Ethical Considerations for the Deployment of Logic-Based Models of Reasoning

Aaron Hunter
Department of Computing, British Columbia Institute of Technology, Burnaby, Canada

Keywords: AI Ethics, Logic-Based AI, Belief Change.

Abstract: Ethical considerations in the development of Artificial Intelligence systems need to be addressed as more systems are deployed in practice. Much of the current work in this area is focused on Machine Learning systems, with an emphasis on issues such as fairness and bias. However, there are also fundamental ethical problems to be addressed in simple logic-based systems, and we do not have solid methods in place to handle these issues. In this paper, we discuss ethical problems that are implicitly introduced in the deployment of systems that formalize reasoning in logic. As a specific example, we focus on logic-based models of belief change. We consider the way belief change operators are defined, and how unintended behaviour can emerge in operators defined with respect to well-known rationality postulates. Preventative measures and potential solutions are discussed.

1 INTRODUCTION

Considerations around the ethics of Artificial Intelligence (AI) have become increasingly important, as more AI systems are deployed in practice. There have generally been two lines of work in this area. One type of work is related to general, forward-looking concerns about the ethical behaviour of machines that make decisions using human-level intelligence (Malle et al., 2019). The other type of work that is commonly seen is explicitly focused on issues around bias and responsibility for systems based on Machine Learning (ML) (Mehrabi et al., 2021). However, it is likely that future AI systems will actually be hybrid systems, that use a combination of learning technology together with models of formal reasoning. For this reason, it is important that we also examine the ethics and biases embedded in traditional logic-based models of reasoning.

In this paper, we look in detail at one specific logic-based model of reasoning: the theory of belief change. We view this as kind of a case study, highlighting different ethical issues that occur in a logical formulation. Note that we are not concerned with broad philosophical concerns about the relationship between logic and ethics. Our concern here is the manner in which simple, logically-defensible choices introduce problems with ethical decision making. The focus is on providing a pathway to move forward with the deployment of hybrid systems that use learning technology and logic in concert.

2 PRELIMINARIES

2.1 Motivating Example

The following example is a variation on a well-known problem originally presented by Darwiche and Pearl (Darwiche and Pearl, 1997). The example here is modified by changing the context to make the connection with moral decision making more clear.

Consider a situation where we have an intelligent system that is processing job applications for a position in engineering. A given applicant is initially identified as being part of a disadvantaged group (D) that should get priority for job interviews; so the system believes D to be the case. On further examination of the application materials, it appears that the applicant is registered as a professional engineer (E), so this is also believed. However, a second system is consulted to verify the credentials, and it indicates that the school the applicant attended is not properly accredited. This leads the system to reject E.

The question at this stage is what happens to the previous beliefs, such as the belief that the applicant belongs to a disadvantaged group. This might be important to maintain, because there might actually be

Hunter, A.
Ethical Considerations for the Deployment of Logic-Based Models of Reasoning.
DOI: 10.5220/0011704500003393
In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023) - Volume 3*, pages 565-569
ISBN: 978-989-758-623-1; ISSN: 2184-433X

other suitable opportunities for this applicant. We will see that some formal models of belief change will actually reject the first information, as if it had never been obtained. On the other hand, some models will actually keep the older information. In the present example, this is appropriate. However, we argue that this can also be a problem in different contexts.

The fundamental problem is that the well-known postulates that describe rational belief change can be satisfied by significantly different processes, in a way that may introduce unintended bias into a system that is intended to provide a *normative* model. While the problems introduced through biased data sets in ML approaches are now being increasingly recognized, this kind of bias in an axiomatic reasoning system needs to be put under the same scrutiny.

2.2 Belief Revision

Broadly, the study of belief revision is concerned with the manner in which agents incorporate new information with their pre-existing beliefs. The most influential approach to belief revision is the AGM approach, in which a set of rationality postulates is specified to define the essential features of the revision process (Alchourrón et al., 1985). This framework is defined in the context of propositional logic. So we assume a fixed vocabulary V, which is just a set of propositional variables. Formulas over V are defined using the usual logical connectives $\neg, \wedge$. The beliefs of an agent are represented by a logically closed set of formulas.

An AGM revision operator is a function $*$ that maps a belief set K and a formula ϕ to a new belief set $K * \phi$. The new belief set represents what an agent should believe after incorporating ϕ. Informally, we want to add ϕ to the belief set, and remove as little as possible while maintaining consistency. An AGM revision operator must satisfy a set of postulates, which we do not list here in the interest of space. However, as an example, one straightforward postulate is the so-called *success postulate*:

$$K * \phi \models \phi.$$

This simply says that new information should be believed; this postulate encapsulates the fundamental assumption of the AGM approach that new information is provided by a reliable source. It is well-known that the semantics of AGM belief revision can be defined with respect to plausibility orderings over propositional interpretations (Katsuno and Mendelzon, 1992).

In this paper, we are not directly concerned with AGM revision. It has a known limitation in that it can not be used for *iterated belief change*. The most influential approach to iterated belief change is an extension of AGM called DP revision (Darwiche and Pearl, 1997). The main difference is that the beliefs of an agent are now represented by an *epistemic state* $\mathbf{E}$, which is essentially a total pre-order over possible interpretations. The minimal elements of this ordering are considered the most plausible states, and we let $B(\mathbf{E})$ denote the set of formulas that are true in all of the minimal states. At the level of formulas, there is a set of postulates for DP revision that essentially guarantee that they behave like AGM revision operators. But then there are four additional postulates that explicitly deal with iterated change:

Darwiche-Pearl Postulates
[DP1] If $\beta \models \alpha$, then $B(\mathbf{E} * \alpha * \beta) = B(\mathbf{E} * \beta)$.
[DP2] If $\beta \models \neg\alpha$, then $(\kappa * \beta) * \alpha = \kappa * \alpha$.
[DP3] If $\alpha \in B(\mathbf{E} * \beta)$, then $\alpha \in B(\mathbf{E} * \alpha * \beta)$.
[DP4] If $\neg\alpha \notin B(\mathbf{E} * \beta)$, then $\neg\alpha \notin B(\mathbf{E} * \alpha * \beta)$.

It turns out that there are many different DP operators that satisfy all of the postulates. We mention the two most extreme examples:

- *Natural Revision*: After revision by ϕ, $B(\mathbf{E})$ becomes the set of states where ϕ is true. For all other states the ordering defined by $\mathbf{E}$ is preserved.
- *Lexicographic Revision*: After revision by ϕ, all states where ϕ is true are moved before the states where ϕ is false. The relative ordering between ϕ-states and $\neg\phi$-states is preserved.

Figure 1 gives a schematic depiction of how these operators work. We remark that there are known issues with these operators, and improvements have been proposed (Booth and Meyer, 2006; Jin and Thielscher, 2007). However, for the purposes of this paper, the significant feature is simply that these two extreme examples satisfy the raitionality postulates.

3 ETHICAL CONSIDERATIONS

3.1 On Prescriptive Theories

For the ethical evaluation of formal belief change, we will consider several approaches: Kant's categorical imperative, utilitarianism, and virtue ethics. These are all well-known approaches, and we refer the reader to (Hill, 2009; Rosen, 2003; Hursthouse, 2001) for a description of each approach, as well as comparisons between them. For the purposes of this paper, the intention is not to draw hard ethical conclusions; our primary intention is simply to highlight where ethical issues can creep into formal models of reasoning.

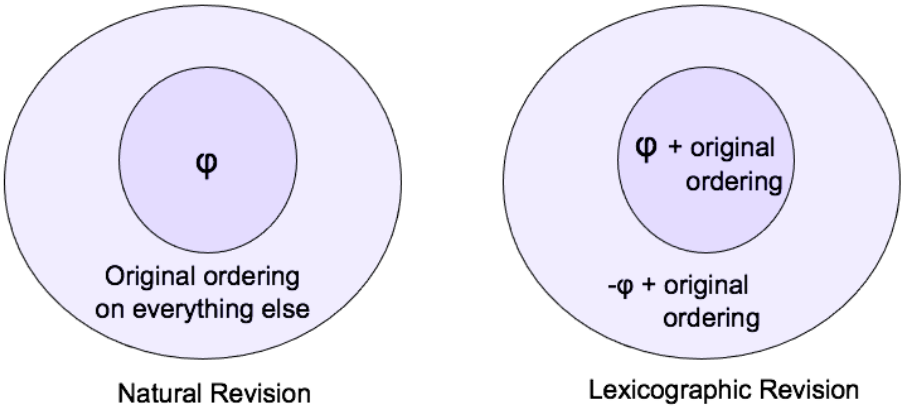

Figure 1: Two Revision Operators.

A distinguishing feature of belief change research is that it is *prescriptive* rather than *descriptive*. It is clear that real humans do not follow AGM-type rationality postulates when they revise their beliefs. For example, human agents do not generally have logically closed belief sets. The fundamental view is that a fully rational agent *should* obey the principles, and it is therefore a human failing to some extent that we do not.

There are clearly ethical concerns around human-defined postulates dictating how beliefs "should" be revised. In most work on belief change theory, the postulates are justified through small sets of motivating examples. The selection and design of these examples can introduce bias, or unjustified assumptions for a general reasoner. This is certainly a problem, although it is not one a particularly novel one. In any scientific endevour, we need to be careful about the consideration of biases of the researcher. We are more concerned with problems that are hidden in the formalism, irrespective of the intentions of the researchers.

3.2 Problems with Ethical Decision Making

We return to our motivating example, regarding the system that is vetting job applicants. It is easy to demonstrate that natural revision will give the following result:

1. After seeing the applicant is from a disadvantaged group (D), the agent believes D.
2. After seeing the engineering degree, the agent believes E.
3. After learning the program is not accredited, the agent revises by $\neg E$. The agent disregards the observation D, as if it had never happened.

This result is ethically problematic:

- *Kantian view*: The applicant in this case provided information about their status in a designated group, because they wanted that information to be considered. By disregarding this information, the sequence of revisions here is not treating this applicant as an end to themselves; they would presumably like to be considered for all available positions.
- *Utilitarian view*: Rules around improving opportunities for disadvantaged groups are generally accepted as being good for society overall. This revision process is therefore leading to a harmful result.
- *Virtue ethics view*: As a general principle, it is a virtuous property to look for all opportunities suitable for a qualified candidate - taking into account all relevant factors. This has not occurred here.

Hence, it is very easy to argue that the system in this case is performing in a way that is ethically problematic. The reason it is problematic is simple: in this particular context, older information is important - yet it is being disregarded. We remark that this approach to revision does respect a large set of rationality postulates, but this problematic behaviour has been observed and addressed previously (Booth and Meyer, 2006). However, the solution in most cases is to introduce more logical postulates. It is difficult to guarantee that operators satisfying the new postulates will be immune to similar problems.

Another important point is that the behaviour of this revision operator is not *inherently bad*. The result of this revision is problematic for this particular context. However, there are other contexts where it is ethically preferable to ignore earlier reports. Consider, for example, a situation where a particular view has been discredited. The view could be a social view

from an early era, or perhaps a scientific theory that has been proved false. In such cases, we may actually want to disregard previous revisions in line with the natural revision approach. The problem is not that this model of revision is "wrong"; the problem is that it can be ethically problematic in certain contexts.

3.3 The Problem with Iteration

In general, the problem with formal models of iterated belief revision is the fact that they need to ascribe value to information from the past. The formal models have to explicitly answer this question: how much weight should be given to past information that has been discredited?

The problem here is not easily addressed. The general approach of the discipline is to use rationality postulates that specify minimum requirements for a suitable operator. But the rationality postulates do not describe a definite approach. Instead, the rationality postulates describe a set of operators. The person that is interested in modelling a particular problem is able to choose the specific operator that is appropriate for their context.

However, we argue that this approach is unlikely to be successful beyond very limited and precisely described contexts. In this paper, we have only looked at the two extreme cases of iterated revision. In natural revision, information obtained in the past carries little weight as compared to new information. On the other hand, in lexicographic revision, the full impact of past revisions is maintained as new information is received. But these are not the only examples. There are many different operators that each value past information differently; each operator may lead to ethical problems in particular contexts.

We need more than rationality postulates. These postulates are excellent for establishing the mathematical properties of the system, but they do little to help determine where particular operators can be deployed. For this reason, we propose that a fundamental new direction for research would be useful. With any formal model of reasoning, one would like to have a precise characterization of the contexts where that model can be trusted to avoid ethically problematic decision making.

4 CHALLENGES

4.1 Discussions with Non-Specialists

One of the challenges with the deployment of logic-based reasoning systems is the fact that the postulates and the logics are not accessible to non-specialists. While the literature focuses on formal characterization results and proofs of correctness, these kinds of results are of little interest to most users of an AI system. All a user would like is a simple tool that models how an agent makes decisions, without worrying about the details.

But this is a problem when there are several different formal models that might behave poorly in different contexts. In order to deploy logic-based reasoning tools in an ethical manner, we therefore propose the following steps:

- *Knowledge acquisition.* We must consult with domain experts to understand the context where the reasoning system will be used. In the case of belief revision systems, one aspect of this process would be to determine how past information is perceived and valued.
- *Validation.* We need to formally validate that the reasoning system we propose behaves well in the given context. This means that we need to move away from the hand-crafted examples produced by AI scientists, and focus on validation with respect to real examples where the system is likely to be employed.

Step 1 here is similar the interview process one might do in the development of an expert system. This is challenging, but necessary. For example, we have seen that some logic-based resasoning systems can lead to unethical behaviour when used in the wrong context. If we want to develop such systems to be deployed in hybrid AI systems, we need to be able to explain to communicate effectively with a non-specialist user.

4.2 Formalizing the Goal

Discussing the deployment setting is an important step, but it will not address the problem. If we look towards AI ethics research in ML as a guide, we can see that issues of bias emerged over time as systems were deployed. As a result, there has been significant effort to address the problem. There are clear principles around key notions of fairness and ethics for learning systems (Mehrabi et al., 2021). These principles might not always be followed, but at least they are specified and under discussion.

On the logic side, the situation is quite different. In many cases, logical reasoning systems are defined with respect to explicit principles that capture some aspect of rational decision making. For example, the rationality postulates in the theory of belief change are specified by a person based on intuition

and simple examples. This is clearly a problem in terms of implicit bias. The validation of the system is generally a mathematical validation addressing issues around syntax, semantics, and efficiency. To some extent it is assumed that a particular belief change operator should only be used in a restricted context, but that context is not specified. We remark that this situation is not unique to the theory of belief change, similar patterns can be seen in other areas of formal AI.

We propose that the solution is to develop a mechanism for specifying context and communicating it to practitioners for deployment. We need to be be able to precisely specify the characteristics of the reasoning problems that can be tackled with a particular model, without running into clear ethical problems. This mechanism will be helpful to those choosing frameworks and models in software, but it is also essential on the theoretical side to formally demonstrate properties around fairness and bias for particular logical models.

5 CONCLUSION

In this position paper, we have outlined a particular challenge for AI ethics. Much of the work on AI ethics today has focused on subtle, difficult problems related to bias in learning systems. However, we argue that there is a much less subtle problem on the logic side, where models are developed by hand and it is easy to find examples where such models do not make ethically sound decisions. We have demonstrated this through a concrete example, looking at formal belief change operators. While the problems in this setting are much easier to see, the solutions are not obvious. We have proposed that explicitly working with domain experts and formalizing context may be an important step towards the eventual safe deployment of hybrid systems that involve formal reasoning.

REFERENCES

Alchourrón, C., Gärdenfors, P., and Makinson, D. (1985). On the logic of theory change: Partial meet functions for contraction and revision. *Journal of Symbolic Logic*, 50(2):510–530.

Booth, R. and Meyer, T. (2006). Admissible and restrained revision. *Journal of Artificial Intelligence Research*, 26:127–151.

Darwiche, A. and Pearl, J. (1997). On the logic of iterated belief revision. *Artificial Intelligence*, 89(1-2):1–29.

Hill, T. (2009). *The Blackwell Guide to Kant's Ethics*. John Wiley and Sons.

Hursthouse, R. (2001). *On Virtue Ethics.* Oxford University Press.

Jin, Y. and Thielscher, M. (2007). Iterated belief revision, revised. *Artificial Intelligence*, 171(1):1–18.

Katsuno, H. and Mendelzon, A. (1992). Propositional knowledge base revision and minimal change. *Artificial Intelligence*, 52(2):263–294.

Malle, B. F., Bello, P., and Scheutz, M. (2019). Requirements for an artificial agent with norm competence. In Conitzer, V., Hadfield, G. K., and Vallor, S., editors, *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 21–27. ACM.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

Rosen, F. (2003). *Classical Utilitarianism from Hume to Mill*. Routledge.