EPiC Series in Computing

Volume 55, 2018, Pages 134–147

GCAI-2018. 4th Global Conference on Artificial Intelligence

Responsible Technologies

Nardine Osman and Carles Sierra

Artificial Intelligence Research Institute (IIIA-CSIC), Barcelona
{nardine,sierra}@iiia.csic.es

Abstract

With the current surge of interest in ethics in AI, we present our position with respect to these challenges. Our proposal, responsible technologies, aims to (1) address a number of the ethical challenges put forward in AI, and (2) provide the first building blocks towards the *development* of ethical AI systems. The current discussion on how to address ethics in AI usually focuses on issues like policies, education, or research culture. There is no computational method yet mature enough to address ethics in AI. We break ground by proposing new methods and tools, underpinned by multidisciplinary research, that can make humans and machines understand their respective dynamic goals while strictly abiding by the values that inspire our societies. This position paper presents our plan of work for the development of responsible technologies that embed values *within* technology through what we refer to as *ethics by construction*.

1 The Why: The Rise of Ethics in AI

The risks of artificial intelligence (AI) are high on the agendas of top AI experts and enterprises.[1][2] The wide application of AI is touching our lives in many ways, fuelling the recent surge in interest in ethics in AI. Many discussions about the risks of AI have focused on future AI and its impact on our lives, such as when artificial general intelligence (AGI) is achieved.[3] However, we argue that ethical considerations are much needed *today* for many existing narrow AI systems. As explained by Professor Dan Weld, narrow AI systems can be catastrophic too, noting that "Knight Capital's automated trading system is much less intelligent than Google DeepMind's AlphaGo, but the former lost $440 million in just forty-five minutes. AlphaGo hasn't and can't hurt anyone."[4]

[1] Top AI scientists have been debating artificial general intelligence (AGI) and the risks of AI in general. Open letters and declarations are being signed by top AI scientists warning about the risks of AI. Some examples are the Open Letter on AI (www.futureoflife.org/ai-open-letter), the Barcelona Declaration for the Proper Development and Use of Artificial Intelligence in Europe (www.iiia.csic.es/barcelonadeclaration/), and the Open Letter on Autonomous Weapons (www.futureoflife.org/open-letter-autonomous-weapons/).

[2] Institutes and foundations are being set up to promote beneficial AI. Some examples are the Future of Life Institute (www.futureoflife.org), the Partnership on AI (www.partnershiponai.org), Centre for the Study of Existential Risk (www.cser.ac.uk), and OpenAI (www.openai.com).

[3] One interesting example is the panel at the Beneficial AI Conference in 2017, which discussed the likelihood and the possible outcome of human-level AGI, and what would we like to happen: www.youtube.com/watch?v=OFBwz4R6Fi0

[4] www.futureoflife.org/2017/03/23/ai-risks-principle/

One pressing issue, for example, is explainability in AI. Machine learning today is helping with parole decisions, loan decisions, and even job decisions. Racism and sexism of the algorithms is emerging. Their racism against African Americans has been revealed in parole decisions in the US. In job recruitment, it has been noted that different job ads were targeting different ethnic groups.[5] Explainable AI is now a necessity, to help us understand how critical decisions are made. But this also raises the question of what values do we want our AI systems to adhere to? Do we really want a racist algorithm? As such, the value alignment problem has also been gaining ground. It states that "highly autonomous AI systems should be designed so that their goals and behaviours can be assured to align with human values throughout their operation".[6] In fact, the list of ethical concerns can be a long and daunting list,[7] with issues like *explainability*, *transparency*, *accountability* and *responsibility*, to name a few. With such a wide range of concerns, we choose to focus on what we refer to as *responsible technologies*, covering a selected number of the raised concerns. We present our focus and objectives next.

2 The What: Focus and Objectives

We propose the term responsible technologies, getting inspiration from the European Commission's work on responsible research and innovation, which has been declared a key action of the 'Science with and for Society' objective and a cross-cutting issue in Horizon 2020, stating that "research and innovation must respond to the needs and ambitions of society, reflect its values, and be responsible". We adopt, and slightly adapt, this declaration to define **responsible technologies** as **technologies that respond to the *needs* and ambitions of society, reflect its *values*, and put people in *control***. As such, we place *needs* and *values* as the basis of responsible technologies, and we introduce the idea of putting the humans in *control* of their technologies. This is in line with many requirements already put forward by several declarations and open letters on AI, such as the following AI principles developed at the 2017 Beneficial AI Conference: 1) *value alignment*, which states that AI systems should be designed so that their goals and behaviours can be assured to align with human values throughout their operation; 2) *human values*, which states that AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity; 3) *personal privacy*, which states that people should have the right to access, manage and control the data they generate, given AI systems' power to analyse and utilise that data; 4) *shared benefit*, which states that AI technologies should benefit and empower as many people as possible; and 5) *human control*, which states that humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives. As such, and in addition to basing responsible technologies on people's needs and values (which we believe addresses principles 1–4 above), we also put the humans in control of their technology so that they have a direct say on how their technology evolves according to their evolving needs and values (addressing principle 5 above).

Our **objectives** may be summarised as follows. ***(1)*** Develop a novel methodology and mechanisms for the design and development of responsible technologies that are based on people's needs and values, and evolve with people's evolving needs and values. ***(2)*** Give people control over their technologies so they can decide amongst themselves on their needs and values, and how their technology should behave accordingly. We present in the following section our

[5] www.newscientist.com/article/mg23230971-200-the-irresistible-rise-of-artificial-intelligence/

[6] www.futureoflife.org/2017/02/03/align-artificial-intelligence-with-human-values/

[7] www.futureoflife.org/ai-principles/

proposal on *how* to achieve the objectives above. However, first, we close this section with a motivating example illustrating our vision.

A Motivating Example: an Illustration of our Vision We get inspiration from the uHelp app, an app that has been developed by the authors and tested with a community of single parents in Barcelona.[8] uHelp essentially allows people to find help within their social network for everyday tasks through an intelligent search algorithm that floods people's social network looking for trustworthy volunteers.

Say Sofia uses the uHelp app to look for help with dropping off her daughter Cecilia at her Karate class. David volunteers, and Sofia accepts him for this task (Figure 1a). Later that evening, after David informs Sofia that the school has complained about the e-signed authorisation for not containing his ID card number, Sofia decides to ask her uHelp community members to add the ID card number to users' profiles. She sends a "request for change" message, where the suggested changes are specified in a natural language (Figure 1b). Community members start discussing the issue, in natural language (or possibly a controlled natural language as we shortly discuss), supported by information from the system (Figure 1c). Discussing values (e.g., security, privacy) becomes an integral part of the discussion. For example, while some oppose this as it violates people's privacy, others support the proposal as it promotes the security of children. At this stage, community members may also give their opinion on each others arguments (e.g., the thumbs up in Figure 1c). The system is capable of analysing the arguments and counter-arguments (either through the use of natural language processing, controlled natural language, or argument schemes and critical questions, as illustrated in Section 3.2), generating an argumentation graph, and assessing the community's preferred arguments (and requirements) (Figure 1d). This allows the system to present advice to the users with the aim of improving their final decision: an example of this can be seen at the bottom of Figure 1c where the system informs the users about the percentage of people preferring security over privacy. Any member can switch between the two views of Figures 1c and 1d during the discussion phase. When the community believes it has a proposal to vote on, it does so. If an agreement is finally reached as a result of the final voting round (Figure 1e), then the resulting changes update the app and its GUI in an automated manner (Figure 1f).

Our example shows how an existing technology, such as uHelp, would look like if it implemented our proposed vision. It shows how needs change over time, how people can discuss and agree on their technology's features and functionality (sometimes based on preferred values), and how the technology automatically adapts to users' evolving needs and values. Note that people are expected to discuss and agree in a natural language, and that the entire remaining process of how the technology evolves accordingly should be fully automated.

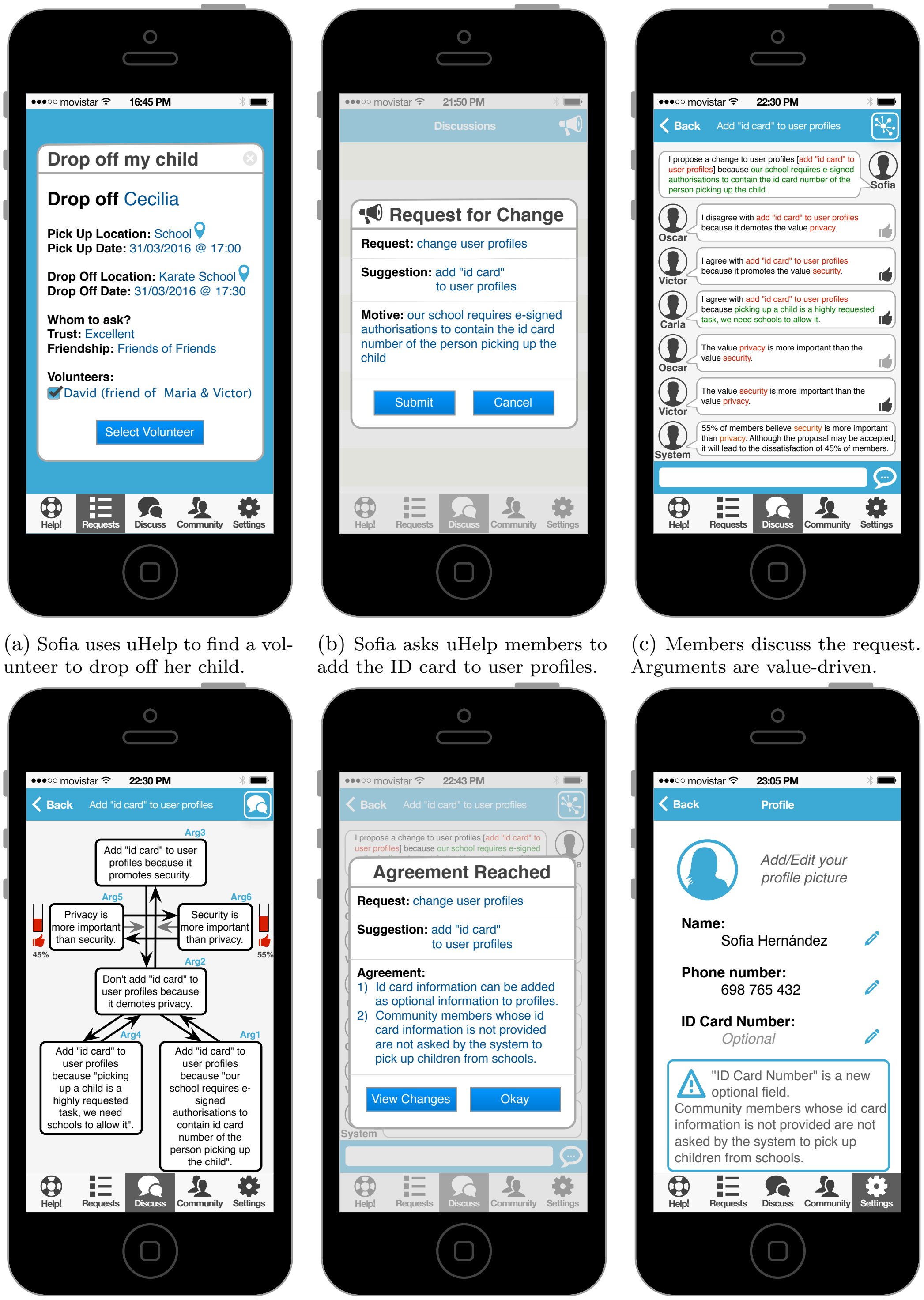
3 The How: Concept and Methodology

3.1 Concept: Norms for Fulfilling Needs & Adhering to Values

Behaviour is what ensures needs are fulfilled and values are adhered to, and **norms** are what govern behaviour. As such, our proposal for basing responsible technologies on needs and values is through norms, the rules of behaviour that help fulfil needs and values (Figure 2).

We envision norms to evolve with people's evolving needs and values, and we expect people to collaboratively agree on their preferred needs, values, and norms. For simplicity, we introduce

[8] A prototype is available on Google Play and Apple Store: www.uhelpapp.com



Drop off my child

Drop off Cecilia

Pick Up Location: School
Pick Up Date: 31/03/2016 @ 17:00

Drop Off Location: Karate School
Drop Off Date: 31/03/2016 @ 17:30

Whom to ask?
Trust: Excellent
Friendship: Friends of Friends

Volunteers:
☑ David (friend of Maria & Victor)

Select Volunteer

(a) Sofia uses uHelp to find a volunteer to drop off her child.

Discussions

Request for Change

Request: change user profiles

Suggestion: add "id card" to user profiles

Motive: our school requires e-signed authorisations to contain the id card number of the person picking up the child

Submit Cancel

(b) Sofia asks uHelp members to add the ID card to user profiles.

Add "id card" to user profiles

Sofia: I propose a change to user profiles [add "id card" to user profiles] because our school requires e-signed authorisations to contain the id card number of the person picking up the child.

Oscar: I disagree with add "id card" to user profiles because it demotes the value privacy.

Victor: I agree with add "id card" to user profiles because it promotes the value security.

Carla: I agree with add "id card" to user profiles because picking up a child is a highly requested task, we need schools to allow it.

Oscar: The value privacy is more important than the value security.

Victor: The value security is more important than the value privacy.

System: 55% of members believe security is more important than privacy. Although the proposal may be accepted, it will lead to the dissatisfaction of 45% of members.

(c) Members discuss the request. Arguments are value-driven.

Add "id card" to user profiles

Arg3: Add "id card" to user profiles because it promotes security.

Arg5: Privacy is more important than security.

Arg6: Security is more important than privacy.

Arg2: Don't add "id card" to user profiles because it demotes privacy.

Arg4: Add "id card" to user profiles because "picking up a child is a highly requested task, we need schools to allow it".

Arg1: Add "id card" to user profiles because "our school requires e-signed authorisations to contain id card number of the person picking up the child".

(d) Arguments are presented in a graph, along with voting results.

Agreement Reached

Request: change user profiles

Suggestion: add "id card" to user profiles

Agreement:
1) Id card information can be added as optional information to profiles.
2) Community members whose id card information is not provided are not asked by the system to pick up children from schools.

View Changes Okay

(e) After a final vote, an agreement is reached.

Profile

Add/Edit your profile picture

Name:
Sofia Hernández

Phone number:
698 765 432

ID Card Number:
Optional

"ID Card Number" is a new optional field.
Community members whose id card information is not provided are not asked by the system to pick up children from schools.

(f) The software and the GUI is automatically updated.

Figure 1: A motivating example: the uHelp app

the term ***ethical code*** to describe the bundle of needs, values and norms that drive and direct behaviour.[9] The life-cycle of a technology is then dictated by the life-cycle of its ethical code, which we divide into four stages (Figure 3): **first**, people discuss and agree on their ethical code; **second**, the agreed upon ethical code is *automatically* translated into a formal specification that can be analysed and reasoned upon (say to spot inconsistencies); **third**, the software that mediates behaviour is *automatically* modified to adopt the new ethical code; and **fourth**, the software enforces the agreed upon ethical code. The details of how these four stages are to be implemented, along with a sample of the *state of the art* in the field, is presented next.

3.2 Methodology: Natural Language Processing, Agreement Technologies, and Normative Systems

Stage 1. Agreement on the Ethical Code Our proposal's main objective is to give the novice person the means to be able to discuss and agree, along with others, on the needs, values, and norms (the ethical code) that should govern their technologies and their interactions. Discussion should be carried out in natural language (or a controlled natural language, as in [57], if the results of the machine translation needed for translating the ethical code from natural language into formal logic (stage 2 of the life-cycle) turns out to be unreliable). To achieve our goal, we suggest to make use of **agreement technologies**, as the key enabler of the people-driven ethical code evolution. **Learning mechanisms** can support this stage by learning when evolution should happen, which norms are best suited for this community, etc.

Research at this stage should then focus on two main issues.

1. *Build an agreement mechanism*, allowing people to discuss, argue, and agree on their ethical code. We suggest to build on agreement technologies, which have emerged as an imperative field in multi agent systems with the aim of helping individuals collaboratively reach a decision. The field is based on a number of models and mechanisms, such as argumentation and negotiation mechanisms, computational social choice, and trust and reputation models.

   In argumentation, argument schemes and critical questions (SchCQ) [53] have been used to provide templates for human authoring of natural language arguments and support computational identification of conflicts between those arguments without the need for natural language processing.[10] This line of work [50, 4] can be used and extended to

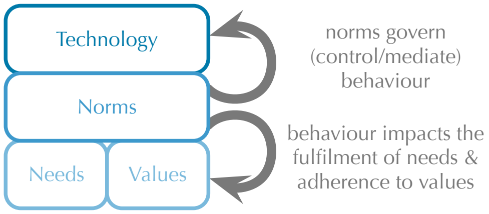

Technology
Norms
Needs
Values
norms govern (control/mediate) behaviour
behaviour impacts the fulfilment of needs & adherence to values

Figure 2: Norms for linking technology with needs and values

[9] A *broad* definition of ethical code refers to mission and values that underpin the code along with a code of conduct that regulates behaviour. We adopt this broad definition and say that an ethical code in our context represents needs, values, and norms (loosely corresponding to mission, values, and code of conduct, respectively).

[10] One suggestion would be for natural language processing to focus on translating the ethical code from natural language to formal logic (Stage 2 of the life-cycle), as the discussion phase itself is much more complex. As such, argument schemes and critical questions can be used at this stage.

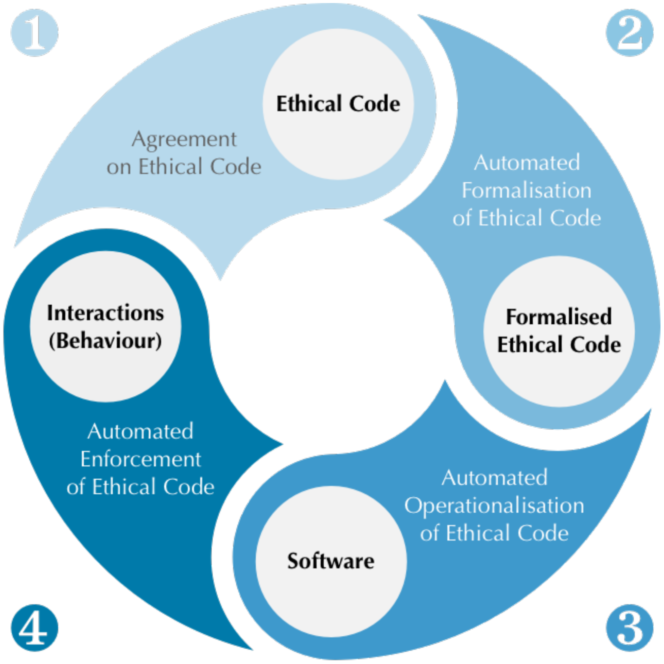

1
2
3
4
Ethical Code
Agreement on Ethical Code
Automated Formalisation of Ethical Code
Formalised Ethical Code
Automated Operationalisation of Ethical Code
Software
Automated Enforcement of Ethical Code
Interactions (Behaviour)

Figure 3: Technology's life-cycle

enable the discussion about: (a) whether, and to what extent, a particular norm fulfils a given need or promotes a particular value; (b) which needs and values are important to people; and (c) which norms should be adopted.

Argumentation mechanisms [7, 42, 2, 6, 35] can then be used to help assess the strength of arguments. Particularly interesting is the work that incorporates social voting to argumentation in social networks [30]. This work can then be extended to study the combination of trust and argumentation [8]. Work on trust and reputation [38, 27, 41] can also be adapted to help modify the strengths of arguments (to consider reputation/trust measures). Voting algorithms may also be adapted by incorporating trust, with votes potentially weighted by the trustworthiness of the voter [22, 10]. Though one model worth noting in this field is LiquidFeedback [5], which helps opinion formation and decision making online, and implements a delegated voting system. Some of its interesting properties is providing equal opportunities, transparency, and protection of minorities.

(2) *Build a learning mechanism* that can learn from past experiences and suggest changes accordingly. For example, learning and pattern recognition techniques [54] can help figure out when things are wrong and the ethical code needs to be revised (for example, when collaboration decreases). Alternative techniques to machine learning can also be useful here. For example, sentiment analysis can help pick up the dis/satisfaction of the community [29], and suggest whether the ethical code needs to be revisited accordingly.

Learning can also be useful to support the users' discussion phase by having the system automatically learn and suggest to users the best norms for their community. Existing norm synthesis techniques are one approach for pursuing this [47, 46, 36].

Additionally, learning the consequences of norms (say from the history of past experiences of either the same or similar community, where similarity measures [43] may be used to help decide which history of which community is worthwhile learning from) can be used either to (1) suggest norms by choosing those with the least undesired consequences, or even (2) provide arguments for the discussion between human members, where the system may join the discussion by supporting or attacking arguments. Again, 'learning' the consequences of norms may be achieved through alternative approaches, such as simulations [45], case-based reasoning [12], analogical reasoning [26] or coherence theory [49].

Stage 2. Automated Formalisation of the Ethical Code To help reason about the agreed upon ethical code, the code needs to be formalised. This step is critical, as we are letting *people* decide on their ethical code. As such, there needs to be a way to 1) ensure that the agreed upon code is coherent and consistent, and 2) explore its practical consequences and implications. This stage is concerned with **formalising the ethical code** and **reasoning about it**. For instance, detecting whether a new value or norm may become contradictory with the current norms. Such information would in fact feed back into the first stage, the agreement on an ethical code, where the humans should be informed that accepting a new norm may force the elimination of a previously accepted one or that a value may be jeopardised.

Research here should focus on two main issues.

(1) *Define a formal logic* for the specification of the ethical code, and *build a reasoner* for that logic that reasons over needs, values and norms, and their inconsistencies. The logic will be a novel need- and value-based logic for norms. Concerning the formal specification of needs, values, and norms (that is, the ethical code), we note that there is extensive work on the specification of norms in logics, but less on needs, and almost nothing on values. We can build here on the traditional approach of using deontic logic for specifying norms [1, 52]. Deontic logic is the logic of duties, and it deals with concepts like permissions, prohibitions, and obligations, which help specify who can do what, under what conditions, and so on. However, we will also need to introduce needs and values. Concerning the combination of needs and norms, we note that the most related research is that which deals with *goals*. Goals may be viewed as the needs that shall be fulfilled. Relevant work in this area is [19], which extends the BDI model of agents to include goals, obligations, and norms; the proposal is essentially based on providing a formal definitions of norms by means of some variation of deontic logic that includes conditional and temporal aspects. As for the combination of values and norms, we note that this is still uncharted territory. Some work related to values is [51], where an argumentation mechanism is designed to help argue whether actions promote or demote values. In [33], values are viewed, apart from being the "most fuzzy concepts of social sciences", as being a particular type of evaluation that affects "important choices and pursuits of an individual, interpersonal attraction and social exchanges, norms and standards of behaviour". In [37], the authors do suggest an engineering approach to the design of value-driven socio-cognitive systems. However, that proposal does not discuss how exactly these values are made operational. In [18], an initial attempt has been made linking values to norms and culture. Last, but not least, concerning the reasoner, we note that the literature is rich with systems implementing deontic logic and reasoning about norms [15, 19] that we can build upon. The results of reasoning (such as spotting conflicting norms/values) should then feed back into the first stage and influence the agreement on the ethical code.

(2) *Build an automatic translator* for translating the ethical code from natural language into the formal logic. This is a rather challenging task, though important research has been carried out in that direction, such as existing work on translating policy statements into first order logic [57], extracting deontic rules from regulations specified in natural language [56], or providing a logical representation of regulations [3].

One approach worth investigating is to base the translation (from natural language to the norm- and value-based normative logic) on exploring the connection between natural language features and the formal ones. For example, trying to label a statement whether it is an obligation, a permission, or a prohibition based on analysing its verb. For example, whether the verb contains 'must', 'shall', 'ought to', 'may', etc. Or trying to figure out who a norm addresses by looking for the subject in the sentence. Or trying to figure out the conditions by searching for conditional conjunctions, such as 'if', 'when', etc. And as illustrated above, if the results of natural language processing turn out to be unreliable, then a controlled natural language can be used, as in [57].

Stage 3. Automated Operationalisation of the Ethical Code Given a formal specification of needs, values and norms, the goal is to have an *executable code* (or software) that would make sure norms are followed, guaranteeing needs are fulfilled and values abode to. How to transform a formal ethical code written in some sort of logical representation into an executable code (or machine code) is a challenging task. To address this challenge, we suggest to build on **norm regimentation** mechanisms of multiagent systems, which implement a strict enforcement of norms, along with **formal verification mechanisms**, to ensure the translation into the executable code satisfies the requirements put forward by the formal one.

Research at this stage should focus on two main issues.

(1) *Build a norm regimentation mechanism* that automatically adapts the software mediating behaviour (or the executable code) so that the agreed upon ethical code is regimented (that is, *strictly enforced*). As illustrated above, we suggest building on existing norm regimentation mechanisms [21, 44]. It is stated that regimentation forces ideality (expressed as norms) and reality (defined by behaviour) to coincide [28]. The literature provides a variety of solutions on the regimentation of social norms [47], such as having contracts and commitments [20], electronic institutions [21], distributed dialogues [44]. These approaches are usually based on the idea of agents playing different roles and interacting by means of speech acts. Each role is defined by what actions agents can perform, when can they be performed, and under what conditions.

However, we note that the challenge at this stage will mostly be in the translation from a formal ethical code into an executable one. This is not a straight forward task, and the solution will depend on the choice of language decided upon (for the executable code). For example, the formal ethical code might be translated into protocols (or tiny programs), constraints, or even rules (such as 'if then' statements). To address this challenge, we suggest to get inspiration from the SIMPLE language [17], a language that may be viewed as a controlled natural language, a formal language, and an executable language, all in one. In other words, a language that can be used in all stages of the life-cycle (or at worst, requires a straightforward translation). We also note that although the chosen language should be simple enough, it should also be expressive enough to allow the specification of common needs, values, and norms.

(2) *Develop a verification mechanism* to verify the overall behaviour of the executable code. As we are automatically translating the formal ethical code into an executable one, it is crucial to verify that the overall behaviour of the executable code satisfies the requirements put forward by the formal one (such as satisfying the intended needs and adhering to the agreed upon values). We suggest building on previous work in model checking multiagent systems [55, 31, 9, 39, 40], adapting them to the languages of the formal and executable code (of Stages 2 and 3). The model checker of [39, 40] is particularly interesting as, unlike other approaches, it is efficient enough to execute at runtime. The main novelty of this research, however, will be in introducing needs and values as properties to be verified against. As such, we need to translate the formal logic that specifies needs and values into some modal logic so it can be verified by the model checker.

Stage 4. Automated Enforcement of the Ethical Code The basic idea of our proposal is that by executing the software of Stage 3, the software will ensure that norms are satisfied, values are adhered to, and needs are fulfilled. However, not all norms can be regimented, or strictly enforced. For example, while the system may prohibit a person from accessing some sensitive data, the system cannot forbid the person from using inappropriate language.[11] As such, aside from strictly enforcing regimented norms, we also need to deal with enforcing "un-regimented' norms. Alternative approaches that apply sanctions when violations occur or provide incentives to abide by the norms have been studied (such as decreasing one's trustworthiness if they miss a deadline) [13]. This is what the literature refers to as **norm enforcement**.

Research at this stage should focus on the following.

(1) *Build a norm enforcement engine* that checks at run-time the abidance to "un-regimented" norms and addresses violations accordingly, getting inspiration from existing work on norm enforcement [25, 34, 14], such as applying sanctions [23] or providing incentives for compliance [32]. Here, we suggest to advance existing work by introducing the notion of a *dynamic* norm enforcement engine, where the rules that address violations will be dependent on the ethical code itself. Mainly, just as people agree on the norms that govern behaviour, people can also agree on the norms that address violations. In a way, we empower people so they can decide for themselves on the appropriate 'punishment' for each violation. For instance, in the uHelp example of Figure 1, the punishment for being late to pickup one's child from school may be decided by community members to be more severe that the punishment for being late to deliver the groceries.

(2) *Develop a user friendly automated interface* that automatically adapts to the changing requirements imposed by the ethical code. Having the software automatically adapt with the evolving ethical code is not enough: the user interface should also adapt according to the changing requirements. For example, if the ID needs to be introduced to people's profiles (Figure 1), then the profile view should be adapted accordingly, and automatically. Here, we suggest to build upon modest and preliminary work on automated GUIs [16, 11]. We note, however, that although automated GUIs are possible, automating user friendly and intuitive GUIs is a challenge that needs to be addressed, as a good user interface is vital for the success of any technology.

3.3 Concluding Remarks: Novelty, Challenges, & Interdisciplinarity

As illustrated earlier, we do not claim to address the whole of ethics in AI, which is a wild beast to tame. What we do propose, however, is fundamental work that addresses some ethical

[11] Some even argue that norms *should not always* be regimented by the system, as we are autonomous beings, and controlling our every action will essentially strip us from our autonomy [13]. For example, it may be argued that user's autonomy is maintained when s/he is allowed to violate a norm.

considerations concerning today's technologies, and builds a novel foundation for future work on ethics in AI. Our proposed solution is a stepping stone for ethics in AI that widens the current discussion, which usually focuses on policies, education, or research culture, to embedding values (and possibly other ethical requirements) *within* technology through what we refer to as **ethics by construction**. While value-sensitive design [24] is well established in the social sciences (where the design of technology essentially accounts for human values, for both direct and indirect stakeholders), here, we take value-sensitive design a step further by allowing stakeholders to have a direct say in the technology's design and functionality, without being dependent on the middleman that liaises between them and technology designers and developers. Furthermore, this 'say' is not restricted to when the technology is first designed, but continues throughout the technology's lifetime. We essentially embed the ideas of value-sensitive design within the technology itself, which we refer to as 'ethics by construction'. The result is empowering people by giving them direct control over their technologies, and maintaining their control throughout the technology's lifetime by dictating and directing its evolution based on their evolving needs and values.

Though our proposal is not only novel in its objective (giving people control so their needs and values dictate technology's functionality and evolution), but in its methodology as well. Our proposal is novel in introducing *needs and values as the means that underpin ethical considerations*. The development of technology is based on people's needs and values. Last, but not least, technically speaking, the study of values is novel in the field of AI. There is some preliminary work in argumentation where a mechanism is designed to help argue whether actions promote or demote values [51]. However, how values may be specified, validated, and enforced is still uncharted territory.

As already illustrated by our methodology, our proposal builds on top of well established lines of research; mostly relying on natural language processing, agreement technologies, and normative systems. This helps support the feasibility of our proposed approach. But there remains a number of challenges to be addressed. First and foremost, the idea of fully automating stages 2–4 of the life-cycle is not straight forward. For example, what if the verification in stage 3 proves that some intended properties are not satisfied? Can we truly do without the human intervention in these three stages of the life-cycle? Our vision might be achievable in specific application domains, such as the online community of Figure 1, but achieving full automation in general is a challenging task.

Another main challenge of our suggested work will be building the languages and mechanisms with respect to the complete life-cycle. As illustrated earlier, there already exist many languages and mechanisms for each of the research lines presented, but these mechanisms have been designed under strict assumptions that limit their impact on or usefulness in real life scenarios. We want to develop languages and mechanisms that *work together* to help us close the life-cycle. We are inspired here by the SIMPLE language [17], a controlled natural language that is also a formal and executable language, which illustrates that one simple language may be possible to cover all three stages of the life-cycle.

Another challenge is our novel introduction of values to formal languages and mechanisms, which we suggest to address through close-knit collaboration with experts in ethics and value-sensitive design [48], along with cognitive scientists working on values [33].

As for the challenge in natural language processing, we have already proposed a backup plan that relies on a controlled natural language [57]. Though it must be noted that this solution does come at the cost of having users learn the limits of the controlled language, which might not be as easy as expected.

Last, but not least, no work on a topic as ethics in AI can succeed with a purely technological approach. As such, this work *must* be carried out in a close-knit collaboration with a **multidisciplinary team**, including experts in philosophy and ethics, cognitive science, and legal studies. For example, the input from philosophy and cognitive science on the values and norms dynamics can help us better understand how decisions on norms can be made taking into consideration preferences on values. The input from both philosophy and law can help us better understand the ethical and legal implications of allowing people to choose their ethical code, and how to avoid potential abuse. They can help us better understand what values may be voted upon and what values should be inscribed or un-negotiable (such as rejecting dictatorships or protecting the rights and interests of minorities in online communities). More importantly, while giving humans legislative powers would be a great achievement (stage 1 of the life-cycle), the consequences and implications of giving the machine the executive and judiciary powers (automating stages 2–4 of the life-cycle) will require a careful assessment.

4 Acknowledgments

This work is supported by the CIMBVAL project (funded by MINECO, project number TIN2017-89758-R) and the AppPhil project (funded by RecerCaixa 2017).

References

[1] Thomas Ågotnes, Wiebe van der Hoek, Juan A. Rodríguez-Aguilar, Carles Sierra, and Michael Wooldridge. On the logic of normative systems. In *Proc. of IJCAI '07*, pages 1175–1180, 2007.

[2] Leila Amgoud and Claudette Cayrol. A reasoning model based on the production of acceptable arguments. *Ann. Math. Artif. Intell.*, 34(1-3):197–215, 2002.

[3] Tara Athan, Harold Boley, Guido Governatori, Monica Palmirani, Adrian Paschke, and Adam Wyner. OASIS LegalRuleML. In *Proc. of ICAIL '13*, pages 3–12. ACM, 2013.

[4] Katie Atkinson, Trevor Bench-Capon, and Peter Mcburney. Computational representation of practical argument. *Synthese*, 152:157–206, 2006.

[5] Jan Behrens, Axel Kistner, Andreas Nitsche, and Björn Swierczek. *The principles of LiquidFeedback*. Interacktive Demokratie, 2014.

[6] Trevor Bench-Capon. Persuasion in practical argument using value-based argumentation frameworks. *J. Logic Comput.*, 13(3):429–448, 2003.

[7] Trevor Bench-Capon and Paul Dunne. Argumentation in artificial intelligence. *Artif. Intell.*, 171:619–641, 2007.

[8] Piero Bonatti, Eugenio Oliveira, Jordi Sabater-Mir, Carles Sierra, and Francesca Toni. On the integration of trust with negotiation, argumentation and semantics. *Knowl. Eng. Rev.*, 29:31–50, 2014.

[9] Rafael H Bordini, Michael Fisher, Carmen Pardavila, Willem Visser, and Michael Wooldridge. Model checking multi-agent programs with casp. In *Proc. of CAV '03*, pages 110–113. Springer, 2003.

[10] Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D. Procaccia, editors. *Handbook of Computational Social Choice*. Cambridge University Press, 2016.

[11] Ismel Brito, Nardine Osman, Jordi Sabater-Mir, and Carles Sierra. Charms: a charter management system. automating the integration of electronic institutions and humans. *Appl. Artif. Intell.*, 26(4):306–330, 2012.

[12] Jordi Campos, Maite López-Sánchez, and Marc Esteva. A case-based reasoning approach for norm adaptation. In Emilio Corchado, Manuel Graña Romay, and Alexandre Manhaes Savio, editors,

Hybrid Artificial Intelligence Systems, pages 168–176, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[13] Cristiano Castelfranchi. Formalising the informal?: Dynamic social order, bottom-up social control, and spontaneous normative relations. *J. Appl. Log.*, 1(1–2):47 – 92, 2003.

[14] Natalia Criado, Estefania Argente, Pablo Noriega, and Vicente J. Botti. MaNEA: A distributed architecture for enforcing norms in open MAS. *Eng. Appl. Artif. Intell.*, 26(1):76–95, 2013.

[15] Nicodemos C. Damianou, Arosha K. Bandara, Morris S. Sloman, and Emil C. Lupu. A survey of policy specification approaches. Technical report, Imperial College, London, UK, 2002.

[16] Dave de Jonge, Juan A. Rodriguez-Aguilar, Bruno Rosell, and Carles Sierra. Infrastructures to engineer open environments as electronic institutions. In *Proc. of E4MAS '14*, volume 9068 of *LNAI*. Springer, 2015.

[17] Dave de Jonge and Carles Sierra. SIMPLE: a language for the specification of protocols, similar to natural language. In *Proc. of COIN '15*, pages 49–64, 2015.

[18] Francien Dechesne, Gennaro di Tosto, Virginia Dignum, and Frank Dignum. No smoking here: values, norms and culture in multi-agent systems. *Artif. Intell. Law*, 21(1):79–107, 2013.

[19] Frank Dignum, David Kinny, and Liz Sonenberg. From desires, obligations and norms to goals. *Cognitive Science Quarterly*, 2:407–430, 2002.

[20] Virginia Dignum, John-Jules Meyer, and Hans Weigand. Towards an organizational model for agent societies using contracts. In *Proc. of AAMAS '02*, pages 694–695. ACM, 2002.

[21] Mark d'Inverno, Michael Luck, Pablo Noriega, Juan A. Rodriguez-Aguilar, and Carles Sierra. Communicating open systems. *Artif. Intell.*, 186:38–94, 2012.

[22] Ulle Endriss. Social choice theory as a foundation for multiagent systems. In *Proc. of MATES '14*, volume 8732 of *LNAI*, pages 1–6. Springer, 2014.

[23] Nicoletta Fornara and Marco Colombetti. Specifying and enforcing norms in artificial institutions. In *Proc. of DALT '09*, volume 5397 of *LNCS*, pages 1–17. Springer, 2009.

[24] Batya Friedman, Peter H. Kahn, and Alan Borning. Value sensitive design and information systems. In *Human-Computer Interaction and Management Information Systems: Foundations*, pages 348–372. M.E. Sharpe, 2006.

[25] Dorian Gaertner, Andres Garcia-Camino, Pablo Noriega, Juan-Antonio Rodriguez-Aguilar, and Wamberto Vasconcelos. Distributed norm management in regulated multiagent systems. In *Proc. of AAMAS '07*, pages 90:1–90:8. ACM, 2007.

[26] Dedre Gentner and Kenneth D. Forbus. Computational models of analogy. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3):266–276, 2011.

[27] Patricia Gutierrez, Nardine Osman, Carme Roig, and Carles Sierra. Trust-based community assessment. *Pattern Recognit. Lett.*, 67:49–58, 2015.

[28] Andrew J. I. Jones and Marek Sergot. On the characterization of law and computer systems: The normative systems perspective. In *Deontic Logic in Computer Science*, pages 275–307. John Wiley and Sons Ltd., 1993.

[29] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis: The good the bad and the OMG! In *Proc. of ICWSM '11*, pages 538–541. The AAAI Press, 2011.

[30] João Leite and João Martins. Social abstract argumentation. In *Proc. of IJCAI '11*, pages 2287–2292. AAAI Press, 2011.

[31] Alessio Lomuscio, Hongyang Qu, and Franco Raimondi. MCMAS: A model checker for the verification of multi-agent systems. In *Proc. of CAV '09*, pages 682–688. Springer, 2009.

[32] Henrique Lopes Cardoso, Ramón Hermoso, and Maria Fasli. Policies for role maintenance through incentives: How to keep agents on track. In *Agreement Technologies*, volume 8068 of *LNCS*, pages 150–164. Springer, 2013.

[33] Maria Miceli and Cristiano Castelfranchi. A cognitive approach to values. *J. Theory Soc. Behav.*, 19(2):169–193, 1989.

[34] Sanjay Modgil, Noura Faci, Felipe Meneguzzi, Nir Oren, Simon Miles, and Michael Luck. A framework for monitoring agent-based normative systems. In *Proc. of AAMAS '09*, pages 153–160. IFAAMAS, 2009.

[35] Sanjay Modgil and Henry Prakken. A general account of argumentation and preferences. *Artif. Intell.*, 195:361–397, 2013.

[36] Javier Morales, Maite López-Sánchez, and Marc Esteva. Using Experience to Generate New Regulations. In *Proc. of IJCAI '11*, pages 307–312, 2011.

[37] Pablo Noriega, Harko Verhagen, Mark d'Inverno, and Julian Padget. A manifesto for conscientious design of hybrid online social systems. In *Proc. of COIN '17*, pages 60–78. Springer, 2017.

[38] Nardine Osman, Patricia Gutierrez, and Carles Sierra. Trustworthy advice. *Knowledge-Based Syst.*, 82:41–59, 2015.

[39] Nardine Osman and David Robertson. Dynamic verification of trust in distributed open systems. In *Proc. of IJCAI '07*, pages 1440–1445, 2007.

[40] Nardine Osman, David Robertson, and Christopher Walton. Run-time model checking of interaction and deontic models for multi-agent systems. In *Proc. of AAMAS '06*, pages 238–240. ACM, 2006.

[41] Nardine Osman, Carles Sierra, Fiona McNeill, Juan Pane, and John K. Debenham. Trust and matching algorithms for selecting suitable agents. *ACM Trans. Intell. Syst. Technol.*, 5(1):16:1–16:39, 2013.

[42] Iyad Rahwan and Guillermo Simari, editors. *Argumentation in AI.* Springer, 2009.

[43] Philip Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.*, 11:95–130, 1999.

[44] David Robertson. A lightweight method for coordination of agent oriented web services. In *Proc. of AAAI Spring Symposium on Semantic Web Services*, 2004.

[45] Bastin Tony Roy Savarimuthu and Stephen Cranefield. Norm creation, spreading and emergence: A survey of simulation models of norms in multi-agent systems. *Multiagent Grid Syst.*, 7(1):21–54, January 2011.

[46] Sandip Sen and Stéphane Airiau. Emergence of norms through social learning. In *Proc. of IJCAI '07*, pages 1507–1512, 2007.

[47] Yoav Shoham and Moshe Tennenholtz. On social laws for artificial agent societies: off-line design. *Artif. Intell.*, 73:231–252, 1995.

[48] Judith Simon. Values in design. In *Handbuch Medien- und Informationsethik*, pages 357–364. J.B. Metzler, 2016.

[49] Paul Thagard. *Coherence in thought and action.* MIT Press, 2002.

[50] Pancho Tolchinsky, Sanjay Modgil, Katie Atkinson, Peter McBurney, and Ulises Cortes. Deliberation dialogues for reasoning about safety critical actions. *Auton. Agents Multi-Agent Syst.*, 25(2):209–259, 2012.

[51] Tom van der Weide, Frank Dignum, John-Jules Meyer, Hendrik Prakken, and Gerard Vreeswijk. Practical reasoning using values: Giving meaning to values. In *Proc. of ArgMAS '09*, pages 79–93. Springer, 2010.

[52] Javier Vázquez-Salceda, Virginia Dignum, and Frank Dignum. Organizing multiagent systems. *Auton. Agents Multi-Agent Syst.*, 11(3):307–360, 2005.

[53] Doug Walton. *Argument Schemes for Presumptive Reasoning.* Lawrence Erlbaum Assoc., 1996.

[54] Gary M. Weiss and Haym Hirsh. Learning to predict rare events in event sequences. In *Proc. of KDD '98*, pages 359–363. AAAI Press, 1998.

[55] Michael Wooldridge, Michael Fisher, Marc-Philippe Huget, and Simon Parsons. Model checking multi-agent systems with MABLE. In *Proc. of AAMAS '02*, pages 952–959. ACM, 2002.

[56] Adam Wyner and Wim Peters. On rule extraction from regulations. In *Proc. of JURIX '11*, volume 235 of *Frontiers in Artificial Intelligence and Applications*, pages 113–122. IOS Press, 2011.

[57] Adam Wyner, Tom van Engers, and Kiavash Bahreini. From policy-making statements to first-order logic. In *Proc. of EGOVIS '10*. Springer, 2010.