Non-Monotonic Reasoning

Konstantin Kueffner

June 20, 2018

1 Introduction

Before introducing non-monotonic logic in a more formal manner, this paper shall develop a intuitive understanding of the different kinds of reasoning such formalisms intend to capture.

1.1 Motivation for Non-Monotonic Reasoning

The initial motivation for the development of non-monotonic formalisms, is strongly intertwined with the field of artificial intelligence. That is, several problems within the field of artificial intelligence, require some form of non-monotonic reasoning, thus driving the development of non-monotonic inference systems. Due to this fact, the initial solutions for capturing non-monotonic reasoning were more or less pragmatic in nature, providing an incentive for the development of a more systematic foundation for non-monotonic reasoning. Hence, some of the formalisms discussed below are strongly shaped and motivated by problems related to the field of artificial intelligence. Bochman 2007. Therefore, in order motivate the following investigation of non-monotonic formalisms some of these problems will be presented. Moreover, these problems shall provide, on an intuitive level, a brief introduction into the kinds of reasoning such non-monotonic formalisms try to capture. This intuition will then be developed further in the following sections.

Tweety-Problem

One of the most prominent example in the literature concerning non-monotonic reasoning is the *Tweety*-Problem Reiter 1980.

Humans employ a myriad of default assumption about the world. For example, if one announces the statement "Look at that bird!!!". The reaction of looking up into the air or at trees, seems to be fairly reasonable. This is, due to the fact that (usually) birds (B) fly (F). Hence, if one intends to capture this fact in classical logic, the following formalisation could be obtained.

$$B(x) \rightarrow F(x)$$

If one announces the fact that there exists a bird called "Tweety" (t), it seems reasonable to use the previous formalisation to obtain the following derivation.

$$\frac{B(t) \rightarrow F(t) \qquad B(t)}{F(t)}\ \text{(Modus Ponens)}$$

So far so good. However, what happens if after this derivation additional knowledge about "Tweety" is obtained. Namely, that "Tweety" is a penguin (P). Furthermore, it is common knowledge that penguins do not fly and that they are birds. Hence, an attempt of formalising these attributes about the world, may result in the following statements.

$$\begin{aligned} &P(t) \\ &P(x) \rightarrow B(x) \\ &P(x) \rightarrow \neg F(x) \end{aligned}$$

Thus, the following can obtained

$$\cfrac{\cfrac{\cfrac{B(t) \rightarrow F(t) \qquad B(t)}{F(t)}\ \text{(MP)} \qquad \cfrac{P(t) \rightarrow \neg F(t) \qquad P(t)}{\neg F(t)}\ \text{(MP)}}{F(t) \wedge \neg F(t)}\ (\wedge i)}{\bot}$$

The above derivation is obviously problematic, as it provides a witness for the inconsistency of the proposed formalisation of the world. In contrast, most humans are able to cope with the fact that there are some birds that can not fly, even if normally it can be assumed that all birds fly. Since, the aim is to model certain aspects of the reasoning employed by humans it is obvious that the presented formalisation of the world is insufficient, as it does not account for the exception penguin. However, this can be accounted for by modifying $B(t) \rightarrow F(t)$ such that

$$B(x) \wedge \neg P(x) \rightarrow F(x)$$

In an subsequent derivation attempt, the following can be observed.

$$\frac{P(t) \rightarrow \neg F(t) \qquad P(t)}{\neg F(t)}\ \text{(MP)}$$

Moreover, by including penguin as an exception and since $P(t)$ is known, the application of $B(t) \wedge \neg P(t) \rightarrow F(t)$ is prevented, thus resolving the issue. However, if one now adds another bird into the formalisation, namely the bird "Robin" (r), one has to ensure that r is definitely not a penguin, i.e. $\neg P(r)$. Otherwise, $F(r)$ can not be inferred. That is, by adding penguin as an exception the rule required to infer $F(r)$ is blocked as long as it is certain that r does not fulfil any exceptions.

This seems manageable, if only few exception have to be made. However, this is not always the case. Even in the current example, we would have to take care of birds like ostrich, emu, kiwi and others. Moreover, what if a new earthbound species of bird is discovered? That is, how should one proceed with the formalisation, if it is impossible to state every possible exception? Therefore, it can be concluded that if no exceptions are introduced, the theory about the world becomes inconsistent. However, accounting for all such exceptions is hightly impractical, if not impossible Reiter 1980; Bochman 2007.

In this example, it can be observed that the kinds of inferences available in classical logic are not sufficient for reasoning about the problem, as a human would do. Namely, a human is capable of jumping to conclusions by assuming normality in order to cope with partial information. Moreover, the conclusions drawn based on such assumptions can be retracted, if evidence to the contrary is presented. Therefore, allowing humans to cope with statements such as

Birds *normally* fly

To conclude, humans are able to make default assumption, use these assumption to formulate conjectures and revise them, if evidence to the contrary is presented. Moreover, since most of those assumptions are based on beliefs, two conflicting conjectures are not be seen as a contradiction. This form of reasoning, i.e. reasoning where one is able to retract inferences, in light of additional information, is called *defeasible reasoning.* The relevance of defeasible reasoning is not restricted to artificially constructed problems such as the Tweety problem, but it occurs frequently in the everyday reasoning of humans. For example, in medical diagnosis or even scientific reasoning, e.g. revising of hypothesis in light of contradicting measurements. Moreover, while classical or intuitionistic logic are characterisations of *deductive reasoning*, which does not accommodate the retraction of inferences, the formalisms contained within the umbrella term of non-monotonic reasoning attempt to characterise defeasible reasoning. Reiter 1980; Bochman 2007; Koons 2017; Strasser and Antonelli 2018.

Missionaries and Cannibals

Another, well known problem in the literature of non-monotonic reasoning is the *Missionaries and Cannibals* problem. A description of this problem can be found in McCarthy 1981:

> *Three missionaries and tree cannibals come to a river. A row boat that seats two is available. If the cannibals out outnumber the missionaries on either bank of the river, the missionaries will be eaten. How shall they cross the river?*

When analysing the puzzle one can observe several issues. Firstly, the properties of the objects, within the riddle are not specified. That is, how would one

know if a boat is an object that enables the crossing of a river or why would crossing the river change the number of missionaries? Hence, before one can actually start an attempt towards solving the riddle, one has to presuppose a certain fraction of common knowledge.
Secondly, the formalisation of the puzzle is not precise enough. That is, the existence of a bridge, of a bigger boat, or a plane were not explicitly excluded. Similarly, no information about possible obstacles, such as the boat being not seaworthy or the existence of a river monster are given. Therefore, when formalising the problem, one would have to be incredibly careful. That is, an excruciatingly detailed list of immense scope excluding all possible exception has to be constructed.
A strong similarity between this example and the previous one, lies within the fact that both of them showcase the same kind of reasoning found employed by humans. That is, humans are able to assume certain things to be true, unless additional information is presented. For example, given the contexed of a riddle they do not need to be told, that there is no bridge, which can be used to circumvent the riddle. Hence, similarly as before, a formalism designed to capture such reasoning has to be able to account for abnormalities, while at the same time dismissing them, if they are not explicitly stated. That is, it has to allow the formalisation of normality in order to assume it as an default. McCarthy 1981; Bochman 2007.

Moreover, this example elucidates the issue of specifying negative information. As seen for example in statements such as, there is no bridge, there is no hole in the boat, there is no plane, and so on. The specification of negative information is among others difficult since, one has to be aware of every possible undesired information. Especially with the Missionary-Problem one would be well advised to simply assume that the information given is the only information that exists (apart from some common knowledge). Hence, in this particular case it would be desirable to have a formalism, which is able to prefer an interpretation of the world where only necessary assumption had been made, e.g. were positive information is minimised.

This issue is closely related to the issue of operating within an open world. Since, in an open world one has to assume that information available is incomplete. That is, in an open world, the absence of information does not imply its negation. For example, when issuing a query to a database, in order to check whether there exists a flight from Vienna to London on the 27^{th} of August, one is confronted with two possible results. Namely, in the positive case, the desired flight exists and in the negative case that there is no flight with the desired attributes found within the database. However, since it was assumed that only partial information is available, it is impossible to conclude that no such flight exists. This is due to the fact that, just because it is not contained within the database it does not mean that it does not exists. Unfortunately, this lack of negative information is an obstacle for when reasoning about the information contained within the database. Fortunately, in the context of databases is reasonable to assume that the database contains all relevant information. Therefore, similarly as in the Missionary-Problem, is is sensible to impose onto the formalism that positive information has to be minimised. Thus in order to avoid the precarious task of stating every negative fact, one simply assumes that, if a fact is not contained within the database, its negation must hold. This assumption is called the *Closed World Assumption* and will be addressed later. Brewka, Dix, and Konolige 1997

Frame Problem

The last problems discussed in this paper, which motivated the development of non-monotonic formalisms is the *frame problem.* It was initially proposed by McCarthy and Hayes 1981, and turned out to be a central problem in the field of artificial intelligence. At its core the frame problem is the problem of modelling actions in a changing world, i.e. a dynamic domain, without explicitly stating every property, which remains unaffected by said action. That is, how can one determine which things stay the same after an action is employed that invokes a change in the world. Therefore, it is closely related to another aspect of human reasoning, namely predictive reasoning, i.e. the type of reasoning employed in tasks such as planning. Bochman 2007; Lifschitz 2015
However, before discussing the frame problem further, some intuition should be developed. This hopefully will be accomplished by the follwoing examples, some of which are taken from Lifschitz 2015.

There exists Alice, Bob and Carol. Furthermore, it is known that Alice is in the room, while Bob and Carol are not. Moreover, Bob has the ability to enter the room. An ability, which in the given scenario, is employed directly after observing the initial state of the world. Given the knowledge that Bob used his ability to enter the room, it is reasonable to assume that both Alice and Bob are now in the room. Hence, it seems desirable that a formalisation of this situation, would allow the same inference. Therefore, in an attempt to do so, a description of the initial state of world, i.e. step 0 could look like

$$IN_0(Alice), \neg IN_0(Bob), \neg IN_0(Carol), ENTER(Bob)$$

Moreover, in order to know the state of the world in step 1, a formula modelling the common fact that entering a room, results of you being in the room has to be added, i.e.

$$\forall x \, (ENTER(x) \rightarrow IN_1(x))$$

Given, this one can now try to reason about the state of the world in step 1. Starting with the fact that is known $ENTER(Bob)$, one can infer that Bob is in the room $IN_1(Bob)$. However, this is the only possible inference. Therefore, the following is known about the world

$$IN_0(Alice), \neg IN_0(Bob), \neg IN_0(Carol), ENTER(Bob), IN_1(Bob)$$

Hence, in state 0 Alice is in the room, while Bob and Carol are not. However, in state 1 Bob is in the room, while Alice and Carol simply disappeared into nothingness. The reason, why this happened is that only change was modelled. That is, it was not specified what remains the same. Hence, in order to impose inertia onto the system, additional axioms have to be added, i.e.

$$\forall x\ (IN_0(x) \rightarrow IN_1(x))$$
$$\forall x\ ((\neg IN_0(x) \wedge \neg ENTER(x)) \rightarrow \neg IN_1(x))$$

This example, depicts the frame problem in a very concise manner. That is, how can inertia be modelled, without an incredible amount of such inertia axioms.
In order to grasp the amount of possible inertia axioms, one could think of all the things that happen in the world when the heart of a human beats once. What changes in the world? Well, the blood in in its veins moves, fresh oxygen is delivered to the subjects cells, some heat will be produced, and so on. However, a simple heart beat does not influence much in the world. Did the subjects heartbeat cause the sun to disappear or did it cause a bag of rice to vanish? Probably not. Hence, for any single action it has to be specified what changes and more importantly what does not. Therefore, if one intends to model state of the world similar to the approach presented above, an immense amount of axioms would be created. Lifschitz 2015; Bochman 2007.
Additionally, due to monotonicity every possible exception to a rule has to be also accounted for. This can be observed beautifully in the following problem, which was stated in Shanahan 2016.

$$\forall x\ (PAINT_n(x,c) \rightarrow COLOUR_{n+1}(x,c))$$
$$\forall x\ (MOVE_n(x,p) \rightarrow POSITION_{n+1}(x,p))$$

This formalisation simply states what changes in the world. Namely, if in step n something (x) is painted $(PAINT)$ the colour c, in step $n+1$ the object x will have the colour c. Similarly, if in step n something (x) is moved $(MOVE)$ to the position p, in step $n+1$ the object x will be at position p. Furthermore, inertia has to be modelled as well, i.e. what does not change in the world if $MOVE$ or $PAINT$ are employed. This is accomplished as follows

$$\forall x\ ((MOVE_n(x,p) \wedge COLOUR_n(x,c)) \rightarrow COLOUR_{n+1}(x,c))$$
$$\forall x\ ((PAINT_n(x,p) \wedge POSITION_n(x,p)) \rightarrow POSITION_{n+1}(x,p))$$

which ensures that paining an object does not change its position, as well as moving an object does not change its colour. However, what about the action of moving the object x into a bucket of paint. In this case by changing the position of the object, the colour of the object was changed, even though $PAINT$ was not applied. Furthermore, these are not the only exceptions, which have to be accounted for. That is, apart from having to account for cases where actions have unintended consequences, there are also exceptions which impose restrictions on the applicability of certain actions. That is, in this scenario one has to ensure that the object to be moved, can actually be moved, e.g. it is not glued to the floor. Therefore, in order to model this properly, we would have to extend our axioms such that the every single exception is accounted for. That is, we are yet again faced with a similar issue as presented in the Tweety problem and Missionaries-Problem. Shanahan 2016; Lifschitz 2015

One central approach for avoiding the excessive use of inertia axioms is the so called *inertia assumption*, i.e. it is simply assumed as a default, that unless explicitly stated everything remains the same. This inertia assumption is an important aspect of the frame problem, as it connects it to non-monotonic reasoning (similarities to the closed world assumption). Bochman 2007
However, while the frame problem is a central problem in the field of artificial intelligence it is actually, only a special case of the more general *temporal projection problem* problem, which encompasses the persistence, ramification and qualification problems. Bochman 2007.

- *The persistence problem / frame problem:* The persistence problem is a reformulation of the frame problem. That is, it is the general problem of predicting the properties that remain the same as actions are performed. If one desires an example, one could think about the curious case of the disappearing Alice as presented above. Bochman 2007

- *The ramification problem:* The general problem of predicting the properties that do change as actions are performed. Hereby, the difficulty lies within predicting the implicit changes that occur as a result of an action. For example, what changes when a bookshelf is moved? Well, the position of the bookshelf changes. However, this is only the direct effect, as among other things additional changes such as the position of the books on the shelf change, a different part of the floor will be covered, maybe access to the room will be prevented, and so on. Given the examples presented above, the ramification problem can be observed in the case of moving an object into a bucket of paint. Ginsberg and Smith 1987.

- *The qualification problem:* The general problem of predicting the properties that have to hold such that an action has its intended effects. For example, could something prevent the movement of the bookshelf? Is the bookshelf too heavy, to fragile?. Additionally, in the case of Carol from the example above. Lets assume that Carol wants to enter the room. However, what if there are conditions that prevent Carol from entering, are her legs broken, is she too big to get through the door or is the room already full? A similar situation arises with accounting for the fact that an object might be glued to the floor, which then prevents the movement of the object. One of the most glaring examples of this problem would be the starting of a car. Because, in order to do so the ignition must work, the battery is not empty, there must be fuel in the tank, the key must not be broken and so on. Ginsberg and Smith 1987; Bochman 2007.

In many instances of this problems it would be advantageous if one could simply make default assumption, which could be retracted if evidence to the contrary arises. That is, one would simply treat inertia as a default, or assume that it is possible to execute a certain action until there is evidence leading to a revision of ones assumptions.

Before moving on a few disclaimers are necessary. Namely, the problems described above where simply motivation factors for the development of non-monotonic formalism, thus it does not follow that these problems can only be accounted for by such formalisms. In fact the non-monotonic formalisms discussed below failed to solve the frame problem. See the Yale-Shooting problem. Moreover, there are also some monotonic approaches, that provide adequate solutions for the frame problem. Lastly, on a technical level for applicative purposes the frame problem is solved to a satisfying degree. Shanahan 2016 To conclude, this subsection attempted to provide some examples responsible for motivating the development of formalisms concerned with non-monotonic reasoning. Moreover, it attempted to provide an intuition for the kind of reasoning one would want to capture with such formalisms. The following subsection will discuss on a more formal level the difference between monotonic and non-monotonic reasoning.

Monotonicity

In order to understand what non-monotonic reasoning is, one is best advised to develop an understanding of monotonic reasoning. This will be accomplished by building an analogy to one instance of monotonicity, most people will be familiar, namely monotonicity with respect to functions over $\mathbb{R}$. This is defined as follows

Definition 1.1 (Citation!!!!). *Let f be a function s.t.*

$$f : \mathbb{R} \to \mathbb{R}$$

This function f is called monotonically increasing if and only if

$$\forall x, y \in \mathbb{R} : x \leq y \implies f(x) \leq f(y)$$

Informally this definition simply states that, if the input of the function is increased its output will either increase or remain the same.

A similar notion can be developed for the logical consequence in formal logical systems. In such a system one wants to be able to derive conclusions from a set of axioms, via a given set of inference rules. Furthermore, it can be desirable that, if a conclusion is obtained given a fixed set of axioms, then the conclusion should is still valid when the original set of axioms is expanded. That is, additional information can not invalidate previously inferred statementsMcDermott and Doyle 1980; Bochman 2007. On a syntactic level this notion is formalised in McDermott and Doyle 1980 as follows

Definition 1.2. *A logic is called monotonic, if the following statement holds.*

Let A and B be two theories, such that $A \subseteq B$ then $Th(A) \subseteq Th(B)$.

Furthermore, $Th(S)$ is defined as the set of sentences derivable from a given theory S by the means of inference provided by the given logic, i.e. $Th(S) = \{p \mid S \vdash p\}$.

That is, similar to an increasing monotonic function, if the input, in this case the set of axioms with respect to ⊆, is increased one can expect that the set of possible inferences, must either increase or remain the same.

Moreover, this notion can easily be applied to the syntactic inference. That is, similar to the definition found in McCarthy 1981 we can restate this as

Definition 1.3. *A logic is called monotonic if the following statement holds.*

Let A and B be two theories, such that $A \subseteq B$ then $Th(A) \subseteq Th(B)$.

Furthermore, $Th(S)$ is defined as the set of sentences that are true under every model of S, i.e. $Th(S) = \{p \mid S \vDash p\}$.

This means that, since $A \subseteq B$ every model of B must also be model of A. Therefore, $A \vDash p$ implies $B \vDash p$ for all models of B.

However, the presented notion of monotonicity is not the only one. That is, one can understand monotonicity as follows. If A implies C then one should be able to expand the statement to $A \wedge B$ imply C. As $A \wedge B$ can only be true, if A is true and since A implies C one must therefore be able to derive C. This notion is called *Strengthening the Antecedent*. As a reminder in classical logic strengthening of the antecedent is characterised by the following tautology.

$$(A \rightarrow C) \rightarrow (A \wedge B \rightarrow C)$$

This notion of monotonicity, is strongly connected with the inference relation. That is, if a system is constructed in a manner such that it permits strengthening of the antecedent, monotonicity is enforced onto the inference relation Bochman 2005.

In this paper the first notion of monotonicity is referred as *global monotonicity*, while the second on will be referred by *local monotonicity*. Both of which seem fairly desirable in a formal system. Especially, if one wants to capture mathematical reasoning. Unfortunately, as seen above reasoning in a monotone system has its shortcomings. For example, since humans seem to have significant problems, when it comes to mathematical reasoning, e.g. the difficulty of employing modus tollens in comparison to modus ponens. It suggests that when trying to capture certain parts of human reasoning in a formal manner the notion of monotonicity may have to be abandoned. Hence, providing the additional motivation for the field of artificial intelligence to develop such formal systems. Evans 2002; Bochman 2007.

1.2 Non-Monotonicity

After the motivation behind the development of formalism for non-monotonic reasoning has been presented, after an intuition for the kind of reasoning captured by such formalisms has been established and after the different notions of monotonicity have been formally introduced, one can now move on to introduce the various notions of non-monotonicity.

As already established non-monotonic reasoning tries, among others, to capture the notion of defeasible reasoning. That is, a form of formal reasoning were one is able to retract previously made inferences. Furthermore, since monotonicity requires that any previously made inference remains valid, even if the set of axioms is increased, it seems reasonable to assume that defeasible reasoning can not be subsumed by monotonic reasoning. As previously introduced there are are two notions of monotonicity. Therefore, by deciding which notions should no longer be enforced, one can obtain two different kinds of non-monotonicity. Namely, *Explanatory Non-Monotonic Reasoning* and *Preferential Non-Monotonic Reasoning* Strasser and Antonelli 2018; Bochman 2007.

On the one hand, the formalisms belonging to the *Explanatory Non-Monotonic Reasoning* approach adhere to global monotonicity. Moreover, they are characterised by the use an operator for restricting the set of possible models to only those models satisfying some closure conditions becoming fixpoints of this specific operator. Some of the formalism belonging to this approach are *Circumscription*, *Default Logic*, *Modal Non-Monotonic Logic* and *Auto-epistemic Logic* Bochman 2005; Brewka, Dix, and Konolige 1997.

On the other hand, the formalisms belonging to the *Preferential Non-Monotonic Reasoning* approach adhere to local monotonicity. Moreover, the are more concerned with the inference relation itself, i.e. they impose a preference relation onto the models and the inference relation upholds the stated preference. Some of the formalism belonging to this approach are *Circumscription*, *Closed World Assumption* and *Prefered Models* Bochman 2005; Brewka, Dix, and Konolige 1997.

While both approaches are distinct, they can be seen as theories of a reasoned use of assumptions. That is, assuming certain information by default, unless information arises, which "defeats" (contradicts) the information assumed by default. While explanatory approach is generally speaking more expressive, the preferential approach is allows Bochman 2005

2 Preferential Non-Monotonic Reasoning

One attempt of formalising defeasible reasoning resulted in the development of *Preferential Non-Monotonic Reasoning.* This mode of reasoning is non-monotonic in the sense that it abandons the concept of strengthening the antecedent. However, it is not necessarily non-monotonic with respect to the ability of expanding the set of axioms upon which inferences are made. That is, it is locally non-monotonic but not necessarily globally non-monotonic. Furthermore, preferenctial non-monotonic reasoning is historically speaking younger than *Explanatory Non-Monotonic Reasoning.* However, even though this approach was developed after the explanatory one, some of its core principals are already present in the introductory paper of Circumscription, i.e. McCarthy 1981 Brewka, Dix, and Konolige 1997; Bochman 2005.

Generally speaking, the formalisms subsumed by this approach are characterised, by the semantic notion of choosing which sets of models are preferred for a given theory. Hence, connecting this approach strongly to the inference relation, thus allowing a more abstracted view on how monotonic inference relation should behave. Therefore, connecting preferential non-monotonic reasoning closely with the study of inference relations and their properties on a meta-theoretical level. Moreover, it has been shown that in some cases it is possible to capture the semantic notion stating a preference for certain models, by means of defining an inference relation with certain properties. Bochman 2007; Brewka, Dix, and Konolige 1997

Lastly, the various formalisms that are subsumed by this approach, can be distinguished based on their level of abstraction and the method of how preference among models is expressed. Some of these formalisms are Brewka, Dix, and Konolige 1997

- *closed world assumption* a proof theoretic approach, where every non-provable ground term is assumed to be false;
- *circumscription* a model theoretic approach, where a second order formula is used to minimise the extensions of a specified set of predicates;
- *model preference logics* a generalisation of circumscription, which imposes a preference relationship onto the models.
- *conditional logics* is an approach, which relies on the a possible worlds relation, as well as on a preference relation between classes of models.
- *meta-theoretic approaches* this approach is similar to conditional logics. However, it uses a top-down approach, i.e. starts with specifying which properties a given inference relation should have.

The following subsections, are concerned with investigating these approaches.

2.1 Closed World Assumption

As already mentioned, the *Closed World Assumption* has its origin in world of database theory and has a strong connection with logic programming. Therefore, as already seen in the flight example, a database operating under the closed world assumption is assumed to be complete. That is, any information which is not contained within the database does not exist in the world. Hence, if a there

is no positive instance of a predicate in the database, its negation is assumed to hold.
For example, if there is no entry in the database expressing that *there is a flight from Vienna to London on the* 27^{th} *of August* then due to the completeness enforced by the closed world assumption, its negation is assumed, i.e. *there is no flight from Vienna to London on the* 27^{th} *of August.* Strasser and Antonelli 2018; Brewka, Dix, and Konolige 1997
Shifting away from traditional databases, towards deductive databases, one can characterise the closed world assumption in a more formal manner. For example, let T be a set of teachers, such that $T \coloneqq \{a, b, c, d\}$. Moreover, let $S \coloneqq \{A, B, C\}$ as set of students and let $TEACH(x, y)$ a predicate expressing whether teacher x teaches student y. Moreover, it is known

$$TEACH \coloneqq \{(a, A), (b, B), (c, C), (d, B)\}$$

If the query "Which teacher does not teach the student B?" is issued, the deductive database, thus tries to prove $\neg TEACH(x, B)$ for any $x \in T$. However, since it is not possible to derive $\neg TEACH(a, B)$ and $\neg TEACH(c, B)$ formal reasoning would fail to provide a proof for the issued query. Hence, one can express the closed world assumption as follows. If there is no proof of a positive ground predicate its negation is assumed Reiter 1981. Presented more formally,

Definition 2.1 (Brewka, Dix, and Konolige 1997)**.** *Let* $P(t)$ *be a ground predicate instance, then*

$$\mathit{CWA(DB)} \coloneqq DB \cup \{\neg P(t) \mid DB \not\models P(t)\}$$

Given this formalisation queries are not issued against DB but against *CWA(DB)* Bochman 2007. Hence, in the above example issuing the query would therefore result in $\{a, c\}$, as one would intuitively expect. Moreover, if one adds expands the database to DB', e.g.

$$TEACH \coloneqq \{(a, A), (b, B), (c, C), (d, B), (a, B)\}$$

CWA(DB') is computed based on the adapted database. However, since $DB' \models TEACH(a, B)$, the ground predicate $\neg TEACH(a, B)$ will not be added by *CWA*. Thus, the query returns $\{c\}$.
Here one can observe, why the closed world assumption is closely related to non-monotonic reasoning. As an expansion of the set of axioms, i.e. the database, resulted in a retraction of the assumption $\neg TEACH(a, B)$. Generally speaking, the closed world assumption already captures a decent fraction of non-monotonic reasoning. However, there are certain well known limitations, namely the implicit requirements of a finite domain and the requirement of unique names Bochman 2007; Brewka, Dix, and Konolige 1997.

Firstly, the closed world assumption implicitly includes the *unique names assumption (UNA)*. That is, in first order logic it is possible, that two syntactically different ground terms reference the same individuals. In the the teacher example, teacher c and teacher d could be the same individual i.e. $c = d$. This notion is formalised by

$$UNA : \{t_i \neq t_j \mid t_i \textit{ and } t_j \textit{ are different ground terms } \}$$

Secondly, the closed world assumption implicitly includes the *domain closure assumption (DCA)*. That is, it assumes the number of individuals in the domain to be finite. An assumption, which is perfectly justified, when operating on a more practical level, such as it is the case with databases. This assumption can be expressed as

$$DCA : \{\forall x\, t_1 = x \vee t_2 = x \vee \cdots \vee t_n = x\}$$

Thirdly, there is another issue with the closes world assumption. Namely, the issue of consistency. That is, how can one ensure that a consistent database DB remains consistent after CWA is applied. For example, let DB be a database which does contain the statement $A \vee B$, but does not contain A and B. If one now applies the strong closure operation CWA, i.e. $CWA(DB)$, one would obtain $\neg A$ and $\neg B$, from which one can infer

$$\begin{aligned} CWA(DB) \vDash \neg A \textit{ and } CWA(DB) \vDash \neg B &\iff CWA(DB) \vDash \neg A \wedge \neg B \\ &\iff CWA(DB) \vDash \neg(A \vee B) \end{aligned}$$

Hence, due to $CWA(DB) \vDash \neg(A \vee B)$ and $CWA(DB) \vDash (A \vee B)$, the database becomes inconsistent. However, such issues only arise, if the database does not possess a least model. Therefore, such inconsistencies can be avoided, if it can be ensured that the database contains a least model, i.e. if the intersection of all models of the database is itelf also a model of the database. Fortunately, this is accomplished by transforming the database into conjunctive normal form and checking, if the clauses contain at most one positive literal. Formally, if the database can be transformed into Horn Clauses, i.e. if the database is definite.Bochman 2007; Brewka, Dix, and Konolige 1997; Reiter 1981

2.2 Circumscription

Circumscription, first proposed in McCarthy 1981, is an augmentation of first order logics, which should allow for non-monotonic reasoning. Even though, it was initially seen as a method for adding additional formulas to a theory, it can be equally seen as a method of restricting the models of a theory. This restriction is based on the notion of minimisation. Possible forms of minimisation could be the restriction to models which are minimal in their extension of some predicates or the restriction to models which are minimal in their domain. Naturally, this implies that there must be several different versions of circumscription and indeed there are. For example, domain circumscription, parallel predicate circumscription, prioritised circumscription, pointwise circumscription and scoped circumscription. Among all of the listed variants, the first form of circumscription is *Domain Circumscription*, which was followed up by the more expressive *Predicate Circumscription* Bochman 2007; Brewka, Dix, and Konolige 1997.

However, reducing circumscription to minimisation, is somewhat misleading. That is, in order to capture non-monotonic reasoing, the approach proposed by McCarthy also heavily relied on *abnormality theory*, which uses abnormality predicates to enable the modelling of default knowledge. Therefore, after the introduction of predicate circumscription, abnormality theory will briefly be discussed. Bochman 2007 Lastly, some of the problems with circumscription will be presented.

2.2.1 Predicate Circumscription

Before moving into the technical details of predicate circumscription, a quick example to hone the intuition. In order to do so a slight modification on the teacher example presented in the subsection concerned with the closed world assumption is used.

Let $T(x)$ be the predicate expressing that x is a teacher. Moreover, let $S(x)$ be the predicate expressing that x is a student. Lastly, let $TEACH(x, y)$ be a predicate expressing that teacher x teaches student y. Furthermore, it is known that

$$\begin{aligned}&S(A) \wedge S(B) \wedge S(C)\\ &\wedge T(a) \wedge T(b) \wedge T(c) \wedge T(d)\\ &\wedge TEACH(a, A) \wedge TEACH(b, B) \wedge TEACH(c, C) \wedge TEACH(d, B)\end{aligned}$$

This example is only concerned with the predicate S, thus the only relevant statement is $S(A) \wedge S(B) \wedge S(C)$. The intent is to express that there are only as many students as required by the previous statement. However, nothing prohibits interpretations, such as

1. Let $\mathcal{M}_1 := (D_{\mathcal{M}_1}, I_{\mathcal{M}_1})$ an interpretation where

   $$I_{\mathcal{M}_1}(A) = I_{\mathcal{M}_1}(B) = I_{\mathcal{M}_1}(C) = \delta \in D_{\mathcal{M}_1}$$

   Moreover, $I_{\mathcal{M}_1}(S) := \{\delta\}$

2. Let $\mathcal{M}_2 := (D_{\mathcal{M}_2}, I_{\mathcal{M}_2})$ an interpretation where

   $$\begin{aligned}I_{\mathcal{M}_2}(A) &= \delta \in D_{\mathcal{M}_2}\\ I_{\mathcal{M}_2}(B) &= \sigma \in D_{\mathcal{M}_2}\\ I_{\mathcal{M}_2}(C) &= \eta \in D_{\mathcal{M}_2}\end{aligned}$$

Moreover, $I_{\mathcal{M}_2}(S) \coloneqq \{\delta, \sigma, \eta\}$

3. Let $\mathcal{M}_3 \coloneqq (D_{\mathcal{M}_3}, I_{\mathcal{M}_3})$ an interpretation where

$$\begin{aligned} I_{\mathcal{M}_3}(A) &= \delta \in D_{\mathcal{M}_3} \\ I_{\mathcal{M}_3}(B) &= \sigma \in D_{\mathcal{M}_3} \\ I_{\mathcal{M}_3}(C) &= \eta \in D_{\mathcal{M}_3} \end{aligned}$$

Moreover, $I_{\mathcal{M}_2}(S) \coloneqq \{\delta, \sigma, \eta, \gamma\}$

(Note: It has to be mentioned that the definition of the models are only restricted to the parts required in the example. Moreover, apart from the definitions above, the considered models are identical.)
With a brief glance at the models, one can easily detect that it is possible to define an ordering between the models, based on the number of individuals, i.e. elements in $D_{\mathcal{M}_x}$, which satisfy S. That is, for $\mathcal{M}_1$ there is only one element satisfying P, while for $\mathcal{M}_2$ there are three and for $\mathcal{M}_3$ there are four. Hence,

$$\mathcal{M}_1 \leq \mathcal{M}_2 \leq \mathcal{M}_3$$

Having this relation in mind, if one would apply circumscription to the predicate S, it would select the model $\mathcal{M}_1$ since it is minimal with respect to the extensions of S. On a syntactical level, circumscription would enforce the desired model by adding additional formulas, which require $\mathcal{M}_1$ in order to be satisfied.
By contrast, the closed world assumption would prefer the model $\mathcal{M}_2$, due to the implicit unique name assumption. Incidentally, if one would use UNA in combination with circumscription one would also obtain $\mathcal{M}_2$ as a preferred model. Because of $S(A) \wedge S(B) \wedge S(C)$ and $A \neq B \neq C$, there would be at least three elements required in $I_{\mathcal{M}_x}(S)$, thus resulting in $\mathcal{M}_2$ being the smallest possible model among those three.
This connection with the closed world assumption, can be formalised as such

Theorem 2.2 (Brewka, Dix, and Konolige 1997)**.** *We have:* $[CWA(A) \wedge DCA] \equiv [A \wedge UNA \wedge DCA]$ *provided that* $CWA(A)$ *is consistent*

That is, if one restricts circumscription to finite domains with assuming that unique names refer to unique elements one can simulate the closed world assumption.

After this brief introduction, the next setp is to formalise the notion of circumscription. This is done by presenting the initial version of predicate circumscription, which was introduced in the paper McCarthy 1981 and can be formalised as follows.

Definition 2.3 (McCarthy 1981)**.** *Let A be a sentence of first order logic containing a predicate symbol $P(x_1, \ldots, x_n)$, written as $P(\overline{x})$. Let $A(\Phi)$ stand for* *the result of replacing all occurrences of P in A by the predicate expression Φ. Then the schema*

$$\forall \Phi \; (A(\Phi) \wedge \forall \overline{x} \; (\Phi(\overline{x}) \rightarrow P(\overline{x}))) \rightarrow \forall \overline{x} \; (P(\overline{x}) \rightarrow \Phi(\overline{x}))$$

is called the circumscription of P.

Firstly, the sentence given in the definition, is a second order sentence, because it quantifies over predicates, i.e. $\forall \Phi$. Secondly, $A(\Phi)$ expresses that the predicate Φ, which replaces the predicate P must satisfy all the condition satisfied by P. Thirdly, $\forall \overline{x} \; (\Phi(\overline{x}) \rightarrow P(\overline{x}))$ requires that the set of tuples satisfying Φ must also satisfy P. That is

$$\{\overline{x} \mid \forall \overline{x} \; \Phi(\overline{x})\} \subseteq \{\overline{x} \mid \forall \overline{x} \; P(\overline{x})\}$$

Lastly, since the whole statement is an implication, one can infer that Φ satisfies both conditions, it must be the case that the set of tuples satisfying P must also satisfy Φ. Hence,

$$\{\overline{x} \mid \forall \overline{x} \; \Phi(\overline{x})\} = \{\overline{x} \mid \forall \overline{x} \; P(\overline{x})\}$$

Even though the formalisation was explained, at least an example is necessary to obtain a more intuitive understanding. Moreover, this example shall provide an argument for the non-monotonic character of this formalism.
Let Γ be the sentence

$$\Gamma = isBlock(A) \wedge isBlock(B) \wedge isBlock(C)$$

which states that A,B and C are blocks. Here, *isBlock* is circumscribed. That is, using the formula ($\forall \Phi$ can be dropped since a specific Φ will be chosen later.)

$$(\Gamma(\Phi) \wedge \forall \overline{x} \; (\Phi(\overline{x}) \rightarrow P(\overline{x}))) \rightarrow \forall \overline{x} \; (P(\overline{x}) \rightarrow \Phi(\overline{x}))$$

and replacing P with *isBlock* and Γ with its definition as well as replacing the occurrences of *isBlock* in Γ with Φ, one obtains the schema

$$\Phi(A) \wedge \Phi(B) \wedge \Phi(C) \wedge \forall \overline{x} \; (\Phi(\overline{x}) \rightarrow isBlock(\overline{x}))) \rightarrow \forall \overline{x} \; (isBlock(\overline{x}) \rightarrow \Phi(\overline{x}))$$

Now it is necessary to specify a Φ, such that it satisfies the required conditions. In this case

$$\Phi(x) = (x = A \vee x = B \vee x = C)$$

First, Φ is substituted with its definition, thus

$$\begin{aligned} &(A = A \vee A = B \vee A = C) \wedge (B = A \vee B = B \vee B = C) \wedge (C = A \vee C = B \vee C = C) \\ &\wedge \forall \overline{x} \; ((\overline{x} = A \vee \overline{x} = B \vee \overline{x} = C) \rightarrow isBlock(\overline{x}))) \\ &\rightarrow \forall \overline{x} \; (isBlock(\overline{x}) \rightarrow (\overline{x} = A \vee \overline{x} = B \vee \overline{x} = C)) \end{aligned}$$

is obtained. Firstly,

$$(A = A \vee A = B \vee A = C) \wedge (B = A \vee B = B \vee B = C) \wedge (C = A \vee C = B \vee C = C)$$

is a tautology. Secondly,

$$\forall \overline{x}\ ((\overline{x} = A \vee \overline{x} = B \vee \overline{x} = C) \rightarrow isBlock(\overline{x})))$$

also holds as $isBlock(A)$, $isBlock(B)$ and $isBlock(C)$ is asserted by Γ and due to the fact that $(\overline{x} = A \vee \overline{x} = B \vee \overline{x} = C)$ only holds if $\overline{x}$ is either A, B or C. Hence, one can use the circumscription sentence to derive

$$\forall \overline{x}\ (isBlock(\overline{x}) \rightarrow (\overline{x} = A \vee \overline{x} = B \vee \overline{x} = C))$$

Thus, the possible extension of the circumscribed predicate $isBlock$ are minimised, i.e. $isBlock$ only holds in the necessary cases, which in this case are A, B and C as asserted by Γ. However, it is important to note that A, B and C must not be unique, i.e. the case $A = B = C$ is not excluded. Moreover, in order to observe the non-monotonic behaviour one has to simply add $isBlock(D)$ to Γ. That is,

$$\Gamma' = isBlock(A) \wedge isBlock(B) \wedge isBlock(C) \wedge isBlock(D)$$

which would, given the previously selected Φ result in

$$\begin{aligned}
&(A = A \vee A = B \vee A = C) \wedge (B = A \vee B = B \vee B = C)\\
&\wedge (C = A \vee C = B \vee C = C) \wedge (D = A \vee D = B \vee D = C)\\
&\wedge \forall \overline{x}\ ((\overline{x} = A \vee \overline{x} = B \vee \overline{x} = C) \rightarrow isBlock(\overline{x})))\\
&\rightarrow \forall \overline{x}\ (isBlock(\overline{x}) \rightarrow (\overline{x} = A \vee \overline{x} = B \vee \overline{x} = C))
\end{aligned}$$

which, due to $(D = A \vee D = B \vee D = C)$ blocks the derivation of

$$\forall \overline{x}\ (isBlock(\overline{x}) \rightarrow (\overline{x} = A \vee \overline{x} = B \vee \overline{x} = C)$$

Hence, an expansion of the theory resulted in the retraction of inferences, thus a counterexample against monotonicity has been provided.
As already stated this from of predicate circumscription is not the only one. Moreover, even though it was the first one, it is not the most popular or the most useful. This honour belongs according to Brewka, Dix, and Konolige 1997 to *Parallel Predicate Circumscription*, which will be discussed next.

2.2.2 Parallel Predicate Circumscription

The previous variant of circumscription, was the first predicate circumscription introduced. It circumscribes only one predicate by means of the adding of formulas, which enforce a certain class of models, i.e. models with minimal extension of predicates. Incidentally, this is the one of the possible approaches to circumscription, namely a method for adding additional formulas to the theory. Hence, *paralell predicate circumscription*, the benefit of which is that it allows one to circumscribe several predicates in parallel, will be additionally introduced along the lines of the second approach to circumscription, which is the more semantic notion of restricting the class of models. Bochman 2007; Brewka, Dix, and Konolige 1997

Let $\mathcal{L}$ be a language and let $P := (P_1, P_2, \ldots P_n)$ and $Z := (Z_1, Z_2, \ldots Z_m)$ be tuples of predicate symbols, such that P contains all predicates that are to be minimised in parallel and Z contains the predicates that are allowed to vary across compared models.
Firstly, one has to define a preference relation, on the basis of which P will be minimised. Therefore, let $A(P;Z)$ be a first-order sentence, containing among others P and Z. A model M_1 is preferred over a model M_2, if M_1 is smaller with respect to predicate P and if both of them have the same interpretation for the so called *fixed symbols*, i.e. symbols of A not in P or Z. This ordering relation can be described as follows

Definition 2.4 (Brewka, Dix, and Konolige 1997)**.** *Let M_1 and M_2 be models of $A(P;Z)$. Let $|M|$ be the universe of M and $M[\![K]\!]$, the interpretation of K in M. Lastly, let $M_1 \leq^{P;Z} M_2$ such that*

$$M_1 \leq^{P;Z} M_2 \iff \begin{cases} |M_1| = |M_2|; & \\ M_1[\![K]\!] = M_2[\![K]\!] & \textit{for all } K \textit{ not in } P \textit{ or } Z; \\ M_1[\![P_i]\!] \subseteq M_2[\![P_i]\!] & \textit{for all } P_i \textit{ in } P; \end{cases}$$

The relation specified by this definition, is a preorder relation, i.e. reflexive, transitive and not necessarily antisymmetric. From there one can define a strict version of this relation, which is defined as

Definition 2.5 (Brewka, Dix, and Konolige 1997)**.**

$$M_1 <^{P;Z} M_2 \iff M_1 \leq^{P;Z} M_2 \textit{ and not } M_2 \leq^{P;Z} M_1$$

Given this relation one can express a minimality. That is, a model M of $A(P;Z)$ is minimal with respect to $\leq^{P;Z}$ if there is no model M' of $A(P;Z)$ such that $M' < M$. Brewka, Dix, and Konolige 1997; Lifschitz 1996

The above described semantic notion of the smallest model, can be expressed with the following second-order formula. However, before doing so some notational conventions have to be introduced. That is, let A and B predicates of the same arity.

$$\begin{aligned} A \leq B &\iff \forall \overline{x}\, A(\overline{x}) \to B(\overline{x}) \\ A < B &\iff A \leq B \land \neg(B \leq A) \\ A = B &\iff A \leq B \land B \leq A \end{aligned}$$

In a more set theoretical depiction one can understand this as

$$\begin{aligned} A \leq B &\iff \{\overline{x} \mid \forall \overline{x}\ A(\overline{x})\} \subseteq \{\overline{x} \mid \forall \overline{x}\ B(\overline{x})\} \\ A < B &\iff \{\overline{x} \mid \forall \overline{x}\ A(\overline{x})\} \subset \{\overline{x} \mid \forall \overline{x}\ B(\overline{x})\} \\ A = B &\iff \{\overline{x} \mid \forall \overline{x}\ A(\overline{x})\} = \{\overline{x} \mid \forall \overline{x}\ B(\overline{x})\} \end{aligned}$$

However, this definition can not be applied directly to the tuple of predicates P and Z. Hence, it has to be generalised such that it accommodates more than one predicate. That is, for $A := (A_1, A_2, \ldots A_n)$ and $B := (B_1, B_2, \ldots B_n)$

$$\begin{aligned} A \leq B &\iff \forall i\ A_i \leq B_i \\ A < B &\iff A \leq B \wedge \neg(B \leq A) \\ A = B &\iff A \leq B \wedge B \leq A \end{aligned}$$

as formulated in Lifschitz 1996 Given this, the second-order formula defining parallel predicate circumscription can be introduced. That is,

Definition 2.6 (Lifschitz 1996; Brewka, Dix, and Konolige 1997)**.** *Let Circum be a second order formula, such that*

$$CIRC(A(P;Z);P;Z) = A(P;Z) \wedge \neg(\exists \overline{p} \exists \overline{z}\ (A(\overline{p};\overline{z}) \wedge \overline{p} < P)$$

where $\overline{p} := (p_1, p_2, \ldots, p_n)$ *and* p_i *is a predicate variable with the same arity as* $P_i \in P$ *with* $P := (P_1, P_2, \ldots, P_n)$*. Analogous for* $\overline{z}$ *and* Z*.*

The above formula, can be understood as follows. Firstly, $A(P;Z)$ requires any model of $CIRC(A(P;Z);P;Z)$ to be also model of $A(P;Z)$. Secondly, $\neg(\exists \overline{p} \exists \overline{z}\ (A(\overline{p};\overline{z}) \wedge \overline{p} < P))$ that the selected model for $CIRC(A(P;Z);P;Z)$ has to be minimal in its extensions of P. That is, there exists no $\overline{p}$ and no $\overline{z}$ satisfying A, such that any $p_i \in \overline{p}$ is smaller in its extension than the corresponding $P_i \in P$. It can be shown that a model for $CIRC(A(P;Z);P;Z)$, will be minimal with respect to the relation $\leq^{P;Z}$, if it satisfies the formula. Lifschitz 1996

In order to improve the understanding of parallel predicate circumscription, a small example could be helpful. Hence, consider $CIRC(A(P;Z);P;Z)$ with $P = \{P_1, P_2\}$, $Z = \varnothing$ and $A(P;Z) = A(P) = A(P_1, P_2) = P_1(a) \wedge P_2(b)$. Hence, if plugged into $CIRC$ one obtains

$$\begin{aligned} CIRC(A(P;Z);P;Z) &= CIRC(P_1 \wedge P_2; P_1, P_2) \\ &= A(P;Z) \wedge \neg(\exists \overline{p} \exists \overline{z}\ (A(\overline{p};\overline{z}) \wedge \overline{p} < P) \\ &= A(P_1, P_2) \wedge \neg(\exists (p_1, p_2)\ (A(p_1, p_2) \wedge (p_1 < P_1 \wedge p_2 < P_2)) \\ &= P_1(a) \wedge P_2(b) \wedge \neg(\exists p_1 \exists p_2\ (p_1(a) \wedge p_2(b)) \wedge (p_1 < P_1 \wedge p_2 < P_2)) \end{aligned}$$

Notice that since $Z = \varnothing$ it does no longer occur in the formula. After the circumscription one obtains for example

$$\forall x\ (P_1(x) \leftrightarrow x = a) \wedge \forall x\ (P_2(x) \leftrightarrow x = b)$$

This is due to the fact that if $CIRC(P_1 \wedge P_2; P_1, P_2)$ is satisfied, then P_1 and P_2 must be minimal in their extensions. That is, ideally there is no element x in the domain such that $P_1(x)$ and no element y such that $P_2(y)$. However, since a model of $CIRC(P_1 \wedge P_2; P_1, P_2)$ must also be a model of $P_1(a) \wedge P_2(b)$ one is forced to consider models that satisfy at least $P_1(a)$ and $P_2(b)$. Thus in order to be minimal one should have a model in which a (b) is the only element for which P_1 (P_2) is true. Hence, in order to enforce such a behaviour, circumscription provides the formula $P_1(x) \leftrightarrow x = a$. This formula, expresses that any element in the domain, which satisfies P_1 must be the element a. Therefore, any model satisfying this formula must be minimal in its extensions of P_1 (analogous for P_2 and b).

Having now introduced the possibility of circumscribing several predicates at once, one can now move on to discussing *Abnormality Theory.*

Abnormality Theory

Abnormality theory, was initially introduced in order to allow a form of default reasoning within the circumscription paradigm. That is, a certain class of predicates is introduced, which contains so called *abnormality predicates*. These abnormality predicates denoted by ab, are designed to model statements such as

The boat can be used to cross the river, unless something is wrong.

Here the *unless something is wrong* can be understood as *unless there is no abnormality.* The idea behind this approach is to specify abnormality for certain aspects of an object, e.g. $Penguin$ is an abnormal bird with respect to Fly, by utilising the abnormality predicates, i.e. ab_x, which are then are circumscribed. Since, there can be multiple abnormalities parallel predicate circumscription is here very helpful. The nature of circumscription then ensures that abnormalities will be minimised in their extension. That is, unless stated otherwise the object in question is assumed to be normal. Bochman 2007; Brewka, Dix, and Konolige 1997

To observe this behaviour in practice, a small example, based on the Tweety-Problem. That is, one approach for formalising *normally birds fly* could be

$$\forall x\ Bird(x) \wedge \neg ab_1(x) \rightarrow Fly(x)$$

Here $ab_1(x)$ refers to element x being abnormal, if it can not fly. If one now applies circumscription

$$CIRC(Bird(Tweety) \wedge (\forall x\ Bird(x) \wedge \neg ab_1(x) \rightarrow Fly(x)); ab_1; Fly, Bird)$$

One would be able to infer $Fly(Tweety)$, because the smallest possible extension of ab_1 will be chosen. That is, since it is not required that $ab_1(Tweety)$, having a model with $ab_1(Tweety)$ would be bigger that one without $ab_1(Tweety)$, thus the one with $\neg ab_1(Tweety)$ is chosen. Here one can beautifully observe that there are two possible methods of invalidating the default reasoning modelled above. That is, one could add $Bird(Tweety) \wedge \neg Fly(x)$ instead of $Bird(Tweety)$, thus directly contradicting the assumption, which allows the conclusion of $ab_1(Tweety)$. On the other hand one could indirectly invalidate the rule, by asserting $Bird(Tweety) \wedge ab_1(x)$ instead of $Bird(Tweety)$, thus blocking the application of the rule, without asserting that $Tweety$ does not fly Brewka, Dix, and Konolige 1997.

Preferential Circumscription

Other forms of Circumscription

Due to the previously introduced abnormality theories various problems within circumscription become apparent. One of which is the problem of specificity. That is, the problem of how to deal with conflicting defaults.

2.3 Meta-theoretical Approach

One approach prevalent in preferential non-monotonic reasoning is the meta-theoretic approach. That is, instead of designing a specific formalism one focuses on the properties of the inference relation. In doing so two possibilities emerge. Firstly, one can investigate the inference relation of specific formalisms and then categorise them based on their properties. Secondly, one can start by specifying what properties an inference relations should have and then investigate their reasoning capabilities. With regard to non-monotonicity, this enables the categorisation of specific formalisms into families, thus providing a better understanding of what non-monotonicity actually means, if expressed in positive terms, i.e. not just as the omission of monotonicity. Moreover, specifying the behaviour of the consequence relation is independent of the underlying logic, thus allowing for a convenient way of choosing the underlying logic based on the problems at hand. Bochman 2007; Kraus, Lehmann, and Magidor 1990; Brewka, Dix, and Konolige 1997

Lastly it has to be noted that even though, classifying the inference relation of specific formalisms based on certain meta-theoretical properties is not restricted to either approach of non-monotonic reasoning, its application is more beneficial to the formalisms subsumed by the preferential approach. This is, among others because the notion of imposing preferences onto the set of possible interpretations corresponds, as shown in Kraus, Lehmann, and Magidor 1990, nicely with certain inference relations, which can be described by a set of meta-theoretical rules. By contrast, some formalisms subsumed by the explanatory approach barely restrict their inference relations. For example, as discussed below, default logic does not even satisfy cumulativity Brewka, Dix, and Konolige 1997; Antoniou and Wang 2007; Kraus, Lehmann, and Magidor 1990.

2.3.1 Semantics of Preferential Non-Monotonic Reasoning

As already stated the core concept of this incarnation of non-monotonic reasoning is stating, which models of a theory should be preferred. One method of accomplishing this is by imposing a preference relation upon all possible models. This results in the definition of model preference logic.

Definition 2.7 (Bochman 2007; Brewka, Dix, and Konolige 1997)**.** *Let $\mathcal{L}$ be a language and let $\prec$ be a well-founded ordering on the set of interpretations. $\mathcal{L}_{\prec}$ is called a model preference logic if the following holds. An interpretation $\mathcal{I}$ is a preferred model of A if it satisfies A and there is no better interpretation $\mathcal{J} \prec \mathcal{I}$ satisfying A. A preferentially entails B (written $A \vDash_{\prec} B$) iff all preferred models of A satisfy B.*

How this preference is established is in general not relevant. For example, as it is the case with circumscription one possible approach is specifying a preference based on the number of certain positive facts. However, in order to build a model preferential logic this relation must be well-founded, i.e. there is no infinite chain of models such as $\ldots \prec \mathcal{I}_3 \prec \mathcal{I}_2 \prec \mathcal{I}_1$.Bochman 2007; Brewka, Dix, and Konolige 1997

While both circumscription as well as the closed world assumption are such model preferential logics. It is still possible to further abstract this concept, in order to obtain the notion of *preferential models.* These were originally defined in Kraus, Lehmann, and Magidor 1990, as a model theoretic counterpart to the preferential consequence relation. That is,

Definition 2.8 (Kraus, Lehmann, and Magidor 1990)**.** *A preferential model W is a triple $\langle S, l, \prec \rangle$ where S is a set, the elements of which will be called states, $l : S \to \mathcal{U}$ assigns an interpretation to each state and $\prec$ is a strict partial order on S (i.e. an irreflexive, transitive relation), satisfying the* ***smoothness*** *condition.*

Here the strict partial order is not entirely necessary, however, models fulfilling this condition behave nicer. Moreover, the smoothness condition is mainly required in order to deal with infinite sets of formulas. Lastly, in order to define the smoothness condition the following definitions are required.

Definition 2.9 (Kraus, Lehmann, and Magidor 1990)**.** *Let $P \subseteq U$ be a binary relation on U. P is smooth iff $\forall t \in P$, either $\exists s$ minimal in P, such that $s \prec t$ or t is minimal in P.*

defines smoothness in general, while defines the smoothness condition

Definition 2.10 (Kraus, Lehmann, and Magidor 1990)**.** *A triple $\langle S, l \prec \rangle$ satisfies the smoothness condition iff $\forall \alpha \in \mathcal{L}$ the set $\widehat{\alpha} := \{s \in S \mid \forall m \in l(s) m \vDash \alpha\}$ is smooth.*

This condition expresses that for any formula α the set states in which every interpretation m associated to s via l does satisfy α must be smooth.

Given this definition one can easily observe that model preference logic is a special case of preference models. That is, one simply defines the set of states S as the set of all possible interpretation, defines l as the identity function and defines $<$ such that it is well-founded. Preference Models are especially relevant as they beautifully correspond with the preference inference relation defined below Kraus, Lehmann, and Magidor 1990.

Axiomatisation of Preferential Non-Monotonic Reasoning

The idea of studying the deduction process by means of investigating the consequence relation itself, can be traced back to Tarski and Gentzen. While the prior studied the consequence with regard to arbitrary sets, the latter one restricted himself to finite sets. Fortunately, due to compactness the difference between these two approaches is small for monotonic consequence relations. Unfortunately, this does is not the case, if monotonicity is rejected. However, since the basis of this subsection is Kraus, Lehmann, and Magidor 1990, only consequence relations with finite sets are considered. In Kraus, Lehmann, and Magidor 1990 several consequence relations are introduced, namely *Cumulative Reasoning, Cumulative Reasoning with Loop, Preferential Reasoning* and *Cumulative Monotonic Reasoning.* However, other systems have been developed, e.g. *Rational Reasoning.* Each of this systems can be characterised by by a certain set of rules, which specify the behaviour of the inference relation. Here the the system **C** (Cumulative Reasoning) and **P** (Preferential Reasoning) will be introduced. Before these systems can be formalised, some preliminary definitions must be made, as well as possible properties of inference relations have to be introduced.

That is, Let $\mathcal{L}$ be the set of all well-formed formulas closed under classical propositional connectives . Let $\vDash$ is defined as in propositional logic. We have for $\varphi, \psi, \chi \in \mathcal{L}$ the following poperties of consequence relation as presented in Kraus, Lehmann, and Magidor 1990

Reflexivity:
Is a notion satisfied by nearly every inference system.

$$\frac{}{\varphi \mathrel{|\!\sim} \varphi}$$

Left Logical Equivalence:
If two formulas are equivalent given the underlying logic, their consequences ought to be the same. Hence, expressing that consequences should depend on the meaning of a formula and not its form.

$$\frac{\vDash \varphi \leftrightarrow \psi \qquad \varphi \mathrel{|\!\sim} \chi}{\psi \mathrel{|\!\sim} \chi}$$

Right Weakening:
If φ is a plausible consequence of χ, i.e. $\chi \mathrel{|\!\sim} \varphi$ and given the underlying logic ψ

is a logical consequence of φ, i.e. $\vDash \varphi \to \psi$ then is is reasonable to require ψ to be also a logical consequence of χ, i.e. $\chi \mid\sim \psi$.

$$\frac{\vDash \varphi \to \psi \qquad \chi \mid\sim \varphi}{\chi \mid\sim \psi}$$

Cut:
This rule states that if χ is the consequence of an extended set of facts $\varphi \wedge \psi$, and ψ is a consequence of the non-extended set of facts, i.e. φ , χ should be a consequence of the non-extended set of facts, i.e. φ. It was accepted since it does not imply monotonicity and because formalisms such as circumscription already confine to cut.

$$\frac{\varphi \wedge \psi \mid\sim \chi \qquad \varphi \mid\sim \psi}{\varphi \mid\sim \chi}$$

Cautious Monotonicity:
If ψ and χ are both plausible consequences of φ then it should be possible to extend the original set of facts to $\varphi \wedge \psi$, while still retaining the previous consequence χ. That is, learning ψ, if ψ was already a plausible consequence, previous conclusions should not be invalidated. Moreover, this formalisation ensures that the additional formula ψ does not lead to a contradiction of φ.

$$\frac{\varphi \mid\sim \psi \qquad \varphi \mid\sim \chi}{\varphi \wedge \psi \mid\sim \chi}$$

Or:
If χ is separately a plausible consequence of two formulas, then it should also be a plausible consequence of their disjunction. This rule was accepted because it does not imply monotonicity.

$$\frac{\varphi \mid\sim \chi \qquad \psi \mid\sim \chi}{\varphi \vee \psi \mid\sim \chi}$$

Given these definitions the system **C** can be defined as follows.

Definition 2.11 (Kraus, Lehmann, and Magidor 1990)**.** *The system **P** consists of **Reflexivity**, **Left Logical Equivalence**, **Right Weakening**, **Cut**, **Cautious Monotonicity***

The system **P** is build upon system **C** and can be defined as follows.

Definition 2.12 (Kraus, Lehmann, and Magidor 1990)**.** *The system **P** consists of **Reflexivity**, **Left Logical Equivalence**, **Right Weakening**, **Cut**, **Cautious Monotonicity** and **Or***

That is, adding **Or** to system **C** will result in **P**. Frankly, all previously mentioned systems use the consequence relation of cumulative reasoning as a basis upon for constructing their own consequence relation. Moreover, if the underlying logic chosen is classical logic the system **C** is *Supraclassical*. That is, any inference made with respect to the rules of inference in classical logic, can also be made in **C** (and its extensions). This can be written concisely as

$$\frac{\varphi \vdash \psi}{\varphi \mathrel{|\!\sim} \psi}$$

Based on the rules characterising the respective system one can now start to build a connection with the model theoretical notions introduced previously, thus implicitly categorising specific formalisms such as circumscription or the closed world assumption. Kraus, Lehmann, and Magidor 1990

Firstly, it has to be noted that system **C** is weaker than **P**. That is, any inference possible by means of $\mathrel{|\!\sim}_{\mathbf{C}}$ can also be made by means of $\mathrel{|\!\sim}_{\mathbf{P}}$. One method of observing this is by looking at the model theoretical notions of model preference logic and preferred models. That is, if the preference relation is well-founded the following can be shown.

Theorem 2.13 (Kraus, Lehmann, and Magidor 1990)**.** *Let $\mathcal{L}_{<}$ be a preference logic, with $<$ being well-founded then $\mathcal{L}_{<}$ is **cumulative**.*

Theorem 2.14 (Kraus, Lehmann, and Magidor 1990)**.** *Let $\mathcal{L}_{<}$ be a preference logic, with $<$ being well-founded then $\mathcal{L}_{<}$ is **preferential**.*

However, this is not an equality. That is, there are cumulative / preferential inference relations that do not correspond to a model preference logic. However, if one takes the abstraction of model preference logic into account, i.e. if one uses preference models, one can obtain the following Kraus, Lehmann, and Magidor 1990; Brewka, Dix, and Konolige 1997.

Theorem 2.15 (Kraus, Lehmann, and Magidor 1990)**.** *A consequence relation is a preferential consequence relation iff it is defined by some preferential model.*

Hence, system **P** fully captures preference models. That is, any preference model does fulfil the properties of system **P** and that in general no additional property can be expressed by preference models, other than **P**. Moreover, since this does not hold for the specific class of preference models, namely model preference logic, it can be concluded that there are additional properties that can be expressed in general with model preference logic. Therfore, there are certain statements that are valid in model preference logic but not in preferential models Brewka, Dix, and Konolige 1997.

Examples of Preferential Inference

In order to hone intuition for how the preferential inference relation could be useful, some examples should be discussed. Firstly, the famous *Nixon-Triangle* shall be discussed. As presented in Strasser and Antonelli 2018 we have *Nixon* (N), *Quaker* Q, *Republican* (R) and *Pacifist* (P). Furthermore, we know that

- Nixon is a Republican;
- Nixon is a Quaker;
- Normally Quakers are Pacifist;
- Normally Republicans are not Pacifists.

This can be formalised, resulting in $\mathcal{T}$

$$\begin{aligned} &N \rightarrow R \\ &N \rightarrow Q \\ &Q \mid\!\sim P \\ &R \mid\!\sim \neg P \end{aligned}$$

Firstly, after applying reflexivity and right weakening

$$\frac{\mathcal{T} \vDash N \rightarrow R \qquad N \mid\!\sim N}{N \mid\!\sim R}$$

(analogue for Q) one obtains

$$\begin{aligned} &N \mid\!\sim R \\ &N \mid\!\sim Q \\ &Q \mid\!\sim P \\ &R \mid\!\sim \neg P \end{aligned}$$

From there it is impossible to derive $N \mid\!\sim \neg P$ or $N \mid\!\sim P$ from $\mathcal{T}$, given the system P. Here one observes, that $\mid\!\sim_{\mathbf{P}}$ uses sceptical reasoning. That is, rather than being able to derive all possible outcomes, namely P and $\neg P$, like a credulous reasoner would do, nothing can be derived. Thus, the information presented in $\mathcal{T}$ is not sufficient for the desired inference. However, as compared to monotonic logic no contradiction was derived, thus allowing inferences of other kinds.
Another example to discuss is related to *Tweety-Problem*. That is, given $\mathcal{K}$ containing

- Penguins are Birds;
- Penguins do not Fly;
- Normally Birds do Fly;

This can be formalised as follows (Note: the material implications are already transformed, as seen in the previous example).

$$\begin{aligned} &P \mid\!\sim B \\ &P \mid\!\sim \neg F \\ &B \mid\!\sim F \end{aligned}$$

Yet again, while classically one would infer a contradiction, it is not possible within the system **P** to infer $P \mathrel{|\!\sim} F$, thus no contradiction arises. However, it is still possible to obtain interesting inferences such as $P \wedge B \mathrel{|\!\sim} \neg F$ via

$$\frac{P \mathrel{|\!\sim} B \qquad P \mathrel{|\!\sim} \neg F}{P \wedge B \mathrel{|\!\sim} \neg F} \text{ (Cautious Monotonicity)}$$

Some Additional possible inferences are $P \vee B \mathrel{|\!\sim} F$, $B \mathrel{|\!\sim} \neg F$, $F \mathrel{|\!\sim} \neg P$ and $P \vee B \mathrel{|\!\sim} \neg P$.

3 Explanatory Non-Monotonic Reasoning

Explanatory Non-Monotonic Reasoning or

3.1 Default Logic

Default Logic is one of the more expressive and popular formalisms of non-monotonic reasoning. Hence, similar to circumscription many different variants of default logic have been introduced. However, if compared with the other two popular non-monotonic formalisms, circumscription and modal non-monotonic logic, default logic takes on a unique position. That is, instead of inducing logical formulas to obtain non-monotonicity, default logic relies on defining inference rules called *defaults*. These defaults can be understood as rules of thumb that enable the reasoner to jump to conclusions given incomplete information. Moreover, default logic even tolerates the existence of inconsistent information, if represented as default. Lastly, the reason why default logic belongs to the fixed-point logics and thus to the explanatory approach of non-monotonic reasoning, is that in its original formulation an operator was used, which uses these defaults to increase the set of derivations until no further expansion is possible. That is, until the set of facts serving as an input for the operator is also the output of that operator Bochman 2007; Reiter 1980; Antoniou and Wang 2007.

Before delving into the technical details of default logic, it is important to develop a strong intuition on how to read default rules. The following examples are based on the examples from Antoniou and Wang 2007. If one wants to formalise the following statement *An event (E) hosted by x takes place (T) with x unless x is not sick (S).* Formalised in classical logic this could look like

$$E(x) \wedge \neg S(x) \rightarrow T(x)$$

Unfortunately, when planning the event one can never be certain that x will not suddenly fall ill. Thus, $\neg S(x)$ will never hold and the event will never take place. Hence, a rule of thumb is required, which can be represented by the following inference rule.

$$\frac{E(x) : \neg S(x)}{T(x)}$$

which expresses that if there is an event hosted by x it is reasonable to assume that it will take place with x. Unless there is explicit information that x is sick, i.e. $\neg S(x)$. That is, given the absence of the information $S(x)$, $\neg S(x)$ can be assumed and one will be able to infer $T(x)$. As soon as there is information that x is sick the assumption of $\neg S(x)$ can no longer be made, thus blocking this inference rule. Moreover, one alternative formalisation of the same problem is the following

$$\frac{E(x) : T(x)}{T(x)}$$

and

$$S(x) \rightarrow \neg T(x)$$

Here the default rule states that given the event, it is simply assumed that the event takes place, thus it can be inferred that the event takes place. Furthermore, the statement in classical logic specifies a condition that blocks the default from assuming $T(x)$ if $S(x)$ is known. That is, exceptions are represented as classical rules. By representing the *exception* in such a manner it was possible to rewrite default rule into a normal default. This is, as will be discussed later, fairly important as defaults in normal form exhibit several desirable properties. Moreover, a default of this form can be used to represent *prototypical reasoning*, i.e. most instances of a concept have some property, which in this case means most events will take place. Antoniou and Wang 2007

Generally, speaking default logic is a powerful tool for formalising normality, exceptions or even no-risk reasoning, i.e. rejecting the highly probable conclusion, in favour of a save conclusion, and can therefore be used in various fields. For example,

- *no-risk reasoning in law*: The concept of *innocent until proven guilty* which if stated as *If x is accused, x is assumed to be innocent, unless evidence for the contrary is supplied* can be formalised as

  $$\frac{Accused(x) : Innocent(x)}{Innocent(x)}$$

  This is, especially important when considering, that even though it may be highly probable that x has committed a crime, one should reject an inference system that does not uphold fundamental rights.

- *prototypical reasoning in biology*: The statement *Normally birds fly* can be formalised as

$$\frac{Bird(x) : Fly(x)}{Fly(x)}$$

- *exceptions in biology*: Given the previous point one can model the exception *Penguins are Birds that do not fly* and as follows

$$Penguin(x) \rightarrow Bird(x) \land \neg Fly(x)$$

Syntax of Default Logic

Having established some kind of intuition on what a default rule is and how it can be understood. The next step is to properly define the syntax used above. First, a default is defined as follows.

Definition 3.1 (Antoniou and Wang 2007)**.** *A default δ has the form*

$$\frac{\varphi : \psi_1, \dots \psi_n}{\chi}$$

or in a compact notation

$$\varphi : \psi_1, \dots \psi_n / \chi$$

with $\varphi, \chi, \psi_1, \dots \psi_n$ being closed propositional formulas for $n > 0$. Moreover,

- *$pre(\sigma) = \varphi$ is called the prerequisite;*
- *$just(\sigma) = \{\psi_1, \dots \psi_n\}$ are called the justifications;*
- *$cons(\sigma) = \chi$ is called the consequent;*

Furthermore, a default theory can be defined as

Definition 3.2 (Antoniou and Wang 2007)**.** *A default theory $\mathcal{T}$ is a pair (W, D), with W being a set of predicate formulas representing the facts/axioms of $\mathcal{T}$ and with D being a set of defaults.*

One glaring inconsistency arises, if one tries to apply the definition of a default to the examples provided above. That is, the definition requires that any formula in the default has to be ground. Yet if one takes the rule

$$\frac{Bird(x) : Fly(x)}{Fly(x)}$$

for example one can easily observe that it is not grounded. However, if this "default" is not understood as a single default, but as a schemata defining a (possibly infinite) class of defaults, such a construction is still coherent with the provided definition. This notion can be formalised as follows.

Definition 3.3 (Antoniou and Wang 2007)**.** *A default schema is a default with $\varphi, \chi, \psi_1, \dots \psi_n$ being arbitrary propositional formulas. Moreover, the set of defaults generated by applying*

$$\frac{\varphi\sigma : \psi_1\sigma, \dots \psi_n\sigma}{\chi\sigma}$$

for all ground substitutions σ that assign values to all free variables occurring in the schema.

For example in the case of $W := \{Bird(Tweety), Bird(Polly)\}$ the schema

$$\frac{Bird(x) : Fly(x)}{Fly(x)}$$

actually represents a set of defaults containing

$$\frac{Bird(Tweety) : Fly(Tweety)}{Fly(Tweety)}$$

and

$$\frac{Bird(Polly) : Fly(Polly)}{Fly(Polly)}$$

Semantics of Default Logic

Properties of Default Logic

Even if default logic is already a fairly expressive system, i.e. it is more expressive than Circumscription and by proxy more expressive than the closed world assumption, it has certain drawbacks and idiosyncrasies that warren discussion.

Existence of Extensions If one reasons within the basic default theory, it is possible no extensions can be found. For example, if $W := \varnothing$ and $D := \{true : p/\neg p\}$, because the default invalidates its own application. This is, in general a value neutral property. On the one hand, if there are not enough meaningful rules to draw any meaningful conclusions, it one should not be required to do so. However, on the other hand, if one wants to build a system of inference that can tolerate deficient pieces of information, i.e. fault tolerant, this property is not beneficial.

Semi-Monotonicity This property is strongly correlated with the existence of an extension and can be defined as follows

Definition 3.4 (Antoniou and Wang 2007)**.** *Let $T = (W, D)$ and $T' = (W, D')$ be default theories such that $D \subseteq D'$. Then for every extension E of T there is an extension E' of T' such that $E \subseteq E'$.*

Which expresses that if the set of defaults increases then every extension obtained with the increased set of defaults will be a superset of the extensions obtained via the smaller set of defaults. Obviously, default logic does not satisfy this in general. Let $T = (\varnothing, \{true : p/p\})$ and $T' = (\varnothing, \{true : p/p, true : q/\neg q\})$ while T has the extension $\{p\}$, T' has no extensions.

Joint Consistency of Justifications In general, default logic does not require justifications to be consistent. That is, let $T = (\varnothing, \{true : p/q, true : \neg p/r\})$ then there is a single extension $\{q, r\}$. Hence, it is possible to reason within default logic even if during the process two contradicting justifications are assumed. However, in day to day reasoning one should be able to consider two contradicting assumptions. For example, one can assume that it may be raining tomorrow, as well as assume that it will be sunny tomorrow. Default logic approaches satisfying this property are among others

Cumulativity and Lemmas

Normal Default Logic

Preferential Default Logic

Other forms of Default Logic

Justified Default Logic

Constrained Default Logic

Cumulative Default Logic

3.2 Autoepistemic Logic

References

Antoniou, Grigoris and Kewen Wang (2007). “Default Logic”. In: *The Many Valued and Nonmonotonic Turn in Logic*. Ed. by Dov M. Gabbay and John Woods. Vol. 8. Handbook of the History of Logic. North-Holland, pp. 517–555. DOI: `https://doi.org/10.1016/S1874-5857(07)80011-2`. URL: `http://www.sciencedirect.com/science/article/pii/S1874585707800112`.

Bochman, Alexander (2005). *Explanatory nonmonotonic reasoning*. Vol. 4. World scientific.

– (2007). “Nonmonotonic Reasoning”. In: *The Many Valued and Nonmonotonic Turn in Logic*. Ed. by Dov M. Gabbay and John Woods. Vol. 8. Handbook of the History of Logic. North-Holland, pp. 557–632. DOI: `https://doi.org/10.1016/S1874-5857(07)80012-4`. URL: `http://www.sciencedirect.com/science/article/pii/S1874585707800124`.

Brewka, Gerhard, Jürgen Dix, and Kurt Konolige (1997). *Nonmonotonic reasoning: an overview*. Vol. 73. CSLI publications Stanford.

Evans, Jonathan St BT (2002). “Logic and human reasoning: An assessment of the deduction paradigm.” In: *Psychological bulletin* 128.6, p. 978.

Ginsberg, Matthew L and David E Smith (1987). “Reasoning about action I: A possible worlds approach”. In: *The frame problem in artificial intelligence*. Elsevier, pp. 233–258.

Koons, Robert (2017). “Defeasible Reasoning”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2017. Metaphysics Research Lab, Stanford University.

Kraus, Sarit, Daniel Lehmann, and Menachem Magidor (1990). “Nonmonotonic reasoning, preferential models and cumulative logics”. In: *Artificial intelligence* 44.1-2, pp. 167–207.

Lifschitz, Vladimir (1996). “Circumscription”. In:

– (2015). “The dramatic true story of the frame default”. In: *Journal of Philosophical Logic* 44.2, pp. 163–176.

McCarthy, John (1981). “Circumscription—a form of non-monotonic reasoning”. In: *Readings in Artificial Intelligence*. Elsevier, pp. 466–472.

McCarthy, John and Patrick J Hayes (1981). “Some philosophical problems from the standpoint of artificial intelligence”. In: *Readings in artificial intelligence*. Elsevier, pp. 431–450.

McDermott, Drew and Jon Doyle (1980). “Non-monotonic logic I”. In: *Artificial intelligence* 13.1-2, pp. 41–72.

Reiter, Raymond (1980). “A logic for default reasoning”. In: *Artificial intelligence* 13.1-2, pp. 81–132.

– (1981). “On closed world data bases”. In: *Readings in artificial intelligence*. Elsevier, pp. 119–140.

Shanahan, Murray (2016). “The Frame Problem”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2016. Metaphysics Research Lab, Stanford University.

Strasser, Christian and G. Aldo Antonelli (2018). “Non-monotonic Logic”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2018. Metaphysics Research Lab, Stanford University.