012] Schwartz, S. H. (2012). An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):2307–0919.

[Sen, 2017] Sen, A. (2017). *Collective Choice and Social Welfare*. Harvard University Press.

[Sen and Airiau, 2007] Sen, S. and Airiau, S. (2007). Emergence of norms through social learning. In *IJCAI*, pages 1507–1512.

[Sensoy et al., 2012] Sensoy, M., Norman, T. J., Vasconcelos, W. W., and Sycara, K. (2012). Owl-polar: A framework for semantic policy representation and reasoning. *Journal of Web Semantics*, 12:148–160.

[Serramia, 2018] Serramia, M. (2018). Ethics in norm decision making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 378–379, New York, NY, USA. Association for Computing Machinery.

[Serramia et al., 2019a] Serramia, M., Ganzer-Ripoll, J., López-Sánchez, M., Rodríguez-Aguilar, J. A., Criado, N., Parsons, S., Escobar, P., and Fernández, M. (2019a). Citizen support aggregation methods for participatory platforms. In Sabater-Mir, J., Torra, V., Aguiló, I., and Hidalgo, M. G., editors, *Artificial Intelligence Research and Development - Proceedings of the 22nd International Conference of the Catalan Association for Artificial Intelligence, CCIA 2019, Mallorca, Spain, 23-25 October 2019*, volume 319 of *Frontiers in Artificial Intelligence and Applications*, pages 9–18, Amsterdam. IOS Press.

[Serramia et al., 2021a] Serramia, M., Lopez-Sanchez, M., Moretti, S., and Rodriguez-Aguilar, J. A. (2021a). On the dominant set selection problem and its application to value alignment. *Autonomous Agents and Multi-Agent Systems (JAAMAS)*.

[Serramia et al., 2020] Serramia, M., Lopez-Sanchez, M., and Rodriguez-Aguilar, J. A. (2020). A qualitative approach to composing value-aligned norm systems. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1233–1241.

[Serramia et al., 2021b] Serramia, M., López-Sánchez, M., and Rodríguez-Aguilar, J. A. (2021b). Algorithm to generate the BIP encoding of a DSSP problem. https://gitlab.iiia.csic.es/marcserr/dssp.

[Serramia et al., 2021c] Serramia, M., López-Sánchez, M., and Rodríguez-Aguilar, J. A. (2021c). Algorithm to generate the BIP encoding of a VANS problem with qualitative input. https://gitlab.iiia.csic.es/marcserr/vans-problem.

[Serramia et al., 2021d] Serramia, M., López-Sánchez, M., and Rodríguez-Aguilar, J. A. (2021d). Algorithm to generate the BIP encoding of a VANS problem with quantitative input. https://gitlab.iiia.csic.es/marcserr/vans-quant.

[Serramia et al., 2019b] Serramia, M., López-Sánchez, M., Rodríguez-Aguilar, J. A., and Escobar, P. (2019b). Optimising participatory budget allocation: The decidim use case. In Sabater-Mir, J., Torra, V., Aguiló, I., and Hidalgo, M. G., editors, *Artificial Intelligence Research and Development - Proceedings of the 22nd International Conference of the Catalan Association for Artificial Intelligence, CCIA 2019, Mallorca, Spain, 23-25 October 2019*, volume 319 of *Frontiers in Artificial Intelligence and Applications*, pages 193–202. IOS Press.

[Serramia et al., 2018a] Serramia, M., López-Sánchez, M., Rodríguez-Aguilar, J. A., Morales, J., Wooldridge, M., and Ansotegui, C. (2018a). Exploiting moral values to choose the right norms. In *Proceedings of the 1st Conference on artificial intelligence, ethics and society (AIES'18)*, pages 1–7.

[Serramia et al., 2018b] Serramia, M., Lopez-Sanchez, M., Rodriguez-Aguilar, J. A., Rodriguez, M., Wooldridge, M., Morales, J., and Ansotegui, C. (2018b). Moral values in norm decision making. In *Proceedings*

of the 17th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'18), pages 1294–1302. International Foundation for Autonomous Agents and Multiagent Systems.

[Sethi and Somanathan, 1996] Sethi, R. and Somanathan, E. (1996). The evolution of social norms in common property resource use. *The American Economic Review*, pages 766–788.

[Shoham and Tennenholtz, 1995] Shoham, Y. and Tennenholtz, M. (1995). On social laws for artificial agent societies: off-line design. *Artificial Intelligence*, 73(1-2):231–252.

[Shoham and Tennenholtz, 1997] Shoham, Y. and Tennenholtz, M. (1997). On the emergence of social conventions: modeling, analysis, and simulations. *Artificial Intelligence*, 94(1-2):139–166.

[Sierra et al., 2019] Sierra, C., Osman, N., Noriega, P., Sabater-Mir, J., and Perello-Moragues, A. (2019). Value alignment: a formal approach. In *Responsible Artificial Intelligence Agents Workshop (RAIA) in AAMAS*, Montreal, Canada. IFAAMAS.

[Słowiński et al., 2002] Słowiński, R., Greco, S., and Matarazzo, B. (2002). Axiomatization of utility, outranking and decision rule preference models for multiple-criteria classification problems under partial inconsistency with the dominance principle. *Control and Cybernetics*, 31(4):1005–1035.

[Stern et al., 1999] Stern, P. C., Dietz, T., Abel, T., Guagnano, G. A., and Kalof, L. (1999). A value-belief-norm theory of support for social movements: The case of environmentalism. *Human Ecology Review*, 6(2):81–97.

[Sugawara, 2011] Sugawara, T. (2011). Emergence and stability of social conventions in conflict situations. In Walsh, T., editor, *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 371–378. IJCAI/AAAI.

[Sun et al., 2019] Sun, F.-Y., Chang, Y.-Y., Wu, Y.-H., and Lin, S.-D. (2019). A regulation enforcement solution for multi-agent reinforcement learning. In *AAMAS Conf.*

[Tielman et al., 2018] Tielman, M., Jonker, C., and van Riemsdijk, B. (2018). What should i do? deriving norms from actions,values and context. In *Tenth International Workshop Modelling and Reasoning in Context*, volume 2134, pages 35–40.

[Timmons, 2012] Timmons, M. (2012). *Moral theory: An introduction.* Rowman & Littlefield Pub.

[Urmson, 1958] Urmson, J. O. (1958). Saints and heroes. In Melden, A. I., editor, *Essays in Moral Philosophy.* University of Washington Press.

[van de Poel and Royakkers, 2011] van de Poel, I. and Royakkers, L. (2011). *Ethics, Technology, and Engineering: An Introduction.* Wiley-Blackwell.

[van der Weide et al., 2009] van der Weide, T. L., Dignum, F., Meyer, J.-J. C., Prakken, H., and Vreeswijk, G. A. W. (2009). Practical reasoning using values: Giving meaning to values. In *Proceedings of the 6th International Conference on Argumentation in Multi-Agent Systems*, ArgMAS'09, page 79–93, Berlin, Heidelberg. Springer-Verlag.

[Vasconcelos et al., 2009] Vasconcelos, W. W., Kollingbaum, M. J., and Norman, T. J. (2009). Normative conflict resolution in multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 19(2):124–152.

[Villatoro et al., 2011] Villatoro, D., Sabater-Mir, J., and Sen, S. (2011). Social instruments for robust convention emergence. In *IJCAI*, pages 420–425.

[von Wright, 1963] von Wright, G. H. (1963). The varieties of goodness. *Ethics*, 74(3):223–225.

[Wallach and Allen, 2008] Wallach, W. and Allen, C. (2008). *Moral machines: teaching robots right from wrong.* Oxford University press.