Towards Requirements Engineering for Superintelligence Safety

Hermann Kaindl
ICT, TU Wien
Vienna, Austria
hermann.kaindl@tuwien.ac.at

Jonas Ferdigg
ICT, TU Wien
Vienna, Austria
e1226597@student.tuwien.ac.at

Abstract

Under the headline "AI safety", a wide-reaching issue is being discussed, whether in the future some "superhuman artificial intelligence" / "superintelligence" could pose a threat to humanity. In addition, the late Steven Hawking warned that the rise of robots may be disastrous for mankind. A major concern is that even benevolent superhuman *artificial intelligence* (AI) may become seriously harmful if its given goals are not exactly aligned with ours, or if we cannot specify precisely its objective function. Both the definition and communication of such goals have to be conducted with extreme caution. Metaphorically, this is compared to king Midas in Greek mythology, who expressed the wish that everything he touched should turn to gold, but obviously this wish was not specified precisely enough. In our view, this sounds like requirements problems and the challenge of their precise formulation. As usual in requirements engineering (RE), ambiguity or incompleteness may cause problems. That is why we actually propose (and partly work on already) yet another RE problem, figuring out the wishes and the needs with regard to a superintelligence, which will in our opinion most likely be a very complex software-intensive system based on AI. To our best knowledge, this view of "AI safety" has not been pointed out yet. Requirements engineering appears to be the right approach to address it first, preferably before serious attempts of building "superhuman artificial intelligence" are being conducted.

1 Introduction

The idea of technology becoming sentient has been a common theme in the literature and movies for decades. One might think of the famous movie "2001: A Space Odyssey" by Stanley Kubrick, the Matrix, or the Terminator movies. These stories seem to be mostly consistent in the impression that a suddenly arising uncontrolled Artificial Intelligence (AI) will not mean us well. Furthermore, the AIs described in books and movies are never dumb machines, but rather potent entities with cognitive superpowers far beyond the capacities of human general intelligence — they are *superintelligent*. In his book "Superintelligence" [Bos14], Oxford philosophy professor Nick Bostrom provides good reasons to believe that we can create such an entity, and he is not the only one (see, e.g., [Yam15]).

In: M. Sabetzadeh, A. Vogelsang, S. Abualhaija, M. Borg, F. Dalpiaz, M. Daneva, N. Fernández, X. Franch, D. Fucci, V. Gervasi, E. Groen, R. Guizzardi, A. Herrmann, J. Horkoff, L. Mich, A. Perini, A. Susi (eds.): Joint Proceedings of REFSQ-2020 Workshops, Doctoral Symposium, Live Studies Track, and Poster Track, Pisa, Italy, 24-03-2020, published at http://ceur-ws.org

In the AI research priorities document published in [RDT15], apart from the desirability of safety, e.g., of self-driving cars, "AI safety" is addressed with regard to "superhuman AI". This is partly based on forecasting work on intelligence explosion and "superintelligence" [Bos14, ELH18], where the importance of given goals to be exactly aligned with those of mankind is emphasized. Even that may be insufficient, however, the system must also somehow be deliberately constructed to pursue them [Bos14].

In terms of requests like "Make paperclips", such a system is envisaged to take everything it can (with high "intelligence") and to make as many paperclips as it can make, whatever it may cost. Much like the example of king Midas or the core theme of the movie with the title "Bedazzled", we view this as a requirement that is not specified precisely enough. While other references regarding "AI safety" can be found in [ELH18] and at `https://vkrakovna.wordpress.com/ai-safety-resources/`, it is mentioned nowhere that it is actually a problem with *requirements*.

In addition to specifying and communicating requirements to a superintelligence, once available, we see a more pressing problem before it may be built. Since mankind may eventually create a "superintelligence", it should rather sooner than later specify the requirements on such a system. Hence, we propose RE on a superintelligence here.

Having a *glass-box view* on a system implementing a superintelligence may be helpful for keeping control of it, such as following an approach to making architectural decisions upfront, e.g., for a generic architecture [MLK+00]. Another approach is to model safety *frameworks*, see, e.g., [EKKL19], where deep neural networks are components of a framework. In contrast, we take a *black-box view* here and propose to focus on requirements.

Strictly speaking, the notion "AI safety" is misleading in our opinion, since it may also include the safety of certain AI-based systems, e.g., self-driving cars. Hence, we prefer the notion of "superintelligence safety". We also prefer this notion over "AGI safety", since the issue is not so much about an "Artificial General Intelligence", but a "superintelligence" based on AI, where an intelligence explosion may arise [Bos14].

The remainder of this paper is organized in the following manner. First, we provide some background material on superintelligence safety, in order to explain the challenge involved. Then we view superintelligence safety from the perspective of RE, and envisage specifying and communicating requirements accordingly. Finally, we propose RE on a superintelligence in the first place.

2 The Challenge of Superintelligence Safety

First, let us be more precise on what "superintelligence safety" really means. This concept includes minimizing the existential risk to humanity posed by superintelligent agents as well as mitigating unwanted social consequences that may arise even if the existential risk has been averted, such as social manipulation, advanced warfare, AI-driven unemployment, status quo preservation or unfair redistribution of resources. In a more technical sense, this means solving the "Control Problem" [Bos14] and designing and incorporating safety mechanisms in a seed AI which are still functional after arbitrarily many iterations of self-improvement by the system. "Ideally, every generation of self-improving system should be able to produce a verifiable proof of its safety for external examination" [Yam15, p.138].

Superintelligence safety also does not only apply to superintelligent systems to-be-built but also extends to the development process, which in itself poses an existential risk. In this context, Yampolskiy proposes the setup of AI research review boards consisting of a team of experts which are to evaluate each research proposal and decide if it can potentially lead to the development of a full-blown AGI [Yam15, p.139]. RE may be applied to the development process, the system to-be-built, to the goal(s) the system will be given and maybe also to the process by which the goal(s) will be selected.

Assuming the creation of a superintelligence is possible, how can we control it and avoid the doomsday scenarios so eloquently depicted by authors and film directors? While an emerging superintelligence might not be inherently malevolent, it still poses a great risk. Without any information about the internal workings of such an entity, we have to assume that its way of 'thinking' might be very different from the way a human thinks. A superintelligence most likely will not share human morals or have a concept of value at all, if the developers did not intentionally design it to have one.

While the most terrifying scenario is a malevolent superintelligence with the intent to eradicate humanity, another potential scenario is a superintelligence with no inherent motivations at all, simply following the instructions of it's creator to the best of its abilities. Bostrom describes some of the abilities a superintelligence could have under the term "Cognitive Superpowers" [Bos14, p. 91]. A superintelligence could excel at strategic planning, forecasting, analysis for optimizing chances of achieving distant goals, social and psychological mod-

eling and manipulation, rhetoric persuasion, hacking into computer systems, design and modeling of advanced technologies, and the list goes on [Bos14, p. 94].

One can see that even equipped with only a subset of these skills, a superintelligence pursuing a goal *to the best of its abilities* and without being constrained by concepts like morality or value, can cause substantial damage to its environment. Bostrom depicts this with his famous thought experiment where a superintelligence is given the simple task to make as many paper clips as possible. While the superintelligence might start off by acquiring monetary resources by predicting the stock market and building paper clip factories, it may not just stop there. Since its goal is the unconstrained maximization of the production of paperclips, it will soon discover that humans pose a hindrance to its endeavor, as we might try to stop the superintelligence from producing more paperclips or even try to shut it off. Since the superintelligence is multiple magnitudes smarter than humanity combined, humanity fails to stop the superintelligence and ultimately the whole world (the whole observable universe, in fact) will be made into paperclips.

Even if this was just a contrived example, one can see the risks of giving a superintelligence an unconstrained optimization problem. This argument is also extended in [Bos14] to constrained optimization problems, but the details are not necessary for our paper.

3 Specifying and Communicating Requirements to a Superintelligence

Under the (usually unrealistic) assumption that we knew the requirements already, 'just' telling the superintelligence about them properly is the problem here. It can be decomposed into specifying and communicating.

According to the Standard ISO/IEC 10746-2:2009 Information technology – Open Distributed Processing – Reference Model: Foundations, 7.4, a *specification* is a "concrete representation of a model in some notation". In order to better understand this sub-problem of specifying requirements, let us follow the observation in [KS10] that requirements *representations* are often confused with requirements *per se*. This confusion is also widespread in practice, as exemplified in the very recent Standard ISO SE Vocabulary 24765:2017. In fact, it defines a requirement both as a "statement that translates or expresses a need and its associated constraints and conditions" and "a condition or capability that must be present in a product ..." (dots inserted).

As it is well-known in RE, specifying requirements can be done informally, say, using some natural language, or formally, e.g., using some formal logic as the notation. Possibilities for semi-formal representations within this spectrum are many-fold. With regard to specifying requirements to a superintelligence, especially

- *ambiguity* and
- *incompleteness*

appear to be major concerns, cf. the paperclip example and king Midas.

In order to avoid *ambiguity* in the course of specifying requirements for a superintelligence, *formal* representation of the requirements should be the choice, of course, possibly in some formal logic. *Grounding the logic* used in the domain will, however, still leave loopholes for the superintelligence to 'misunderstand' the given specification of our requirements. For example, for some predicate on paperclips, a grounding of what paperclips are in the real world will be necessary.

Also *incompleteness* of the specification remains an open issue. Unfortunately, no general solution appears to be feasible at the current state of the art in RE.

Communicating a given requirements specification to a superintelligence may be investigated according to [JMF08], based on *speech-act theory* [Sea69], under the premise that stakeholders communicate information in the course of requirements engineering. This view was not taken for the sake of really communicating requirements in [JMF08], but for using the semantics of various speech acts to theoretically (re-)define the *requirements problem* originally defined by Zave and Jackson [ZJ97] (see also below). Everything is assumed to be communicated by speech acts here. However, we can only communicate *representations* of requirements, and not requirements *per se*. It is clear that the text given in the examples in [JMF08] *represents* something by describing it, the very confusion addressed in [KS10], where specific examples of this confusion are given.

Anyway, for communicating specifications of requirements to a superintelligence, speech-act theory could be helpful for annotating the specific kind of speech act, e.g., *Question* or *Request*. For example, the text "Can you produce paperclips?" may be interpreted either way, unless specified more precisely.

Still, we have to face that 'perfectly' specifying and communicating requirements to a superintelligence may not be possible. Even if we could, this would not help in case of malevolent superintelligence that would not

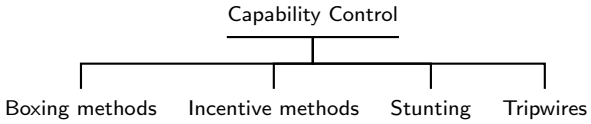

Figure 1: Overview of capability control techniques (based on [Bos14]).

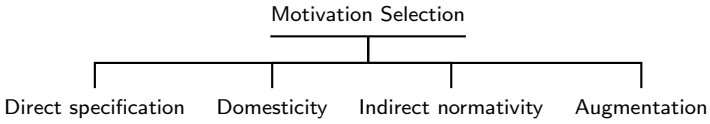

Figure 2: Overview of motivation selection techniques (based on [Bos14]).

necessarily satisfy requirements as specified and communicated. Hence, let us consider building a superintelligence in such a way that these problems can be mitigated.

4 RE on a Superintelligence

In our view, building a superintelligence should certainly include RE (much as building any non-trivial system). For specifically addressing superintelligence *safety*, we may ignore defining *functional requirements* on its superintelligent abilities. Still, some functionality may have to be defined for operationalizing certain "non-functional" safety requirements, which will in the first place be cast as *constraints* on the superintelligence.

Note, that such requirements can and should be defined, elaborated and implemented for AI-based systems already before these may actually become superintelligent. This may happen all of a sudden, and then it may be too late.

Working on these requirements could be done as usual according to best practice of RE. This will involve major stakeholders (in particular, AI researchers), requirements elicitation may be done including questionnaires in the Web, etc. While the core questions will be about requirements on superintelligence safety, of course, a bigger question should be asked in our opinion, what the goals and requirements are regarding AI in the first place. As its stands, most AI system are created and deployed without doing any RE at all.

We expect that RE for *superintelligence safety* will pose even greater challenges than the usual practice of RE. After all, it is not just about conceiving such a superintelligence but about making sure that its creation will not raise uncontrollable safety risks. The currently most widely used approach to *functional safety* of software-intensive systems [SS11, VCMG18, LL03, HKR+16] will not be sufficient. It is primarily concerned with hazards in the environment of the system and related safety risks, and as dealt with, e.g., in the current automotive standard ISO 26262, it focuses on failures of functions and how to manage them. Superintelligence safety will have to make sure that the superintelligent system will not create hazards even if none of its functions has a failure. This will have to be dealt with in a new kind of *safety requirements*.

Let our RE endeavor be also informed by some preliminary thought by the author who raised the issue of superintelligence safety. In [Bos14, Chapter 9], *controlling* a superintelligence is discussed, in order to deal with this specific safety problem. Two broad classes of potential methods are distinguished — *capability control* and *motivation selection*. Within each of them, several specific techniques are examined, see Figures 1 and 2 for overviews.

Viewed from an RE perspective, it seems as though most of what is discussed here could be elaborated as *constraint requirements*. Both current theory and practice of RE will most likely be sufficient for dealing with *capability control* as laid out here.

The ideas of *motivation selection* will be much harder to elaborate according to the current state of the art in RE than the ideas on capability control above. This may even involve extending the current theoretical formulation of the RE problem.

5 Conclusion and Future Work

In this paper, we address the potentially very important issue of "AI safety", in the sense of *superintelligence safety* [Bos14], from an RE perspective. To our best knowledge, this is the first RE approach to this issue, although it may seem obvious that RE is the very discipline of choice here (for someone being aware of RE).

We actually distinguish two different approaches:

- Specifying and communicating requirements for specific problems to a concrete superintelligence, and
- Doing RE in the course of building a superintelligence in the sense of an AI-based software-intensive system.

Based on common wisdom of RE at the current state of the art, we tentatively conclude that 'perfectly' specifying and communicating requirements to a superintelligence may not be possible. And even if this became possible in the future, it would not help in case of malevolent superintelligence that would not necessarily satisfy requirements as specified and communicated.

Hence, we already pursue RE on a superintelligence to be built. This RE endeavor is informed by some preliminary thought on *controlling* a superintelligence in [Bos14]. The first approach through *capability control* may be dealt with properly through *constraint requirements*. The second approach for controlling through *motivation selection*, however, appears to go beyond the current theory of RE. In particular, we raise the challenge of extending Goal Oriented Requirements Engineering (GORE) with *dynamic goals of a superintelligence*. In [KF19], we motivate how the theory of the *requirements problem* should be extended to cover goals of a superintelligence.

More specifically, at the time of this writing we are already preparing for gathering requirements on an AI-based superintelligence through Web-questionnaires from a wide variety of stakeholders from different business sectors, age groups and educational backgrounds. These should contribute to a better understanding of the wishes and (perceived) needs of different stakeholders with regard to a superintelligence, especially related to safety of humans or even mankind.

Our road map includes

- Analyzing the filled-in questionnaires;
- Developing a first baseline of a requirements specification with a focus on safety;
- Distributing this document to the ones having filled-in the questionnaires for receiving further feedback; and
- Integrating this feedback into a second baseline of a requirements specification on superintelligence safety.

In the long run, this may lead to a more elaborate approach along the lines of the moral machine experiment [ADK+18].

References

[ADK+18] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-Francois Bonnefon, and Iyad Rahwan. The Moral Machine experiment. *Nature*, 563:59–64, Nov 2018.

[Bos14] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, 2014.

[EKKL19] Tom Everitt, Ramana Kumar, Victoria Krakovna, and Shane Legg. Modeling AGI safety frameworks with causal influence diagrams. In *Proceedings of the Workshop on Artificial Intelligence Safety 2019*, volume Vol-2419. CEUR Workshop Proceedings, 2019.

[ELH18] Tom Everitt, Gary Lea, and Marcus Hutter. AGI safety literature review. Technical Report arXiv:1805.01109 [cs.AI], arXiv, 2018.

[HKR+16] B. Hulin, H. Kaindl, T. Rathfux, R. Popp, E. Arnautovic, and R. Beckert. Towards a common safety ontology for automobiles and railway vehicles. In *2016 12th European Dependable Computing Conference (EDCC)*, pages 189–192, Sep. 2016.

[JMF08] I. Jureta, J. Mylopoulos, and S. Faulkner. Revisiting the core ontology and problem in requirements engineering. In *2008 16th IEEE International Requirements Engineering Conference*, pages 71–80, Sep. 2008.

[KF19] Hermann Kaindl and Jonas Ferdigg. Superintelligence safety: A requirements engineering perspective. Technical Report arXiv:1909.12152 [cs.AI], arXiv, 2019.

[KS10] Hermann Kaindl and Davor Svetinovic. On confusion between requirements and their representations. *Requirements Engineering*, 15(3):307–311, Sep 2010.

[LL03] Axel Van Lamsweerde and Emmanuel Letier. From object orientation to goal orientation: A paradigm shift for requirements engineering. In *Radical Innovations of Software & System Engineering, Montery'02 Workshop, Venice(Italy), LNCS*, pages 4–8. Springer-Verlag, 2003.

[MLK+00] Mike Mannion, Oliver Lewis, Hermann Kaindl, Gianluca Montroni, and Joe Wheadon. Representing requirements on generic software in an application family model. In William B. Frakes, editor, *Software Reuse: Advances in Software Reusability*, pages 153–169, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.

[RDT15] Stuart Russell, Daniel Dewey, and Max Tegmark. Research priorities for robust and beneficial artificial intelligence. *AI Magazine*, 36(4), 2015.

[Sea69] John R. Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, England, 1969.

[SS11] David J. Smith and Kenneth G. L. Simpson. *Safety Critical Systems Handbook: A Straightfoward Guide to Functional Safety, IEC 61508 (2010 Edition) and Related Standards, Including Process IEC 61511 and Machinery IEC 62061 AND ISO 13849*. Elsevier, 2011.

[VCMG18] J. Vilela, J. Castro, L. E. G. Martins, and T. Gorschek. Assessment of safety processes in requirements engineering. In *2018 IEEE 26th International Requirements Engineering Conference (RE)*, pages 358–363, Aug 2018.

[Yam15] Roman V. Yampolskiy. *Artificial Superintelligence: A Futuristic Approach*. CRC Press, 2015.

[ZJ97] Pamela Zave and Michael Jackson. Four dark corners of requirements engineering. *ACM Trans. Softw. Eng. Methodol.*, 6(1):1–30, January 1997.