Overview and challenges of normative reasoning

William Bitsch
M.Sc Student of Mathematics
Free University of Berlin
Department of Mathematics and Computer Science
E-mail: william.bitsch@fu-berlin.de

In this article we will give an overview of normative reasoning, and state some of its challenges. First a motivation regarding machine ethics is given. We then define normative reasoning and consider different approaches to its formalisation. In the end we will discuss challenges in this field and state further questions.

1 Motivation: Machine Ethics

Even before the first computers were invented the first humans thought about creating intelligence similar to ours. Most of those thoughts revolved around myths or thought experiments (e.g. Laplace's demon). This changed though by the first half of the twentieth century with the invention of the electrical computer and the transistor. The ability to perform complex calculations reinvigorated the hope to be able to simulate the whole process of human thought one day even though artificial intelligence as capable as human intelligence still seems far away today. Nevertheless the research in artificial intelligence made enormous progress in the last decades. This development can be mostly traced back to the exponentially increasing processing power of computers and the research on ever more complex algorithms that harness this power. Historically there were many different approaches to construct intelligence. Today especially machine learning algorithms are heavenly studied and get integrated into more and more applications we use everyday. The general idea behind machine learning is that instead of giving a computer explicit instructions on how to do a task we give it lots of data on how the task is done and a way to learn on this data. This self learning approach brings great advantages since it gives computer scientists the capability to give programs some kind of intuition for previously unsolvable tasks. Programs based on those algorithms bring great benefits by not only automatising or improving tedious tasks as labeling images, analysing speech, making product recommendations or recognising spam e-mails but also seem capable of doing more complex tasks as playing sophisticated games, driving a car or trading assets. This approach also brings disadvantages, since programmers pass their responsibility and control on how a task is done down to the data they program their models on. Especially biased data can lead to models that discriminate people or make rigged choices with far reaching consequences. Problems like this naturally motivate the question on how can we prohibit problems like this. How can we design artificial intelligence to behave lawfully and morally?

2 Normative reasoning

In a general context *normative* means to attribute a value. This gives rise to the terms of the *norm* which is a standard we valuate against and *normativity* which is the appearance of norms in a society. We use norms often in our daily live when we say how something should or ought to be. The concept of a norm is sometimes defined more precise depending on the context. *Ethical norms*, *legal norms* and *social norms* are commonly used in

their respective disciplines. In this paper we will not restrict ourselves to one of those definitions but rather work with the most general one.

Normative reasoning is concerned with how we inference with norms. It seems promising to formalize this process for several different reasons. First we could get a grip on the problems of algorithms being unethical stated earlier by implementing moral behaviour into our system. Second formalized processes of normative reasoning could be compared with actual human behaviour which extends our knowledge on normativity in a practical sense. Lastly formalizing strengthens our understanding of general desriptive theories on reasoning which is especially interesting philosophically. We call such a formalized model a *normative system* or a *normative theory*. While it seems tempting to try to define one standardized normative system to use for all researchers the community is divided into many different frameworks and normative theories. Often models can be embedded into each other but this often lacks utility. Models are designed very differently because they are made to handle different types of normative scenarios and therefore also model those scenarios more understandably and efficient.

Before we dive into those different approaches it is useful to point out some subtleties. First of all we need to distinguish between normative and practical reasoning. Practical reasoning is concerned with intentions. The difference between a norm and an intention gets noticeable by considering the following. If A implies B and I intend to do A then I do not automatically intend B but if it ought to be A it also ought to be B. For example just because I intend to live healthy I do not automatically intend to drink the spirulina smoothies my girlfriend offers me every morning. But if I ought to live healthy I also ought to drink the smoothie. Thus norms are more absolute than intentions in the way that logical implications also imply normative implications and therefore modeling norms should be conducted differently than intentions. Also it should be noted that normative theories can not say what is right or wrong in the same manner logic can not tell what is true or false. Both disciplines are concerned with the process of inference and not the actual values itself. We could also study normative systems that are equipped with the norms of discrimination being good and human rights being bad. To find the right norms for a normative systems to apply to real world challenges still remains an important question in philosophy.

3 Approaches to normative reasoning

Norms have different properties that get considered when we try to formalize them. In Nick Chater and Mike Oaksfords paper „Normative Systems: Logic, Probability, and Rational Choice "the authors identify three different characteristics of normative beliefs which give rise to three different approaches to model those properties. First we could want to model the consistency of our beliefs i.e. we would like our norms to be compatible with each other. The most direct way to handle this is to use modal logics. *Deontic logic* is a subbranch of logic that is concerned with such norms but it should be noted that because of our broad definitions of norms there can be norms that are not deontic [**?**]. Nevertheless we will restrict our self to model normative behaviour in the context of deontic logic. Another trait norms often posses is that they don't need to be absolute. They also can be equipped with some degree of how hard we believe into them. This observation gives rise to *probabilistic methods* used to model norms. Also norms are very commonly used when we have a set of choices and we want to pick the normative right ones. A theory concerned with questions of selecting the right option from a set of choices is called *rational choice theory*. All approaches have advantages and disadvantages we will describe more detailed.

3.1 Logic

Formal logic gives us the tools to specify the reasoning process we use in our natural language in terms of a formal language equipped with strict rules of how to do inference. One of our main challenges is that our own normative thinking is heavily influenced by background beliefs. If we have to evaluate some situation we intuitively draw a lot of knowledge from the back of our minds. While this is very useful it also makes the process of formalizing our natural reasoning very hard. We therefore restrict our self to model our beliefs lo-

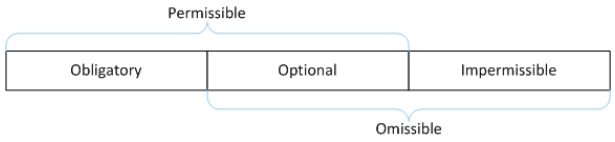


Abbildung 1. Traditional threefold classification of deontic modalities, https://plato.stanford.edu/entries/logic-deontic/ [21.02.20]

cally. Deontic logic can be defined as a subbranch of modal logic. Modal logic is an area of logic that is concerned with statements that not only can be true or false but also can have a mode that qualifies the truth value. We call such statements *modals*. For example we could not only say „The cake *is* on the kitchen counter. "We could also say „The cake *is possibly* on the kitchen counter", „The cake *is permitted to be* on the kitchen counter "or „The cake *is believed to be* on the kitchen counter ". While the first statement is a simple statement about where the cake is, the following statements are *modals* about where it is possible for it to be, where it should be and where we know it to be. Modals can have different logical properties, for example if the cake is on the kitchen counter then it also needs to be true that it is possible for it to be on there but it does not imply that anyone knows about the location of the pastry. Thus we divide modal logic into different branches that are concerned with different types of modality. Modalities of possibility are called athletic, modalities of obligation/permission are called deontic, modalities of belief doxastic, and there are many more. To show that a statement is a modal we use modal operators that qualify a statement in some way. In deontic logic we traditionally use **PE** for permissions and **OM** for omissions. For example denote $light(Gizmo)$ to be the statement that Gizmo is exposed to light. We then write **PE** $light(Gizmo)$ for the statement that Gizmo *is omissed to be* exposed to light. Following from this we can also define **OM** for omissions, **OP** for optionals and **OB** for obligations from these modal operators. More generally we can first introduce any of the five deontic operators and define the others from it. For example can define the operator **OM** from **PE**. For this just consider that a statement p is omissible if and only if not p is permissible, i.e. $\mathbf{OM}p \leftrightarrow \mathbf{PE}\ \neg p$. Now knowing the basics on how to present a deontic statement how can we inference with them. Consider the following example:

All mogwais ought not get wet
Gizmo is a mogwai
Gizmo ought not get wet

which we could formalize in first order predicate calculus extended with the deontic operators in the following way:

$$\frac{\forall x : mogwai(x) \rightarrow \mathbf{OB}\neg wet(x) \quad mogwai(Gizmo)}{\mathbf{OB}\neg wet(Gizmo)}$$

First note that we not only use deontic statements when we want to reason deontically. Next note that $\mathbf{OB}\neg wet(Gizmo)$ is equivalent to the reduced statement $\mathbf{IM}wet(Gizmo)$. Even though this inference seems to be rather simple the definition of rules to reason with our modal operators remains controversial to this day and is a topic too big to be covered here. Thus while there is one common system named standard deontic logic (SDL), there is not one deontic logic standard but rather many that inherit different aspects from the deontic logic presented here, but discard and refine others.

3.2 Probability

Since many problems in the real world contain uncertainty, also normative beliefs can have different degrees. Interpreting probability as a degree of belief we can model this phenomenon using probability theory. This point of view is called the *subjective interpretation of probability* in philosophy and the *bayesian approach* in statistics. Bayesian statistics uses Bayes' law, an elementary theorem of mathematics to update beliefs with obtained data. Let A and B be some events. In stochastics we write $P(A)$ for the probability of the event A happening, $P(B)$ for the probability of the event B happening and $P(A \cap B)$ for the probability that both events A and B happen. We can then define $P(A|B) = P(A \cap B)/P(B)$ for $P(B) \neq 0$ and $P(A|B) = 0$ otherwise, as the conditional probability of A under B. This probability can be interpreted as the probability of A happening under the assumption that B happened. This give rise to Bayes' theorem:

Theorem 1 (Bayes' Theorem)**.** *Let A and B be events with $P(B) \neq 0$, then*

$$P(B|A) = \frac{P(A|B)P(A)}{P(B)} \tag{1}$$

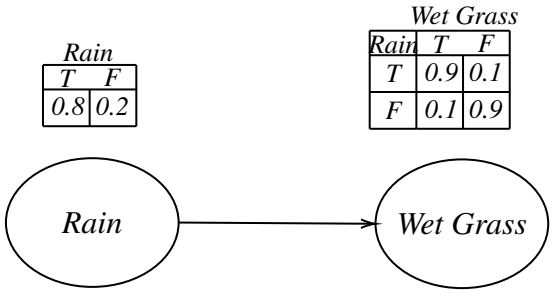
Proof. The statement is trivial for $P(A) = 0$ so let us consider $P(A) \neq 0$. We have $P(A|B) = P(A \cap B)/P(B) \iff P(A|B)P(B) = P(A \cap B)$ and therefore symmetrically also $P(A \cap B) = P(A|B)P(B)$ which yields $P(A|B)P(B) = P(A|B)P(B)$ and thus $P(B|A) = \frac{P(A|B)P(A)}{P(B)}$ since $P(B) \neq 0$.

Bayes theorem now helps us to derive unknown probabilities from known ones. A tool that uses this method are the so called *bayesian networks*, which were popularized by Judea Pearl and are widespread in the field of artificial intelligence.

Definition 1. *A bayesian/belief network is an acyclic directed graph $G = (V,E)$ which is probabilistic. That means that the vertices correspond to random variables and the arcs to probabilistic dependencies between the vertices.*

The acyclic and directedness conditions allow for efficient algorithms for interference and learning in those networks. Before we dive into how we can use this models to model normative beliefs we will first consider the following example:

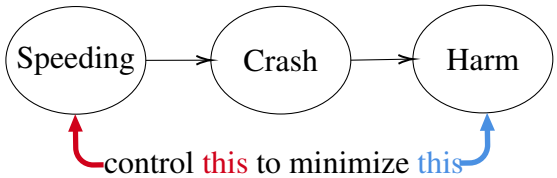
Example 1.



In this example we have two different events. The first being that it rains (Rain) and the second being that the grass is wet(Wet Grass). As one would assume that the grass is wet does not effect the weather directly but the rain definitely affects if the grass is wet. We can now use this model to update probabilities given data. Assume it rains i.e. $P(Rain = T) = 1$ and $P(Rain = F) = 0$. What is the probability of the grass being wet? Just consider $P(Wet\ Grass = T)$

$$\begin{aligned}
&= P((Rain = T) \cap (Wet\ Grass = T)) \\
&+ P((Rain = F) \cap (Wet\ Grass = T)) \\
&= P(Wet\ Grass = T | Rain = T) P(Rain = T) \\
&+ P(Wet\ Grass = F | Rain = F) P(Rain = F) \\
&= P(Wet\ Grass = T | Rain = T) = 0.9
\end{aligned}$$

We can use this reasoning for very large networks to complete various tasks. One possible application would be the following. One common problem with self driving cars is that they frequently have crashes were people behind them speed and drive into the rear end of the autonomous vehicle [**?**]. Normally a driver could accelerate to a non-legal speed to prevent this crash but the self driving car is programmed to always comply to the traffic laws. How could we bypass this behaviour in a situation where it could harm the passengers without giving the car the possibility to always go rogue? Consider the following simplified bayesian network:



Speeding is the degree of belief of the car that itself should drive below the speed limit. Crash is the probability that the car will have an accident which gets evaluated by the autonomous driving system of the car. Harm measures the probability that the passengers will get injured. In our network speeding has an effect on the probability of a crash which has an effect on the probability that the passengers get harmed. When our car drives on the road the probability for a crash changes over time. This effects the probability that the passengers get harmed. Now if the harm variable reaches a certain threshold we could allow the car to bypass the norm that prohibits speeding if this minimizes the probability of a crash. The reduced probability would then again decrease the risk that the occupants get hurt. This model is very simplified but captures the main essence we would need to build a utilitarian autonomous vehicle.

3.3 Rational choice theory

Microeconomics is an important research area in economics that is concerned with the decision

making of different actors regarding the allocation of scarce resources. The need for an abstract model for individual decision making to gain a solid foundation for this discipline is one of the main reasons rational choice theory developed to where it is today. The resulting theory is therefore very applied. *Rational choice theory (RCT)* could be defined as a collective term for different frameworks to model the preferences from a rational agent over a given set of choices where the preferences are assumed to be complete and transitive. Rational means that the actor always chooses what is in their best interest. Complete means that given a binary choice the individual is always able to tell which option they prefer or if it prefers neither of those options. Transitive means that given three choices A, B and C we know that if the person prefers A over B and B over C then he or she also prefers A over C [?]. While the principles of RCT are mostly used in economics we can also use them when reasoning normative. Instead of using inherent preferences we can use norms to evaluate given choices. While the basic theory has a reasonable scope once the choices are evaluated it still remains challenging to find a function that evaluates moral choices.

4 Challenges

The formalization of normative reasoning is still researched and there are not only the three different approaches we have covered here but also hybrids and probably also ones that do not fit in any of those schemes. Thus when normative reasoning is formalized for a given task it is still challenging to find the right system for this purpose. For example in systems of deontic logic there can be paradoxes, i.e. situations that run against our understanding of valid reasoning. A great example for this is *Ross'sparadox*. In standard deontic logic we want that if $p \to q$ then also $\mathbf{OB}p \to \mathbf{OB}q$ as we discussed before in the smoothie example. But if we consider the statements „I send the letter "(p) and „I send the letter or burn the letter"$(p \vee q)$. It is clear that $p \to (p \vee q)$ but it sounds unreasonable that $\mathbf{OB}p \to \mathbf{OB}(p \vee q)$ since the statement $\mathbf{OB}p$ should probably not entail any obligation that can be fulfilled by burning the letter.

Another major challenge lies in the *frame problem* where we want to model how different aspects of a model are entangled and change if new information is provided. For example if I ought to clean my house but my house gets burned down by a fire then my obligation gets non sensible. This entanglement is not only hard to model but also gives proper reasoning a much higher complexity. Generally there is still a lot of research to be conducted in finding efficient algorithms for normative reasoning.

Next we should ask how can we deal with inconsistency. First our norms can be conflicting. Second a lot of situations can arise that force us to act against our norms. Especially in classical logic we then can get a logical ëxplosionßince by using the law of the excluded middle we can conclude every statement from our premises.

To study systematically in which ways different definitions of norms differ logically and how norms can be learned from a given set of data seems promising for further research.

Acknowledgements

This paper was written as part of a seminar on normative reasoning and machine-ethics organized by Prof. Dr. habil. Christoph Benzmüller and conducted in the winter semester of 2019/2020 at the Free University of Berlin.

Literatur

[1] Nick Chater, M. O. "Normative systems: Logic, probability, and rational choicey". *10.1093/oxfordhb/9780199734689.013.0002*.

[2] Broome, J. "Normative practical reasoning i". *10.1111/1467-8349.00085*.

[3] McNamara, P. "Stanford encyclopedia of philosophy: Deontic logic". *https://plato.stanford.edu/entries/logic-deontic/*.

[4] Jan Broersen, Stephen Cranefield, e. "Normative reasoning and consequence". *10.4230/DFU.Vol4.12111.33*.

[5] Faltin F., K. R. "Encyclopedia of statistics in quality reliability, bayesian networks". *http://www.eng.tau.ac.il/ bengal/BN.pdf*.

[6] Jonathan Levin, P. M. "Introduction to choi-

ce theory". *http://web.stanford.edu/ jdlevin/Econ%20202/Choice%20Theory.pdf*.

[7] McNamara, P. "Challenges in defining deontic logic". *https://plato.stanford.edu/entries/logic-deontic/challenges.html*.

[8] Crook, N., 2018. Koncrete podcast: Nathan crook. Youtube, November. https://www.youtube.com/watch?v=elXMAFc7joU.