Ethical Requirements for AI Systems

Renata Guizzardi[1], Glenda Amaral[2], Giancarlo Guizzardi[2], and John Mylopoulos[3]

[1] NEMO/UFES, Espírito Santo (UFES), Brazil
rguizzardi@inf.ufes.br
[2] CORE/KRDB, Free University of Bozen-Bolzano, Bolzano, Italy
{gmouraamaral,giancarlo.guizzardi}@unibz.it
[3] University of Ottawa, Ottawa, Canada
jm@cs.toronto.edu

Abstract. AI systems that offer social services, such as healthcare services for patients, driving for travellers and war services for the military need to abide by ethical and professional principles and codes that apply for the services being offered. We propose to adopt Requirements Engineering (RE) techniques developed over decades for software systems in order to elicit and analyze ethical requirements to derive functional and quality requirements that together make the system-to-be compliant with ethical principles and codes. We illustrate our proposal by sketching the process of requirements elicitation and analysis for driverless cars.

Keywords: Requirements Engineering · Ethical Requirements · AI Systems.

1 Introduction

The advent of Artificial Intelligence (AI) technologies, including machine learning, computer vision and natural language processing, has made it possible to build autonomous cyber-physical systems (CPSs), systems consisting of software and physical components, for example robots. Some CPSs being developed, including driverless cars and autonomous weapons, have raised ethical questions and even calls for their banning altogether [2]. Since AI is often built to stand in situations where human decision-making would otherwise be required, a big aspect one takes into account in decision-making processes is one's own ethics. Thus, systems should likewise be built based on ethical principles. But ethical questions about CPSs that socially interact with humans are not limited to AI systems and apply to all CPSs, including car cruise control systems, drones and photo cameras. It seems that the publicity surrounding AI systems has focused the limelight on a neglected dark corner of Software Engineering (SE): *Ethical Requirements*.

Requirements Engineering (RE) is the area of research within SE concerned with the elicitation and analysis of requirements for a system-to-be (for our purposes, an AI system). Requirements are elicited from stakeholders: persons,

groups, or organizations that are actively involved in the design of the system-to-be, may be affected by its outcomes, or can influence its outcomes. Analysis of stakeholder requirements leads to a specification for the system-to-be, consisting of functional and quality constraints the system-to-be must satisfy in order to meet the needs of its stakeholders.

Ethical requirements are requirements for AI systems derived from ethical principles or ethical codes (norms). They are akin to *Legal Requirements* [8], i.e., requirements derived from laws and regulations[4]. We are interested in characterizing the sources of ethical requirements, ethical principles and ethical codes, also sketching a systematic process for deriving requirements from such sources. The AI systems built on the basis of our proposal are not ethical agents who can reason and act on the basis of ethical principles. Rather, they are software systems that have the functionality and qualities to meet ethical requirements, in addition to other requirements they are meant to fulfill. We illustrate our initial proposal with a case study involving a driverless car. The main thesis of this paper is that techniques developed in RE that have been practiced for decades can also be used for making AI systems compliant with ethical principles and codes.

Defining ethical requirements allows ethical issues to be considered from the beginning in the CPSs development process. Hence, first of all, developers and stakeholders (e.g. those paying for the development of the system or the actual users of the system) shall include these issues during requirements elicitation, aiming at achieving a consensual agreement in their regard. Moreover, during requirements validation activities, i.e., when it is time to evaluate if each requirement is met by the system, a focus on ethical aspects is assured.

The remainder of the paper is structured as follows. Section 2 introduces ethical principles and codes, while section 3 sketches a systematic process for identifying ethical requirements. By leveraging on this process, section 4 briefly discusses the case of driverless cars, discussing their compliance to ethical considerations. Section 5 discusses related work, and section 6 presents some final considerations.

2 Ethical Principles and Codes

Ethical principles are general principles of conduct towards others. For example, The European Commission's draft ethical guidelines for trustworthy AI [5] lists five such principles: *Autonomy* (respect for human dignity), *Beneficience* (doing good to others), *Nonmaleficence* (doing no harm to others), *Justice* (treating others fairly), *Explicability* (behaving transparently towards others). For example, from the Principle of Autonomy one may derive "Respect for a person's privacy", and from that an ethical requirement "Take a photo of someone only after her consent" for a phone camera. As another example, from Nonmaleficence, we

[4] But note, there are ethical requirements that are not legal, and legal ones that are not ethical.

may derive a functional requirement "Do not drive fast past a bystander" for a driverless car.

Ethical principles are generally domain-independent and rather abstract, so they require some analysis to fit them to a particular domain so as to derive ethical requirements. Ethical codes specialize ethical principles into particular domains, such as codes of conduct for employees of an organization, and codes of professional conduct for members of a professional society. The medical profession has adopted elaborate rules for an ethical code of practicing doctors, and so have research organizations for the conduct of research. There are codes of conduct for the military, by national jurisdiction, and numerous ethical codes for drivers in regional or municipal jurisdictions depending on driver responsibilities (such as taxi/track/school bus driving). Notably, Germany is the first country to adopt an ethical code for driverless cars [10]. Finally, and perhaps most importantly for autonomous weapons, there are international conventions for the conduct of war, the use of weapons, the treatment of civilians and prisoners, etc.

3 Deriving Ethical Requirements

The key concept to deriving ethical requirements is that of *Runtime Stakeholders*. These include those stakeholders that are using, affected by, or influencing the outcomes of a system as it is operating. Traditional RE often limits runtime stakeholders to just users of the system-to-be. However, for AI systems this needs to be extended to other parties. For example, for a driverless car, runtime stakeholders include passengers – i.e., the users of the car – but also pedestrians, whose path may cross that of the car and shouldn't be hit; bystanders, who shouldn't be scared or splashed as the car drives by; nearby drivers, who as a courtesy, should be allowed to cut in front in the car's lane; and fellow drivers in general, who might benefit from information about an accident that just happened in the vicinity of the car.

Runtime stakeholders are often ignored in classical RE as they are perceived to lack a concrete "stake" in the system-to-be. But the intrusion of AI systems in social settings is dictating a shift in the theory and practice of RE to include also these somehow indirect stakeholders into the RE process. Considerations such as the examples given above may seem trivial in the dawn of a new technological era. But they aren't! Think of ten thousand driverless cars added to a local setting, say Ottawa (population approximately 1,000,000), who are aggressive and inconsiderate in their driving in the sense that they don't fulfill simple ethical requirements, such as the above. Wouldn't this constitute an act of maleficience towards local drivers and pedestrians alike? Manufacturers of driverless cars should produce cars that can do more than meet legal, safety, security and other requirements: the cars they produce must be *good* drivers. And what constitutes good driving is defined in terms of ethical requirements, to be derived from ethical principles and codes.

We could categorize Ethical Requirements for an artificial system as types of *Ecological Requirements*, in the sense that they are necessarily requirements

that are derived from the whole ecosystem in which the system is included. From an ontological perspective, there is a fundamental reason why this is the case, namely, given that these requirements are derived from assessments of *value* and *risk*. In a nutshell, value can be seen as a relational property, emerging from a set of relations between the intrinsic properties of a *value object*(or a value experience) and the goals of a *Value Subject* [9]. Roughly speaking, the value of an object (or experience) amounts to the degree to which the properties (*affordances*) of that object positively contribute (help, make) to the achievement of the value subject goals. Mutatis Mutandis, risk can also be seen as a relational property, emerging from a set of relations between the intrinsic properties of an *Object-at-Risk* (vulnerabilities), as well as *Threat Objects* and *Risk Enablers* (capacities, intentions) and the goals of a *Risk Subject* [9][5]. Again, roughly speaking, the risk of an object-at-risk given threat objects and risk enablers amounts to the degree to which the properties of those entities can be enacted to negatively contribute to denting (hurt, break) the risk subject goals. Now, ontologically speaking, affordances, vulnerabilities, capacities, intentions are all types of *dispositions*, which are themselves ecological properties, i.e., those that essentially depend on their environment (context) for their manifestation [9].

For example, given that we (as a society) value life, we would of course like to reduce as much as possible the risk of serious accidents with threats to human life (humans being the object at risk). For this, we must both consider the vulnerabilities of cars and their passengers, as well as the possible threats posed by other entities (e.g., other cars, road conditions). We must also endow driveless cars with a number of security features, but we must also do that for the entire platform in which driveless cars operate, including the consideration of features for roads, coordination points (the digital equivalent of traffic lights and road signs).

Given a set of runtime stakeholder types with their associated value and risk assessments, the next step is to introduce functional requirements that ensure that the car-to-be can actually recognize with adequate accuracy when it encounters instances of each type, under different weather and lighting conditions. In addition, we need functional requirements for recognizing notable events in the traffic environment of a car, such as accidents, slow/fast/very fast moving vehicles. Reports from different driverless car projects suggest that this is a step that has been recognized and adopted by driverless car manufacturers. Ethical requirements are functional and quality requirements elicited from runtime stakeholders in accordance with the five ethical principles discussed above.

4 The Case of Driveless Cars

We can now conduct an analysis of how to apply ethical principles, such as those listed above, to the case-at-hand. Explicability towards passengers may lead to a functional requirement for the driverless car to engage in conversations to

[5] In [9], the focus is on *use value* as opposed to *ethical value*. However, we believe the analysis still holds, in particular, regarding the connection between value and risk.

explain the route it is following and why. Explicability towards nearby drivers, pedestrians and bystanders leads to a functional requirement for the car to signal on turns and changes of lane. Explicability towards society in general benefits from the type of analysis aforementioned in which requirements can be traced back to the explicit identification of stakeholders, and an explicit and semantically transparent analysis of their values and risk[6]. Respect for human dignity calls for the car to stop in case it encounters a runtime stakeholder in need of assistance. Beneficience calls for the car to let a nearby driver cut in front, also to notify traffic authorities of an accident. Nonmaleficence calls for the car to slow down in the presence of nearby pedestrians and bystanders, independently of any speed limits that might apply. And in the case of two lanes merging into one, Justice calls for treating drivers from the other lane fairly, rather than in a me-first manner.

This analysis can be made more concrete and guided if it is based on an ethical code that applies for the system-to-be. Firstly, ethical codes often identify some of the runtime stakeholders, also include concrete applications of ethical principles that make the derivation of ethical requirements more direct and less controversial.

5 Related Work

In [6], the authors offer an excellent discussion on the incorporation of ethics into AI systems in the context of driverless cars. Two approaches are considered: (a) Make the AI system an ethical agent who can reason top-down from first principles to an ethical problem-at-hand and choose a suitable action; (b) Have the AI system learn bottom-up the most suitable ethical choice in different circumstances. Both alternatives are found to be problematic and both assume that for an AI system to comply with ethical principles or codes, it must be capable of reasoning on its own about the ethical merits of alternative decisions.

The US Department of Defense directive on autonomous weapon system [1] adopts a human-in-the-loop approach to such weapons. It also proposes policies that emphasize thorough testing and Verification & Validation for all semi-autonomous weapons to ensure that they function as designed. Arkin [4] discusses the merits and pitfalls of autonomous weapons, emphasizing that they could end up saving civilian lives. On the other hand, O'Connell [7] considers the politics of banning autonomous killing altogether.

[6] Notice that transparency w.r.t. the entities that compose an ecosystem regarding their capabilities, intentions, vulnerabilities, and goals strongly connects also to the notion of *trust*. In a nutshell, trust amounts to a set of relations connecting the beliefs of a (trustor) agent regarding the capabilities, vulnerabilities and intentions of a trustee insomuch as they can affect that agent's goals [3]. From this we directly have that: (1) trustworthiness assessment can and should be grounded in the explicit assessment of these aspects; (2) trustworthiness is not an absolute property of a system, but one that depends on all these aspects. To put it bluntly, it is meaningless to speak of trustworthy systems in an unqualified manner.

6 Conclusions

We have argued that RE techniques can be applied in the design of AI systems, such as driverless cars and autonomous weapons, to ensure that they comply to ethical principles and codes. It is important to emphasize that the solution we propose doesn't render such systems autonomous in ethical decision-making, since ethical matters are dealt with by their designers and built into the systems. Our proposal, however, does suggest a way to go forward with AI systems where technology is available, but we don't know how to deal with the ethical implications of their outcomes.

As to the implementation of functional and quality requirements derived from ethical requirements, it is important to emphasize that the system-to-be should be able to perform as well as well-trained humans performing the same task. For instance, similarly to a medical doctor, who writes a detailed report explaining her findings, AI systems should explain their reasoning rather than only providing results and taking decisions. This has important implications because some of the most successful AI technologies, notably Machine Learning ones, cannot currently deal well with explainability and other transparency-related requirements.

Acknowledgments

This research is supported by the Strategic Partnership Grant "Middleware Framework and Programming Infrastructure for IoT Services".

References

1. Autonomy in weapon system, dod directive. Tech. rep., Department of Defence (2012), https://tinyurl.com/vh2qhej
2. Artificial intelligence: Mankind's last invention (2019), https://tinyurl.com/y9moo26c
3. Amaral, G., Sales, T.P., Guizzardi, G., Porello, D.: Towards a reference ontology of trust. In: Proc. Coopis 2019. Springer (2019)
4. Arkin, R.: Lethal Autonomous Systems and the Plight of the Non-Combatant. AISB Quarterly (137), 1–9 (July 2013)
5. High-Level Expert Group on Artificial Intelligence, E.C.: Draft ethics guidelines for trustworthy ai, draft document (2018)
6. Etzioni, A., Etzioni, O.: Incorporating Ethics into Artificial Intelligence. Journal of Ethics **21**(4), 403–418 (December 2017)
7. O'Connell, M.E.: Banning Auttonomous Killing. In: Evangelista, M., Shue, H. (eds.) The American Way of Bombing: How Legal and Ethical Norms Change. Cornell University Press (2013)
8. Otto, P.N., Antón, A.I.: Addressing legal requirements in requirements engineering. In: Proc.15th IEEE RE 2007, October 15-19th, 2007, New Delhi. pp. 5–14 (2007)
9. Sales, T.e.a.: The common ontology of value and risk. In: Proc. ER 2018. pp. 121–135. Springer (2018)
10. Tuffley, D.: At last! the world's first ethical guidelines for driverless cars. The Conversation (September 2017), https://tinyurl.com/u4gbskh