Utrecht University
Graduate School of Natural Sciences
Philosophy of AI

Janneke van Lith, Natasha Alechina, Sven Nyholm, Mehdi Dastani,
Richard Starmans, Jan Broersen, Rachel Boddy

Semester II

Rethinking AI Responsibility
From Fixed Objectives to Learning Preferences

Otto Mättas
Artificial Intelligence
Year I

o.mattas@students.uu.nl
10/04/2020
Student ID: 6324363

My thesis is that the responsibility gap observed in connection to Artificial Intelligence systems can be bridged by rethinking how we build those systems to achieve goals, namely replacing fixed objectives with learning preferences.

By the responsibility gap, I am referring to the complexity of distributing responsibility and liability, when AI systems are effectively taking part in real-world decision making processes with having consequences. One can think of self-driving cars for example. If we allow AI systems to make decisions that have a direct impact on human life and health, we should be able to ask - who is responsible for the decisions the AI system makes? Right now, there is no direct answer.

Many if not most such modern systems are built with a clear objective in mind. One can think of famous examples like the IBM-developed chess-playing system Deep Blue which beat Garry Gasparov, a reigning world champion in chess in 1997. Another more recent example is from 2015, when Google-backed system AlphaGo beat Lee Sedol, a top-ranking professional player in the game of Go. Those systems were specifically built to achieve one fixed objective - beat your opponent in a game with clearly defined rules and expected outcomes.

Fixed objectives are introduced to the system most often by its developer, who follows their rationale. This approach does not leave any freedom for the goal to change or be changed as the environment changes. After the system is turned on, it optimises for achieving its goal. Systems tied to closed environments like a chess or Go board have exhibited amazing results in beating humans with a seemingly higher level of intelligence. On the other hand, in a changing environment like traffic, results are not so promising. With every additional parameter, the situation is becoming increasingly complex. We are still able to implement a clear goal to the system but the actions to reach it are not so clearly defined anymore. A self-driving car might select its route to destination for example.

Next to fixed objectives, there is another approach - let the system learn its preferences from observing humans and create an internal belief system. Not knowing the explicit goal is essential to this method, so the system moves

from situation to situation and decides on the course of action based on the preferences it has learned from humans, case by case.

The endless search for ethics in machines

We are finding limitations to AI development not only in the physical world but also in a more philosophical sense. We keep discussing and arguing endlessly on topics like ethics and responsibility, trying to find the one answer to every question often before getting started with the technical implementation. One can think of questions like the following … Whether or not a machine should be held accountable? What is consciousness and can it be artificial? Who to blame in case systems act unexpectedly?

Misalignment between implementations and ideals

Let's assume that current views on Machine Responsibility stem from the past thirty years of AI developments where fixed-objective implementations have spectacularly succeeded in solving many challenges - for example Deep Blue and AlphaGo which both were considered greatly ahead of their time. Those practical developments and endless philosophical discussions are not supporting each other in many cases. This results in tearing the responsibility gap even greater. Practically minded engineers are often avoiding philosophical discussion and philosophers often confuse the relevancy of topics in the real world by not addressing them directly.

In this essay, I will argue that rethinking goal setting for AI systems has many benefits, one of which is the ability to bridge the responsibility gap with minimal effort, already today. In the first part of the essay, I will explain the philosophical concepts of responsibility. This is to set the foundation for coming up with solutions to distribute responsibility effectively. In the second part, I will give a brief overview of some of the proposed solutions out there. In the third part of the essay, I will give an account of what would rethinking responsibility entail. In the fourth part, I will expand on the latter and propose a solution. The essay closes with a view forward in its fifth and final part - what are some of the effects if we changed our way of building AI systems.

I

This section is mainly for introducing the reader to the philosophical notion of responsibility and what it might hold. For the sake of argument, let us just look at one of the more widely known concepts without diving too deep into it.

I have chosen the classical account of responsibility developed by Aristotle in the Nicomachean Ethics. He has given us two conditions which need to be met to exhibit responsibility:

* Firstly, the agent acting in the real world can be responsible only if it is choosing its actions;
* Secondly, the agent has to be consciously aware of the content of its actions to be responsible. (Coeckelbergh, 2016)

What is responsibility put in context?

To make it explicit, let us follow the example of driving a car. For the human driver to be considered a responsible agent, we now have two conditions that need to be met. Firstly, the driver has to be able to choose her actions. Secondly, the driver has to know what she is doing, when she is doing it. If we look at self-driving cars, the same applies. The agent chooses its actions and also has to be aware of the nature of its actions.

To generalise, there is more to knowing what you are doing than just knowing how to operate an interface or a machine. The environment and the situation in which the action is taken, for example, road conditions and oncoming traffic are also essential. Only this way would the agent have a complete picture of the situation to make decisions and eventually be held responsible. If a self-driving car chooses its actions but has no awareness of the consequences of its actions to the traffic flow or other drivers around it, then we can not hold the car responsible. (Coeckelbergh, 2016)

As we now have a better understanding of responsibility, I am going to further expand on the responsibility gap as a problem we want to solve. Also, I am going to tie it back to the way we have built our AI systems up until now. Let’s look at both conditions of responsibility separately:

* Even though a self-driving car might have the freedom to make its decisions and take action accordingly, can we be sure that those actions will not harm humans? As most of current popular intelligent systems are working on a sub-symbolic level, so without specific representations of knowledge, there is no way to verify or even prevent it. Another example - we are tasking the agent to cure cancer while we have no control over how it will achieve this. Without proper knowledge representation about the real world it is acting in, it might decide to optimise for an action in a way that is not preferred by us. This approach could include a task to spread cancer to more people to have the biggest pool of available test subjects which, can be argued, is the fastest road that leads to the cure.

* Right now, AI systems are not intelligent enough to be able to match human cognitive abilities. Even when it knows what is the consequence of turning the wheel in the scope of the car itself - both front wheels are turned into the same direction after the action is carried out - it does not have enough knowledge on the real world around it to be considered responsible. (Matthias, 2004)

II

In this section, I am giving a brief overview of some of the proposed solutions out there today. This is for the reader to better understand the situation and possibly lead towards better adoption of the alternative idea of rethinking how we build our systems in the first place.

Determinism

First of all, determinism is the philosophical belief that all events are determined completely by previously existing causes. This approach has already proven not to lead to any feasible solutions for today that can be implemented in the real world due to the complexity it involves. We need to be able to detect all the existing causes of all the events to assign responsibility. This is a problem on its own without a solution yet. Another critical aspect to consider is that determinism goes directly against the first condition of exhibiting responsibility, namely the agent's freedom and autonomy in decision making. (Van de Poel & Vincent, 2011)

Intentionality and moral luck

Kant developed a concept where we need to look at reasons behind any action the agent takes. If the intention is good, then the action can be considered good. The consequences of any action are subject to good or bad luck, though and should not be affecting our moral judgement. One can think of a self-driving car. The system has the good intention of bringing their passenger to the end destination while keeping them safe and as fast as possible. Whilst driving, the car does not stop for yellow lights and runs over a pedestrian. Under this concept, we would still consider the self-driving car to have acted well and in line with its good intentions. Who do we assign responsibility for harming the run-over pedestrian, then? This is where the approach falls short. A good aphorism to summarise some of the criticism on this - the road to hell is paved with good intentions. (Nagel, 1979)

III

Many approaches have been developed to try and give an answer to what responsibility should mean in terms of AI making decisions and taking actions. Based on some accounts given before, we can approximate that the proposed solutions are either incomplete or just very complex to understand and implement.

Replacing fixed objectives with preferences

Should an intelligent agent be able to manage its goals autonomously? Artificial General Intelligence (AGI), a system that matches or exceeds human intelligence in every aspect, is certainly expected to behave in such a manner. As of now and considering intelligent solutions already implemented in the real world, we are simply not ready to answer such questions yet. We just lack the understanding of what is involved to give a complete answer.

Instead of making machines better at reaching fixed objectives, we could benefit from machines that operate under uncertainty and on their internal belief system. This internal structure is made of preferences and updated via receiving feedback from humans. The concept should be quite universally understandable to us as it is comparable to children growing up and how parents guide them in reaching maturity. The agent (in the role of a "child") will create and update their internal belief system based on interactions with the human (in the role of a “parent"). If the task is deemed too critical and/or not enough information is known to make a decision autonomously, the system will resort to asking for preferences before continuing with the task at hand. Just as a child would ask their parent. In a sense, from this perspective Artificial Intelligence has started to resemble human intelligence.

Human Compatible: AI and the Problem of Control

Stuart Russell has come up with a formal solution that I am exploring to employ for bridging the responsibility gap. First, let's look at the solution from the original perspective - the problem of control. Without going further into it, in Artificial Intelligence and philosophy, the control problem of AI is the issue of how to build a superintelligent agent that will aid its creators, and avoid inadvertently building a superintelligence that will harm its creators. (Russell, 2019)

Russell explains how the way we have developed our intelligent systems can lead to situations where harm comes to the creators of the systems. One can think of the following example - we task AI with finding the cure for cancer. This is a fixed objective, so the system could decide to infect every human with cancer, after finding this be the optimal approach for curing the disease - the more subjects there are to test treatments on, the faster it reaches the objective. It is just optimising towards the solution. We can not be sure there is a solution at all until we find it …

There are attempts to come up with workarounds to those issues while none have fully solved them. Ultimately, the fixed objective still takes priority over everything else. Also, the course of action is still not predictable, even if fixed imperative rules are also incorporated to such agents. We need something to make up the internal belief system and provide the agent with goals via preferences so that the agent can reason on the better course of action and adjust it based on the situation at hand.

To decide whether or not to take a particular action, the system must have a preference - either do action (*X)* or do action (*Y*) or do not care, so choose randomly. We are irrational beings, often affected by our emotions so the learning of those preferences will induce the full spectrum of our behaviours to those agents.

Solution to the Problem of Control

Russell has proposed a solution consisting of three principles to solve the problem of control.

1. The robot's only objective is to maximise the realisation of human values.

What he means by this, is to change the perspective of how we perceive AI in general. He proposes to look at intelligent agents as something that would take the role of personal assistants to humans. Those agents would get their internal value system directly from the humans "they work for". This principle relies on the Completeness Axiom from Utility Theory. The agent must have a preference, one over the other or the agent is indifferent. Nothing else can not be assumed true in this situation. Confronted with two actions/options, you must either prefer (*X*), prefer (*Y*), or not care. Humans have preferences, and the machine's sole purpose is to fulfil those, whatever they may be. The agent doesn't attach any intrinsic value to its well-being or existence. This approach would result in Purely Altruistic Machines which anchor its pursuits entirely to our desires.

2. The robot is (initially) uncertain about what those values are.

Assuming perfect knowledge of the objective decouples the machine from the human: what the human does, no longer matters, because the machine knows the goal and is able to pursue it on its own. As soon as it knows what we want, we are not needed any longer.

To make it clearer - in the real world, what makes a person proceed with caution? Essentially, it is the uncertainty about the outcome of actions. The condition also brings about humility - if the agent is uncertain, it will defer to the human and will allow itself to be switched off (if that's the preference).

Additionally, this principle also solves the problem of switching of intelligent agents. With fixed objectives, the agent has a goal (*X)*. Being shut off defeats the goal (*X*). So the agent will never allow itself to be turned off, or at least until the objective is reached. With preferences, though, the agent's goal is to maximise the realisation of human preferences. As the agent is not sure what human preferences are, it is able to reason that the human will switch it off only if it's doing something wrong - that is, doing something contrary to

human preferences. By the first principle, the agent wants to avoid taking such action. But, by the second principle, it knows that is possible because it doesn't know exactly what "wrong" is. In other words, the machine has a positive incentive to allow itself to be switched off.

3. The best source of information about human values is human behaviour.

The best source of information about human preferences is not what we say, it is observed from our behaviour. We are irrational and we can't fully articulate our preferences in principles or even descriptions. This is simply because since we don't know all the options out there. We might be able to articulate a menu option at a restaurant but there are so many preferences out there that it would be impossible to word them all out explicitly. We don't know them. We also act on emotion and those actual choices will be different than what one would report if I asked "What's your value system?" This also feeds into the notion of fixed objectives and how we make mistakes by implementing them into intelligent systems.

So the agent needs to observe the choices, and that will reveal preferences, even if they are very complicated. One can think of an example - who do I wish to vote for president?

Following these observations, the agent is building a model of preferences. We as humans do it constantly - our guide is to test outcomes against our preferences, real or imagined. An intelligent agent will start building a model of what the person wants, the values, the preferences, and all that by watching the person make a great number of decisions every day.

Considerations with Russel's solution

Russel argues there is a great economic incentive to build robots that are responsive to human needs. Another positive aspect in support of his solution - we have abundant raw data indicating human preferences - all the records of human action and responses, also preferences. A lot of human essence has

been captured in books and on film, even in music. The agents could easily learn from such data.

From another perspective - it would take some effort to find stability in the real world with such agents present. People are different, companies developing agents are different and nation-states are different - all have their goals and motives which can conflict with each other.

IV

I am proposing to use a framework that’s already very familiar to us as humans - parenting. This approach incorporates many of the ideas already discussed before without the need to create a completely new concept to understand them in practice.

Parenting your AI

My proposal for bridging the responsibility gap is essential to assume the responsibility until the AI is ready to assume it for itself.

Let us look at real-world parenting. The parent might be accountable for how the seventeen-year-old child behaves in certain domains. We would not hold the parent responsible in the sense in which we might hold the child responsible — or might hold the parent responsible, had he or she been the agent. The grounds on which someone can be held accountable are much less demanding than the grounds on which they can be held responsible. (Pettit, P)

But does it make sense considering where we are now? There have already been accidents involving AI systems that resulted in death. People are worried as a result and looking for the responsible party in those situations where AI systems are involved. While we are looking for someone to blame, we could maybe benefit from rethinking AI responsibility altogether.

The responsibility gap can be finally bridged as the actions are directly informed and won't make for a situation open to interpretation. The human “parent” will always be responsible for their agent “child” until it achieves maturity. For the sake of argumentation, we could say maturity is achieved when AGI is created so that the system can decide on its own to take the responsibility and we are sure the machine can be trusted. Even though we have created the problem of defining maturity we have still improved. Together with the responsibility gap, we are also taking away the unreasonable expectations towards implementations we sport today. We can then start focussing more on finding the common ground on what a mature agent is and

when should we be able to trust it to take the responsibility. This is the question we can answer in the real world today.

Also, the approach would save us from creating complex structures for handling responsibility for AI. We could just use our legal structures today. Those structures are not by no means perfect but developed over many generations so we rely on them and in some cases, even trust in the legal structures.

Even though the responsibility gap seems to be bridged now, there are some challenges with this proposed solution as well:

- the solution assumes that we can reach human-level Artificial Intelligence or AGI;
- we will get to the point where the solution takes full effect only gradually and the liability is already there in the real world today;
- allowing machines to make mistakes to let them learn on their own.

V

Future

As Stuart Russell has said looking ahead - we will need to become good at being human. While most governing could be eventually done by machines, we still want to have control and responsibility towards our future. If we rethink how we develop AI technologies, we arguably might still succeed in this. If we build machines that ask for direction in uncertainty rather than rely on imperative goals, there's a good chance we will avoid unwanted scenarios going forward. Additionally, I believe we can become better humans in the process because that someone out there is not directly under our control but acts like us. This would entice us to improve ourselves. (Russell, 2019)

Conclusion

We have seen that the current way of developing AI is not feasible in the long run. At the same time, we are still discussing the same questions as in the early days of AI research in terms of morality and ethics which also encompasses responsibility. Not much real-world progress has been achieved and we have seen the potential reasons for it - incomplete and complex philosophical solutions to a problem that feels natural to solve to humans.

We have also seen an alternative that involves rethinking how we build intelligent systems altogether. Instead of fixed objectives, we might be better off letting AI systems learn preferences from observing humans. This will allow agents to operate in the real world much like humans do, so we also bridge the responsibility gap.

Sources

* Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control
* Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. IEEE intelligent systems, 21(4), 18-21
* Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and information technology, 6(3), 175-183
* Van de Poel, I., Vincent N. A. (2011). In Moral responsibility, Beyond Free Will and Determinism, 1–13, Springer
* Nagel, T. (1979). Moral Luck. Mortal Questions. Cambridge: Cambridge University Press. pp. 24–38. OCLC 4135927
* Pettit, P. (2007). Responsibility incorporated. Ethics, 117, 171–201
* Coeckelbergh, M. (2016). Responsibility and the moral phenomenology of using self-driving cars. Applied Artificial Intelligence, 30(8), 748–757