Integrating Reflective Equilibrium and Structured Argumentation: A Logical Approach to Norm Identification

Tianwen Xu[1], **Jieting Luo**[2]

[1]School of Philosophy, Nanjing University
[2]School of Philosophy, Zhejiang University
tianwen.xu@nju.edu.cn, luojieting@zju.edu.cn

Abstract

In open multi-agent systems without centralized authority, agents cannot rely on predefined norms and must instead learn them through decentralized norm identification. Existing data-driven and machine-learning methods are vulnerable to data quality, lack explainability, and often require labeled datasets, limiting their applicability in open environments. To address these issues, this paper proposes a logic-based approach to norm identification by integrating the method of reflective equilibrium and structured argumentation. Identifying norms is then conceived as constructing a consistent moral theory, and a moral theory is represented as an argumentation theory. This allows us to inductively construct a consistent moral theory from moral judgments reflecting prevalent morality in a community, and any potential conflict within the theory are resolvable by its priority structure comprised of specificity and firmness orderings. The theory can also develop arguments that justify an agent's decision by such theory. We prove that, whenever the firm part of the constructed theory applies, the agent can always derive a definite moral answer, fulfilling the dual demands of consistency and justification. Moreover, when presented with a set of *prima facie* norms as inputs, our method ensures the path-independent identification of any firm core that exists.

Introduction

Norms are one of the important elements that agents need to consider for decision-making apart from their belief, desire and intention. Research on norm-based reasoning investigates how to resolve the conflicts between goal achievement and norm compliance (Broersen et al. 2001; Knobbout and Dastani 2012; Criado et al. 2014), or how to design appropriate sanction-based norms in order to enforce desired system behavior (Dastani, van der Torre, and Yorke-Smith 2017; Knobbout, Dastani, and Meyer 2016; Dell'Anna, Dastani, and Dalpiaz 2020; Akinkunmi and Babalola 2020). All this research assumes that agents are aware of norms so that agents can decide whether to comply with the norms or not. Yet, in open multi-agent systems without a central authority, this awareness cannot be pre-programmed; it must be learned. This brings us to the fundamental area of *norm identification*. Here, scholars investigate how agents can discern the norms already prevalent in a society through their own experiences and observations of others, thereby making the initial assumption a reality through decentralized learning (Cranefield and Dhiman 2021).

Currently, norm-identification techniques rely on big data and machine learning. For example, Awad et al. (2020) extract ethical principles from dilemma vignettes using inductive logic programming (ILP) , and Cranefield and Dhiman (2021) identify norm candidates from a normative language using Markov Chain Monte Carlo (MCMC) search, and Woodgate, Marshall, and Ajmeri (2025) create ethical norm-learning agents that operationalise maximin in their reinforcement learning processes by balancing societal well-being with individual goals. However, these data-driven methods are vulnerable to the input data quality, in that data has to be labeled consistently and the right data properties need to be described in a machine-processable way to obtain an accurate training of machines (Tolmeijer et al. 2020). Data-driven methods are also impossible to defend agents' norm-based decisions, since there is no argument behind the identified norms.

Alternatively, norms can be identified through logical reasoning. Typically for moral norms that are not written in text but are enforced by personal beliefs and public opinion, agents receive moral judgments through social practice and transfer them to moral norms for constructing their *coherent* moral theory. Some attempts have been done by Johnston and Governatori (2003) on theorizing a legal database using induction of defeasible logic, but the superiority relation over a set of rules is given instead of being derived empirically.

In this paper, we see norm identification as a problem of theory construction using the method of *reflective equilibrium*. Inspired by structured argumentation, a moral theory is represented by an argumentation theory, to capture the defeasibility of normative reasoning and the priority structure therein. The argumentation theory is inductively constructed to include moral judgments (as *prima facie* norms) from the environment and to generate the priority structure among them by means of specificity ordering and firmness ordering. The result is a set of identified norms which are as consistent as possible and able to justify the agent's decisions through derived arguments. We prove that, as far as the firm part of the constructed moral theory can apply, the agent can

always derive a definite moral answer. This fulfills the task of norm identification, in particular the expectation of consistency (as much as possible) and capability of moral defense. Besides, given a set of moral norms that are uncontroversially followed and practiced by agents within the community, the agent can always path-independently identify them through our method of construction. Our method does not rely on consistent data and can provide justification for agents' norm-based decisions. Moreover, compared with existing learning-based approaches that allow agents to identify norms during training before execution, our logic-based approach allows agents to identify norms directly during execution, making it more possible to deploy autonomous agents in open and unknown environments.

Seeing Norm Identification as Reflective Equilibrium

In a broad sense, norms of a society provide the rules of encounter under which that society's members interact. They can be legal norms that are promulgated and enforced by authorities, or social norms that are shared and considered acceptable by majority of a society, or moral norms that are inherently used to judge actions, behaviors and states of affairs. Different from legal norms that are fixed in laws, codices, regulations, orders etc., the violation of which leads to punishment enforced by public authority, social norms and moral norms are not written but are enforced by personal beliefs and public opinion. Among these, social norms are upheld by convention and the pressure of public opinion, while moral norms operate as deeper standards used to judge actions, behaviors, and states of affairs.

The norms that we consider in this paper are moral norms that have been accepted and used for judgment by a community, thus it makes sense to identify them from what agents in the community approve or disapprove, namely their moral judgments. The moral judgments might be expressed directly in words, or indirectly by sanctioning signals such as emotions and behavior. In this paper, We treat judgments as *prima facie* norms that serve as inputs to our logical system and do not care about the way through which agents receive moral judgments from other agents.

Now the question arises as to what method an agent can use to identify moral norms from the received moral judgments. In general the method should serve as a device for theorizing the moral system that underlies the moral judgments. To speak more specifically, it should be able to generalize the moral norms that govern the moral judgments, and to resolve the inconsistencies between moral judgments and moral norms that the agent believes govern the judgments. The inconsistencies may arise because norms might have exceptions but they haven't been identified by the agent (queue jumping can be allowed for emergency), or the agent might interpret other agents' moral judgments in a wrong way, or stereotype the whole community with local attitudes of a minority, and etc. For considerations as such, in this work we propose that the agent with the task of norm identification can use the method of reflective equilibrium, proposed by John Rawls (Rawls 1951, 1974, 1999) for explicating moral judgments and by Ronald Dworkin for interpreting social practices (Dworkin 1986), to identify moral norms from moral judgments and to resolve the inconsistencies therein.

Reflective equilibrium is a method that we use to construct a consistent moral theory. It works back and forth among moral judgments and general moral principles that are supposed to justify the judgments, revising any of these elements wherever necessary in order to achieve an equilibrium in which all judgments are justified by principles. A *moral judgment* is some particular belief or feeling that we might have as to whether a certain action is right or wrong in some situation. For example, if we find out that a person has been murdered, we usually think that this is terrible. Given this moral judgment, we then try to generalize and come up with a general *moral principle* that explains this judgment. In this particular case, we can have "we should never kill any people" as a principle for explaining why we feel bad when knowing a person gets murdered. As we practice in the society, we come up with many such moral principles. According to the method of reflective equilibrium, we should check whether these principles are compatible with each other and consistent with the other judgments we have. If they come into conflict then we can either revise the principle in order to keep our moral judgments, or accept a principle and revise our judgments instead.

For illustration, let us continue to discuss the above example. Suppose that we encounter another case where a person killed another person for self-defense. Most people would probably consider murder for self-defense as morally right. So we now have another moral judgment. But if we apply the moral principle that we should never kill any people to this case, inconsistency between the moral judgment and the moral principle arises. To resolve it, one option is to modify the original principle as, for example, we should not kill any people unless we do it for self defense. We can also keep the original principle and see self defense as a wrong act. This illustrates that, to understand morality is to theorize our intuitive judgments into a consistent moral theory, as consistent as it can be. This process of reflective equilibrium can go on and on, in order to achieve an equilibrium in which all judgments are justified by principles, as depicted by Fig.1.

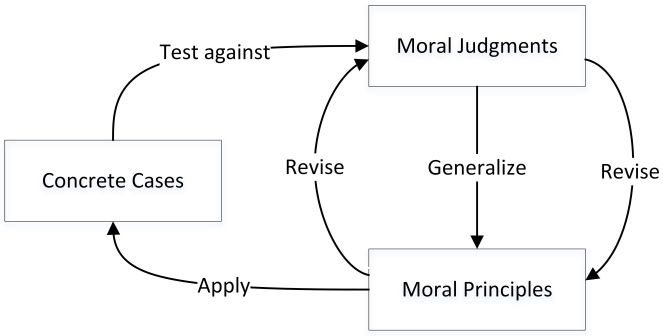

Figure 1: The method of Reflective Equilibrium.

Norm identification closely parallels aforementioned ethi-

cal deliberation. Both examine moral norms that emerge organically rather than being decreed by an authority. Their shared objectives are to clarify what these norms are, understand the mechanisms of compliance, and establish a sound justification for decisions based on them. Seeing moral principles as abstract moral norms (more general than moral judgments), an agent can use reflective equilibrium for identifying norms. It can start from having moral judgments. Different from the process we introduced above, the moral judgments do not come from the agent itself, but are collected from environments, e.g., received from other agents in a community. The moral judgments can be communicated to the agent in words, or through sanctioning signals such as emotions and behavior. Here we assume that the agent already has data-driven mechanisms for language processing and emotion recognition and is able to infer moral judgments from them. It then looks for moral norms that justify the judgments. Whenever inconsistencies are encountered, the agent can either revise its beliefs of moral judgments, or modify its beliefs of moral norms. This is done incrementally, until finally reach an equilibrium between the judgment set and the norm set. At this final stage, the moral norms are identified because they can explain and justify the judgments of approval or disapproval for certain actions.

Modeling Moral Theory and Reasoning through Structured Argumentation

We formalize a moral theory and reasoning with the theory by means of *structured argumentation*, in particular the renown *ASPIC*$^+$ framework (Modgil and Prakken 2013, 2014). In a nutshell, *ASPIC*$^+$ is a non-monotonic reasoning formalism comprised of customizable representational language, monotonic/non-monotonic inference rules and knowledge base. From these components it accounts for the formation of argumentative structures within and between arguments that represent (non-)monotonic proof sequences. The result of (non-monotonic) reasoning is then determined by argument acceptability semantics founded by Dung (1995). In what follows, we will define a simplified *ASPIC*$^+$ formalism, which differs from the original *ASPIC*$^+$ but maintains its core idea. We start with our model of moral theory as a special type of argumentation theory: *contextual argumentation theory*.

Definition 1 (Contextual argumentation theory)**.** *A contextual argumentation theory CAT is a tuple* $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C})$*, where:*

- $\mathcal{L}$ *is a logical language closed under negation and has expressivity rich enough to express deontic formulas in the form of* $O\varphi, P\varphi$*, etc.*
- $\mathcal{R}^s$ *is the set of strict inference rules of the form* $\varphi_1, \ldots, \varphi_n \rightarrow \psi$ *reflecting the inferences of a monotonic proof system* $\mathcal{S}_\mathcal{L}$ *over* $\mathcal{L}$*, such that:*

$$\varphi_1, \ldots \varphi_n \rightarrow \psi \in \mathcal{R}^s \text{ iff } \{\varphi_1, \ldots \varphi_n\} \vdash_{\mathcal{S}_\mathcal{L}} \psi.$$

  with $\mathcal{S}_\mathcal{L}$ *satisfying:*
  - *Propositional reasoning is conducted by the classical Propositional Calculus.*
  - *Contains the classical definitional scheme between deontic operators.*
- $\mathcal{R}^d$ *is a finite set of defeasible inference rules of the form* $\varphi_1, \ldots, \varphi_n \Rightarrow \odot\psi$*, where* $\odot$ *is a general reference of deontic operators expressed in* $\mathcal{L}$ *(e.g.,* $O\varphi$*,* $P\varphi$*,* $F\varphi$*).*
- $\mathcal{C} = \{c \mid c \in 2^{\mathcal{L}^-} \text{ and } c \nvdash_{\mathcal{S}_\mathcal{L}} \bot\}$ *is a collection of contexts, where* $\mathcal{L}^-$ *is the fragment of* $\mathcal{L}$ *free of deontic formulas.*

All defeasible rules are representation of norms of the form: $\varphi_1, \varphi_2, \cdots \Rightarrow \odot\psi$.

Defined as such, a *CAT* intends to reflect some key features of a moral theory.

1. The background reasoning over facts and deontic notions. Rules of such reasoning as included in $\mathcal{R}^s$.
2. Norms that comprised of condition to apply and normative consequence, by which the reasoning is mostly defeasible. $\mathcal{R}^d$ represents moral norms with this structure and defeasible character. Elements in $\mathcal{R}^d$ can be considered as norms with different degree of generality, from the most concrete moral judgments to the more abstract moral principles.
3. Contexts that have moral relevance. Intuitively speaking, a moral theory always have a set of contexts in mind that it wants to give moral answers. For instance, a theory of AI ethics will take into account various scenarios where AI system interact with human being, such as data collection, recommendation, user profiling. In each scenario (or in our term, a 'context'), the theory must answer how an AI system should behave: when a user is visiting a website, the cookies should only be obtained by the website subject to user's consent; when a user expresses an intention to suicide to a language model, the model must not encourage the user to actually conduct suicide. The set $\mathcal{C}$ collects these contexts. To construct a moral theory then requires taking into account relevant contexts as many as possible. If readers are familiar with *ASPIC*$^+$, then $\mathcal{C}$ can be considered as a set of knowledge bases $\mathcal{K}$ that only contains axiomatic premises (as elements of a context are factual description of the context, therefore are not supposed to be challenged).

Next we define arguments based on a *CAT*. These arguments signify different types of reasoning under a moral theory. For simplification, we require that each argument contains *at most* one defeasible rule. This suffices for expressing the common type of moral reasoning that derives each moral argument from a single norm and consider moral disagreement as attack relation between moral arguments. As is the case in original *ASPIC*$^+$, for an argument we use `Prem` to return its premises, `Conc` its conclusions, `Sub` its sub-arguments, `DefRule` the defeasible rules it used and `TopRule` the inference rule it used in its last argument step .

Definition 2 (Arguments)**.** *Let CAT* $= (\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C})$*. Let* $c \in \mathcal{C}$ *be a context. An argument* A *defined in CAT with respect to c is:*

- φ *if* $\varphi \in c$ *with:* $\texttt{Prem}(A) = \{\varphi\}$*,* $\texttt{Conc}(A) = \varphi$*,* $\texttt{Sub}(A) = \{\varphi\}$*,* $\texttt{DefRule}(A) = \emptyset$*,* $\texttt{TopRule}(A) =$ *undefined;*

- $A_1, \ldots, A_n \rightarrow \psi$ *if* $A_1, \ldots, A_n$ *are arguments such that there exists a strict rule* $\texttt{Conc}(A_1), \ldots, \texttt{Conc}(A_n) \rightarrow \psi$ *in* $\mathcal{R}^s$*, and:*
  $\texttt{Prem}(A) = \texttt{Prem}(A_1) \cup \ldots \cup \texttt{Prem}(A_n)$,
  $\texttt{Conc}(A) = \psi$,
  $\texttt{Sub}(A) = \texttt{Sub}(A_1) \cup \ldots \cup \texttt{Sub}(A_n) \cup \{A\}$,
  $\texttt{DefRule}(A) = \texttt{DefRule}(A_1) \cup \ldots \cup \texttt{DefRule}(A_n)$,
  $\texttt{TopRule}(A) = \texttt{Conc}(A_1), \ldots, \texttt{Conc}(A_n) \rightarrow \psi$.
- $A_1, \ldots A_n \Rightarrow \psi$ *if* $A_1, \ldots A_n \Rightarrow \psi$ *is an argument such that there exists a defeasible rule* $\texttt{Conc}(A_1), \ldots \texttt{Conc}(A_n) \Rightarrow \psi \in \mathcal{R}^d$*, and:*
  $\texttt{Prem}(A) = \texttt{Prem}(A_1) \cup \ldots \cup \texttt{Prem}(A_n)$,
  $\texttt{Conc}(A) = \psi$,
  $\texttt{Sub}(A) = \texttt{Sub}(A_1) \cup \ldots \cup \texttt{Sub}(A_n) \cup \{A\}$,
  $\texttt{DefRule}(A) = \{\texttt{Conc}(A_1), \ldots, \texttt{Conc}(A_n) \Rightarrow \psi\}$ *(a singleton as each argument contains at most one defeasible inference rule.)*
  $\texttt{TopRule}(A) = \texttt{Conc}(A_1), \ldots, \texttt{Conc}(A_n) \Rightarrow \psi$.

We use $\mathcal{A}^c$ *to denote the smallest set of finite arguments* [1] *generated by CAT with respect to a context* c*.*

Given that $\texttt{DefRule}(A)$ is always a singleton in our formalization. For the rest of the paper, instead of writing $\texttt{DefRule}(A) = \{r\}$, we will slightly abuse the notation and directly write $\texttt{DefRule}(A) = r$.

In our simplified *ASPIC*$^+$, since we do not specify a naming function and the premises of arguments are facts of a contexts that cannot be attacked, an argument can only attack another argument through rebutting.

Definition 3 (Attacks)**.** *Argument* A *attacks argument* B *if and only if* $\texttt{Conc}(A) = -\psi$ *for some* $B' \in \texttt{Sub}(B)$ *of the form* $B'_1, \ldots, B'_n \Rightarrow \psi$*.*

We use '$-$' for the case where $\texttt{Conc}(A)$ is contradictory to $\texttt{Conc}(B)$.

Definition 4 (Contextual normative conflict)**.** *Let CAT* $= (\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C})$*. Let* $\mathcal{A}^c$ *be the smallest set of finite arguments generated by CAT based on a context* $c \in \mathcal{C}$*. For any* $x, y \in \mathcal{R}^d$*,* x, y *are in contextual conflict with respect to* c*, denoted as*

$$x \leftrightsquigarrow_c y$$

if there are $A, A' \in \mathcal{A}^c$*, such that* $\texttt{Conc}(A) = -\texttt{Conc}(A')$*, and* $\texttt{DefRule}(A) = x, \texttt{DefRule}(A') = y$*.*

Note that by this definition, the argument A, A' attacks each other. This match the underlying intuition that when we say two norms are in conflict, we are referring to a case where the two norms can be used to derive incompatible normative decisions, i.e., arguments that rebut each other's conclusion. Note also that the relation $\leftrightsquigarrow_c$ is symmetric.

Besides reasoning mechanism $(\mathcal{R}^s, \mathcal{R}^d)$ and contexts $(\mathcal{C})$, a moral theory typically contains priority structure of the norms. This structure is crucial to determine which norm prevails over another when they are in conflict. This is captured by the following definitions of *prioritized contextual argumentation theory* and attack/defeat relation between moral arguments.

Definition 5 (Prioritized contextual argumentation theory)**.** *Let* $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C})$ *be a contexual argumentation theory.* $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C}, \trianglelefteq)$ *is a prioritized contextual argumentation theory, notation PCAT, if* $\trianglelefteq \subseteq \mathcal{R}^d \times \mathcal{R}^d$*.*

Let $x, y \in \mathcal{R}^d$. $x \trianglelefteq y$ is read as y *is at least as prioritized as* x. If $x \trianglelefteq y$ and $y \not\trianglelefteq x$, then we write $x \triangleleft y$, meaning y *is more prioritized than* x. If $x \trianglelefteq y$ and $y \trianglelefteq x$, then we write $x \sim y$, meaning x *is as prioritized as* y*.*

Again, as every argument in our formalization contains at most one defeasible rule, instead of defining another relation between singletons $\texttt{DefRule}(A)$ and $\texttt{DefRule}(A')$, we will slightly abuse the notation and just write $\texttt{DefRule}(A) \trianglelefteq \texttt{DefRule}(A')$.

Definition 6 (Structured contextual argumentation framework)**.** *Let PCAT* $= (\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C}, \trianglelefteq)$*. A structured contextual argumentation framework with respect to PCAT and a context* $c \in \mathcal{C}$*, notation SCAF*^{c}*, is a tuple* $(\mathcal{A}^c, att, \preceq)$*, where:*

- $\mathcal{A}^c$ *is the smallest set of finite arguments generated by PCAT under* c*.*
- *For any* $A, A' \in \mathcal{A}^c$*,* $(A, A') \in att$ *if and only if* A *attacks* A'*.*
- $\preceq \subseteq \mathcal{A}^c \times \mathcal{A}^c$ *is an ordering over arguments.*

The strict version $\prec$ and equal version $\simeq$ of $\preceq$ is defined as usual. Since our formalism only has axiomatic premises and each argument contains at most one defeasible rule, $\preceq$ is determined rather straightforwardly by the ordering $\trianglelefteq$ over $\mathcal{R}^d$.

Definition 7 (Determination of $\preceq$ by $\trianglelefteq$)**.** *Let PCAT* $= (\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C}, \trianglelefteq)$*. Let SCAF*^{c} $= (\mathcal{A}^c, att, \preceq)$ *be defined by PCAT with respect to a context* $c \in \mathcal{C}$*. For any* $A, A' \in \mathcal{A}^c$*:*

- *If* $\texttt{DefRule}(A) = \emptyset$ *then* $A \not\preceq A'$*;*
- *If* $\texttt{DefRule}(A) \neq \emptyset$ *and* $\texttt{DefRule}(A') = \emptyset$*, then* $A \preceq A'$*; else*
- $A \preceq A'$ *if and only if* $\texttt{DefRule}(A) \trianglelefteq \texttt{DefRule}(A')$*.*

Proposition 1. *Let PCAT* $= (\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C}, \trianglelefteq)$*. Let SCAF*^{c} $= (\mathcal{A}^c, att, \preceq)$ *where* $c \in \mathcal{C}$ *and* att *is determined by* $\trianglelefteq$ *according to Definition 7. This produce the same argument ordering as those defined by weakest link or last link principle, using either elitist or democratic preference lifting in Modgil and Prakken (2014).*

Definition 8 (Defeat)**.** *Argument* A *defeats argument* B *if and only if* A *attacks* B *and* $A \not\prec B$*.*

Proposition 2. *Argument* A *defeats argument* B *if and only if* A *attacks* B *and:*

- $\texttt{DefRule}(A) = \emptyset$*, or*
- $\texttt{DefRule}(A) \not\trianglelefteq \texttt{DefRule}(B)$*.*

Definition 9 (Contextual argumentation frameworks)**.** *Let PCAT* $= (\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C}, \trianglelefteq)$*. Let SCAF*^{c} $= (\mathcal{A}^c, att, \preceq)$ *be defined by PCAT with respect to a context* $c \in C$*. A contextual argumentation framework with respect to the context* c*, notation CAF*^{c}*, is a pair* $(\mathcal{A}^c, \mathcal{D}^c)$*, where* $(X, Y) \in \mathcal{D}^c$ *iff* X *defeats* Y*.*

[1] An argument is finite if the set of strict and defeasible rules that used by the argument is finite. This is a default setting of *ASPIC*$^+$ (Modgil and Prakken 2014).

This argumentation framework reflects the moral arguments and the argumentative structure in-between when we are using a moral theory to derive moral answers, faced with specific contexts,. As mentioned earlier, the answer depends on which arguments we accept according to the semantics in Dung (1995), or using Dung's terminology, whether an argument is in a particular 'extension' defined based on an acceptability notion. The next theorem shows that the way we define *CAF*^{c} framework meets the rational postulates to avoid bizarre behaviors in non-monotonic reasoning.

Theorem 1 (Satisfaction of rationality postulates)**.** *Let CAF*$^c = (\mathcal{A}^c, \mathcal{D}^c)$ *be a contextual argumentation framework based on Definition 9. For any complete extension* E *of CAF*^{c} *under the Dung's semantics,* E *satisfies the following rational postulates proposed by Caminada and Amgoud (2007).*

- *Subargument closure;*
- *Closure under strict rules;*
- *Direct consistency;*
- *Indirect consistency.*

Now we can formalize how a moral theory may give definite moral answers. Intuitively, a definite answers is derived from arguments that are either uncontroversial or can be defended by uncontroversial arguments. In terms of Dung's semantics, this means that the answer should be derived from arguments in *grounded extension*. We prove that if a moral theory is *well-structured*, then all of its extensions are equivalent to grounded extension. This means that the theory is 'clear enough' such that its answer is always acceptable with respect to any acceptability standard. Based on Theorem 30 in (Dung 1995), it amounts to show that the contextual argumentation framework defined by a well-structured theory is *well-founded*, and that there is no infinite sequence $A_1, \ldots, A_n, \ldots$ such that A_{i+1} defeats A_i for each i.

Definition 10 (Well-structured argumentation theory)**.** *Let PCAT* $= (\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C}, \trianglelefteq)$*. PCAT is well-structured, if*

- $\trianglelefteq$ *is a preoder, and*
- *for any* $x, y \in \mathcal{R}^d$*, if* $x \leftrightsquigarrow_c y$ *for some* $c \in \mathcal{C}$ *then either* $x \triangleleft y$ *or* $y \triangleleft x$*.*

Proposition 3. *Let PCAT* $= (\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C}, \trianglelefteq)$ *be well-structured. Let* $c \in C$ *be a context and CAF*$^c = (\mathcal{A}^c, \mathcal{D}^c)$*. For any sequence of defeasible arguments* $A_1, \ldots, A_k, \ldots$ *from* $\mathcal{A}^c$*, if* A_{i+1} *defeats* A_i *for each* A_i *in the sequence* ($i \leq n-1$ *when the sequence only contains finite* n *arguments), then* $\texttt{DefRule}(A_i) \triangleleft \texttt{DefRule}(A_{i+1})$*.*

Theorem 2. *Let PCAT* $= (\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C}, \trianglelefteq)$*. If PCAT is well-structured, then CAF*^{c} *is well-founded for any* $c \in \mathcal{C}$*.*

Theory Construction by Reflective Equilibrium

Now question arises as to how such well-structured argumentation theory (conceptually a moral theory) can be actually constructed, so as to fullfil the task of norm identification through theory construction. Inspired by reflective equilibrium, we first introduce *specificity ordering* and *firmness ordering* on which $\trianglelefteq$ is based, and then we design a theory construction procedure to obtain a well-structured *PCAT* in finite steps of construction.

Specificity ordering. Specificity ordering reflects a common principle of ranking moral norms: more specific norm takes priority. Given a defeasible rule r, let $\texttt{body}(r)$ denote the conjunction of its antecedent, and $\texttt{head}(r)$ its consequent. For example, if r is $\varphi_1, \varphi_2, \varphi_n \Rightarrow O\psi$, then $\texttt{body}(r) = \varphi_1 \wedge \varphi_2 \wedge \ldots \wedge \varphi_n$, and $\texttt{head}(r) = O\psi$.

Definition 11 (Specificity ordering)**.** *Let CAT* $= (\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C})$*. A specificity ordering* $\trianglelefteq^{\texttt{s}}$ *over* $\mathcal{R}^d$ *is defined as: For any* $x, y \in \mathcal{R}^d$*,*

$$x \trianglelefteq^{\texttt{s}} y \text{ iff } \texttt{body}(y) \rightarrow \texttt{body}(x) \in \mathcal{R}^s$$

Further, $x \trianglelefteq^{\texttt{s}} y$ *if* $x \trianglelefteq^{\texttt{s}} y$ *but not* $y \trianglelefteq^{\texttt{s}} x$*,* $x \sim^{\texttt{s}} y$ *if* $x \trianglelefteq^{\texttt{s}} y$ *and* $y \trianglelefteq^{\texttt{s}} x$*.*

It is easy to see that $\trianglelefteq^{\texttt{s}}$ is reflexive, transitive and antisymmetric, i.e., a partial order.

Definition 12 (Structural coherence)**.** *Let CAT* $= (\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C})$*.* $\mathcal{R}^d$ *is structurally coherent under* $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C})$*, if for any* $x, y \in \mathcal{R}^d$*, if* $x \leftrightsquigarrow_c y$ *for some* $c \in \mathcal{C}$ *then either* $\texttt{body}(x) \rightarrow \texttt{body}(y) \in \mathcal{R}^s$ *or* $\texttt{body}(y) \rightarrow \texttt{body}(x) \in \mathcal{R}^s$*.*

Proposition 4. *Let CAT* $= (\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C})$*. If* $\mathcal{R}^d$*is structurally coherent, then any* $\Delta \subseteq \mathcal{R}^d$ *is also structurally coherent.*

Proposition 5. *Let PCAT* $= (\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C}, \trianglelefteq)$*. If* $\mathcal{R}^d$ *is structurally coherent and* $\trianglelefteq$ *is the specificity ordering on* $\mathcal{R}^d$ *according to Definition 11, then PCAT is well-structured.*

Together the two propositions reflect the intuition that if any contextual conflict within a moral theory is resolvable by the specificity ordering over norms, then the moral theory always gives definite answers.

Firmness ordering. The lesson from reflective equilibrium is that we can conceive moral theory construction as searching and expanding the reliable part of our moral sense. By reliable we mean that one would be more confident to hold a moral judgment if it is supported by a principle, and moral judgments with their supporting principles together form a *firm* part of our moral sense. This lead us to divide $\mathcal{R}^d$ into $\mathcal{R}^{d+}$ and $\mathcal{R}^{d-}$ signifying respectively the firm and uncertain part of our moral sense, and specify that rules in the former take priority over those in the latter.

Definition 13 (Firmness ordering)**.** *Let* $\mathcal{R}^{d+}, \mathcal{R}^{d-}$ *be two set of defeasible rules. Let CAT* $= (\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C})$*, where* $\mathcal{R}^d = \mathcal{R}^{d+} \cup \mathcal{R}^{d-}$*. A firmness ordering* $\trianglelefteq^{\texttt{f}}$ *over* $\mathcal{R}^d$ *is defined as: For any* $x, y \in \mathcal{R}^d$*,*

$$x \trianglelefteq^{\texttt{f}} y \text{ iff } y \in \mathcal{R}^{d+} \text{ and } x \in \mathcal{R}^{d-}.$$

Further, $x \trianglelefteq^{\texttt{f}} y$ *if* $x \trianglelefteq^{\texttt{f}} y$ *but not* $y \trianglelefteq^{\texttt{f}} x$*,* $x \sim^{\texttt{f}} y$ *if* $x \trianglelefteq^{\texttt{f}} y$ *and* $y \trianglelefteq^{\texttt{f}} x$*.*

On condition that $\mathcal{R}^{d+} \cap \mathcal{R}^{d-} = \emptyset$, it is easy to verify that $\trianglelefteq^{\texttt{f}}$ is asymmetric. In this case, $x \trianglelefteq^{\texttt{f}} y$ implies $y \trianglelefteq^{\texttt{f}} x$.

Theory construction process. With reflective equilibrium in mind, in our work a prioritized contextual argumentation theory $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C}, \trianglelefteq)$ is inductively constructed through inputs of moral judgments and operations on defeasible rule set according to the supportive relation between norms (judgments and principles are both norms but differs in degree of generality). We first set out some auxiliary notations.

- The construction process is represented as a sequence of stages $Q = (0, \ldots, n)$. The number 0 denotes the initial stage where no construction is made.
- For each $i \in Q$, a moral judgment is input in the form of $c_i \Rightarrow \odot\varphi$, where c_i is the context of the judgment. This input rule is denoted as r_i. In this way, each judgment can be conceived as conveying *prima facie* norms. We further require that there are no $r_i, r_j (i, j \in Q)$ in the form $c_i \Rightarrow \odot\varphi, c_j \Rightarrow \odot\psi$ such that $\vdash_{\mathcal{SL}} c_i \leftrightarrow c_j$ and $\vdash_{\mathcal{SL}} \odot\varphi \leftrightarrow \odot\psi$.This means that input rules are supposed to be distinct moral judgments. [2]
- The theory $PCAT_i$ to be constructed at stage i is denoted as $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d_i, \mathcal{C}_i, \trianglelefteq_i)$.

$PCAT_i$ in each $i \in Q$ is a result of inductive construction. For each $i \in Q$, the construction intuitively goes as follows:

1. When a moral judgment r_i is input at i, we need to verify two things: (1) whether the judgment, as a defeasible rule, has any contextual conflict with the firm set $\mathcal{R}^{d+}_{i-1}$ at preceding stage, and whether the conflict can be resolved by specificity ordering; (2) whether the context c_i of r_i will let us discover new unresolvable (by specificity) conflicts between norms already in $\mathcal{R}^{d+}_{i-1}$. For both cases, if we find contextual conflicts within $\{r_i\} \cup \mathcal{R}^{d+}_{i-1}$ and the conflicts are not resolvable by specificity ordering, then we discover vagueness in our moral sense, and the conflicting norms should tentatively be considered unreliable, and move to the uncertain set $\mathcal{R}^{d-}_i$.
2. If r_i survives the aforementioned verification, then it can be added to $\mathcal{R}^{d+}_{i-1}$, meanwhile retrieves those norms it 'support' from the $\mathcal{R}^{d-}_{i-1}$ to $\mathcal{R}^{d+}_{i-1}$, so long as they do not have unresolvable conflicts with $\mathcal{R}^{d+}_{i-1}$ after its vague norms are removed. The reason for doing so is that the retrieved norms now find a support from a norm in the firm part of our moral sense, just like a moral principle justifying a moral judgment in reflective equilibrium.
3. For any norm r' in $\mathcal{R}^{d+}_{i-1}$, if all the norms that serve as the reason for it to be considered reliable are removed due to vagueness, then r' should be moved to $\mathcal{R}^{d-}_{i-1}$ due to lack of 'ground'. In our work, a norm losing ground means it was previously retrieved to the firm set and any norm that supports it is already removed.
4. The new firm set $\mathcal{R}^{d+}_i$ and uncertain set $\mathcal{R}^{d-}_i$ is formed according to the operations describe earlier.

Now we give formal definition of the operations mentioned above: *supportive retrieval*, *vagueness removal* and *ground removal*. After that we will define theory construction based on these notions. And we verify some of their important properties which are crucial to a rational theory construction.

Definition 14 (Support between norms)**.** *Let CAT* $= (\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C})$. *For any* $x, y \in \mathcal{R}^d$, x *supports* y *(notation* $x \gg y$*), if* $\texttt{body}(y) \rightarrow \texttt{body}(x) \in \mathcal{R}^s$ *and* $\texttt{head}(x) \rightarrow \texttt{head}(y) \in \mathcal{R}^s$.

Proposition 6. *Let CAT* $= (\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C})$. *The support relation* $\gg$ *on* $\mathcal{R}^d$ *is reflexive and transitive.*

Proposition 7. *Let* $\mathcal{L}$ *be a logical language,* $\mathcal{R}^s$ *the set of strict rules reflecting monotonic inferences over* $\mathcal{L}$*. Let* $\mathcal{R}^d$ *be a set of defeasible rules, and* $\gg$ *the support relation on* $\mathcal{R}^d$ *based on* $\mathcal{R}^s$*. For any* $r \in \mathcal{R}^d$*, the set* $\{x \in \mathcal{R}^d \mid r \gg x\}$ *contains no contextual conflict under* $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C})$ *for any set of contexts* $\mathcal{C}$*.*

Definition 15 (Supportive retrieval)**.** *Let* $\mathcal{L}$ *be a logical language,*$\mathcal{R}^s$ *the set of strict rules reflecting the monotonic inferences over* $\mathcal{L}$*,* $\mathcal{C}$ *a set of contexts, and* Δ, Λ *be two defeasible rule sets. Let* r *be a defeasible rule. The supportive retrieval by* r *from* Δ *to* Λ *with respect to* $\mathcal{C}$ *is defined as follows:*

$$\begin{aligned}\texttt{ret}(r, \Delta, \Lambda, \mathcal{C}) = \{x \in \Delta \mid\ & r \gg x, \textit{and for any } y \in \Lambda,\\ & \textit{if } x \leftrightsquigarrow_c y \textit{ for some } c \in \mathcal{C}\\ & \textit{under } (\mathcal{L}, \mathcal{R}^s, \Lambda \cup \{x\}, \mathcal{C})\\ & \textit{then either } \texttt{body}(x) \rightarrow \texttt{body}(y) \in \mathcal{R}^s\\ & \textit{or } \texttt{body}(y) \rightarrow \texttt{body}(x) \in \mathcal{R}^s\}\end{aligned}$$

Proposition 8. *Let* $\mathcal{L}$ *be a logical language,* $\mathcal{R}^s$ *the set of strict rules reflecting the monotonic inferences over* $\mathcal{L}$*. Let* Δ, Λ *be two defeasible rule sets, and* r *a defeasible rule, then:*

1. *For any set of contexts* $\mathcal{C}$, $\texttt{ret}(r, \Delta, \Lambda, \mathcal{C})$ *contains no contextual conflict under* $(\mathcal{L}, \mathcal{R}^s, \texttt{ret}(r, \Delta, \Lambda, \mathcal{C}), \mathcal{C})$.
2. *If* Λ *is structural coherent under* $(\mathcal{L}, \mathcal{R}^s, \Lambda, \mathcal{C})$*, then* $\Lambda \cup \texttt{ret}(r, \Delta, \Lambda, \mathcal{C})$ *is structural coherent under* $(\mathcal{L}, \mathcal{R}^s, \Lambda \cup \texttt{ret}(r, \Delta, \Lambda, \mathcal{C}), \mathcal{C})$.
3. *Let* $\Gamma_0 = \texttt{ret}(r, \Delta_0, \Lambda_0, \mathcal{C})$*, where* $\Delta_0 = \Delta$ *and* $\Lambda_0 = \Lambda$*. Let* $\Gamma_{k+1} = \bigcup_{y \in \Gamma_k} \texttt{ret}(y, \Delta_{k+1}, \Lambda_{k+1}, \mathcal{C})$*, where* $\Delta_{k+1} = \Delta_k - \Gamma_k$ *and* $\Lambda_{k+1} = \Lambda_k \cup \Gamma_k$*. For any* $i \in \mathbb{N}$*,* $\Gamma_i \subseteq \texttt{ret}(r, \Delta, \Lambda, \mathcal{C})$.

The third item needs some special attention. It shows that the ret is 'closed' in the sense that any norm that can be retrieved by the norms retrieved by r are also retrieved by r, due to the transitivity of support relation.

Definition 16 (Vagueness removal)**.** *Let* $\mathcal{L}$ *be a logical language,* $\mathcal{R}^s$ *the set of strict rules reflecting the monotonic inferences over* $\mathcal{L}$*,* $\mathcal{C}$ *a set of contexts, and* Δ *a defeasible rule set.The vagueness removal* $\texttt{rem}^{\texttt{v}}$ *on* Δ *with respect to* $\mathcal{C}$ *is defined as follows.*

$$\begin{aligned}\texttt{rem}^{\texttt{v}}(\Delta, \mathcal{C}) = \{x \in \Delta \mid\ & \textit{there is a } y \in \Delta, \textit{such that}\\ & x \leftrightsquigarrow_c y \textit{ for some } c \in \mathcal{C} \textit{ under } (\mathcal{L}, \mathcal{R}^s, \Delta, \mathcal{C})\\ & \textit{and it is not the case that either}\\ & \texttt{body}(x) \rightarrow \texttt{body}(y) \in \mathcal{R}^s \textit{ or}\\ & \texttt{body}(y) \rightarrow \texttt{body}(x) \in \mathcal{R}^s\}\end{aligned}$$

[2] This can prevent an irrational belief change: using a norm which is essentially equivalent to another norm (already deemed unreliable) as input. It reconsiders the latter without any reason.

Proposition 9. *Let $\mathcal{L}$ be a logical language, $\mathcal{R}^s$ the set of strict rules reflecting monotonic inferences over $\mathcal{L}$, $\mathcal{C}$ a set of contexts, and Δ a defeasible rule set, then:*

1. *For any $x, y \in \Delta$ that $x \leftrightsquigarrow_c y$ for some $c \in \mathcal{C}$ under $(\mathcal{L}, \mathcal{R}^s, \Delta, \mathcal{C})$, $x \in \mathtt{rem}(\Delta, \mathcal{C})$ if and only if $y \in \mathtt{rem}(\Delta, \mathcal{C})$.*
2. *If Δ is structurally coherent under $(\mathcal{L}, \mathcal{R}^s, \Delta, \mathcal{C})$, then $\mathtt{rem}^{\mathtt{v}}(\Delta, \mathcal{C}) = \emptyset$.*
3. *$\Delta - \mathtt{rem}^{\mathtt{v}}(\Delta, \mathcal{C})$ is structurally coherent under $(\mathcal{L}, \mathcal{R}^s, \Delta - \mathtt{rem}^{\mathtt{v}}(\Delta, \mathcal{C}), \mathcal{C})$.*

The ground removal is rather special, it is defined with respect to a theory construction process. We will formulate ground removal when defining the theory construction process.

Definition 17 (Theory construction through reflective equilibrium)**.** *Let $Q = (0, \ldots, n)$ be a sequence of construction stages. For each stage $i \in Q$, denote the input rule $c_i \Rightarrow \odot\psi$ as r_i. The prioritized contextual argumentation theory at stage i, notation $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}_i^d, \mathcal{C}_i, \trianglelefteq_i)$ where $\mathcal{R}_i^d = \mathcal{R}_i^{d+} \cup \mathcal{R}_i^{d-}$, is subject to inductive construction as follows.*

- *$i = 0$. In this case: $\mathcal{C}_0 = \emptyset$, $\mathcal{R}_0^{d+} = \emptyset$, $\mathcal{R}_0^{d-} = \emptyset$, $\trianglelefteq_0 = \emptyset$.*
- *$i = k + 1 (0 \leq k < n)$. In this case:*
  - *$C_{k+1} = C_k \cup \{c_{k+1}\}$*
  - *$\mathcal{R}_{k+1}^d = \mathcal{R}_{k+1}^{d+} \cup \mathcal{R}_{k+1}^{d-}$, where:*
    - *If $\mathcal{R}_k^{d+} \cup \{r_{k+1}\}$ is structurally coherent under $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}_k^{d+} \cup \{r_{k+1}\}, \mathcal{C}_{k+1})$, then:*

$$\begin{aligned}\mathcal{R}_{k+1}^{d+} =& \mathcal{R}_k^{d+} \cup \{r_{k+1}\} \cup \\ & \mathtt{ret}(r_{k+1}, \mathcal{R}_k^{d-}, \mathcal{R}_k^{d+}, \mathcal{C}_{k+1}) \\ \mathcal{R}_{k+1}^{d-} =& \mathcal{R}_k^{d-} - \mathtt{ret}(r_{k+1}, \mathcal{R}_k^{d-}, \mathcal{R}_k^{d+}, \mathcal{C}_{k+1}).\end{aligned}$$

    - *Else:*
      (1) if $r_{k+1} \in \mathtt{rem}^{\mathtt{v}}(\{r_{k+1}\} \cup \mathcal{R}_k^{d+}, \mathcal{C}_{k+1})$, then

$$\begin{aligned}\mathcal{R}_{k+1}^{d+} =& (\mathcal{R}_k^{d+} - \mathtt{rem}^{\mathtt{v}}(r_{k+1} \cup \mathcal{R}_k^{d+}, \mathcal{C}_{k+1})) - \\ & \mathtt{rem}^{\mathtt{g}}(\mathcal{R}_k^{d+}, Q) \\ \mathcal{R}_{k+1}^{d-} =& \mathcal{R}_k^{d-} \cup \mathtt{rem}^{\mathtt{v}}(r_{k+1} \cup \mathcal{R}_k^{d+}, \mathcal{C}_{k+1}) \cup \\ & \mathtt{rem}^{\mathtt{g}}(\mathcal{R}_k^{d+}, Q)\end{aligned}$$

      (2) if $r_{k+1} \notin \mathtt{rem}^{\mathtt{v}}(\{r_{k+1}\} \cup \mathcal{R}_k^{d+}, \mathcal{C}_{k+1})$, then

$$\begin{aligned}\mathcal{R}_k^{d+'} =& (\mathcal{R}_k^{d+} - \mathtt{rem}^{\mathtt{v}}(\{r_{k+1}\} \cup \mathcal{R}_k^{d+}, \mathcal{C}_{k+1})) - \\ & \mathtt{rem}^{\mathtt{g}}(\mathcal{R}_k^{d+}, Q) \\ \mathcal{R}_{k+1}^{d+} =& \mathcal{R}_k^{d+'} \cup \{r_{k+1}\} \cup \\ & \mathtt{ret}(r_{k+1}, \mathcal{R}_k^{d-}, \mathcal{R}_k^{d+'}, \mathcal{C}_{k+1}) \\ \mathcal{R}_{k+1}^{d-} =& (\mathcal{R}_k^{d-} - \mathtt{ret}(r_{k+1}, \mathcal{R}_k^{d-}, \mathcal{R}_k^{d+'}, \mathcal{C}_{k+1})) \cup \\ & \mathtt{rem}^{\mathtt{v}}(\{r_{k+1}\} \cup \mathcal{R}_k^{d+}, \mathcal{C}_{k+1}) \cup \mathtt{rem}^{\mathtt{g}}(\mathcal{R}_k^{d+}, Q)\end{aligned}$$

      where $\mathtt{rem}^{\mathtt{g}}$ is the ground removal on $\mathcal{R}_k^{d+}$ defined as follows:

$$\begin{aligned}&\mathtt{rem}^{\mathtt{g}}(\mathcal{R}_k^{d+}, Q) = \{x \in \mathcal{R}_k^{d+} \mid \text{there is a } j \in Q \\ &\text{such that } j \leq k \text{ with } x \in \mathtt{ret}(r_j, \mathcal{R}_{j-1}^{d-}, \mathcal{R}_{j-1}^{d+}, \mathcal{C}_j) \\ &\text{and } x \in \mathcal{R}_{[j,k]}^{d+}, \text{ and for any } y \in \mathcal{R}_k^{d+}, \text{if } y \gg x \\ &\text{then } y \in \mathtt{rem}^{\mathtt{v}}(r_{k+1} \cup \mathcal{R}_k^{d+}, \mathcal{C}_{k+1})\}\end{aligned}$$

  - *$\trianglelefteq_{k+1} = \trianglelefteq_{k+1}^{\mathtt{f}} \cup \{(x, y) \mid x, y \in \mathcal{R}_{k+1}^d$ and $x \trianglelefteq_{k+1}^{\mathtt{s}} y$ but not $y \trianglelefteq_{k+1}^{\mathtt{f}} x\}$, where $\trianglelefteq_{k+1}^{\mathtt{f}}$, $\trianglelefteq_{k+1}^{\mathtt{s}}$ and respectively the firmness and specificity ordering defined on $\mathcal{R}_{k+1}^d$ according to definition 11, 13.*

Proposition 10. *Let Q be a sequence of constructions, and $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}_i^d, \mathcal{C}_i, \trianglelefteq_i)$ where $\mathcal{R}_i^{d+} = \mathcal{R}_i^{d+} \cup \mathcal{R}_i^{d-}$ be defined as in Definition 17. For each $i \in Q$:*

1. *$\mathcal{R}_i^{d+} = \bigcup\{r_i\}$*
2. *$\mathcal{R}_i^{d+} \cap \mathcal{R}_i^{d-} = \emptyset$.*
3. *$\mathcal{R}_i^{d+}$ is structurally coherent under $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}_i^{d+}, \mathcal{C}_i)$.*
4. *$\trianglelefteq_i$ is a preorder.*

Can such construction process leads to a well-structured moral theory, and give definite moral answers? The following theorem says yes! With regard to those contexts considered in the construction, so long as they can make use of norms in the identified firm set to deal with moral disagreement, then the relevant moral theory is always well-structured, and therefore leads to well-founded argumentation framework whose extensions are all grounded. This exactly matches the core idea of reflective equilibrium: by finding principles that support our moral judgments, we make clear of our moral sense and can give definite answer by way of this clear part of moral sense.

Theorem 3. *Let $Q = (0, \ldots, n)$ be a construction sequence, and PCAT$_i = (\mathcal{L}, \mathcal{R}^s, \mathcal{R}_i^d, \mathcal{C}_i, \trianglelefteq_i)$ be constructed at stage i according to Definition 17. CAF$_i^c = (\mathcal{A}^c, \mathcal{D}^c)$ is well-founded, for any set of contexts $\mathcal{C} \subseteq \mathcal{C}_i$ satisfying:*

for any $c \in \mathcal{C}$ and any $A, A' \in \mathcal{A}^c$, if A attack A' then $\mathtt{DefRule}(A) \in \mathcal{R}_i^{d+}$ or $\mathtt{DefRule}(A') \in \mathcal{R}_i^{d+}$

where CAF$_i'^c = (\mathcal{A}^c, \mathcal{D}^c)$ is defined by PCAT$_i' = (\mathcal{L}, \mathcal{R}^s, \mathcal{R}_i^d, C, \trianglelefteq_i)$ with respect to a $c \in \mathcal{C}$, according to Definition 6 and Definition 9.

Now we have one last thing to verify. Suppose that in a community, there are a set of moral norms that are uncontroversially followed and practiced by agents within the community. Can these norms be identified by our method of theory construction? Again, the answer is yes! Moreover, the positive result is also independent of the order that the judgments expressing the norms are input to construction! This means that our method has path-independence with regard to identifying a 'clear' morality existing in a community. This result is shown in the final theorem, with the aid of some auxiliary definitions.

Definition 18 (Conflict closure)**.** *Let CAT $= (\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C})$. A set $\Delta \subseteq \mathcal{R}^d$ is a conflict closure under CAT if and only if, for any $x \in \Delta$ if there is a $y \in \mathcal{R}^d$ such that $x \leftrightsquigarrow_c y$ for some $c \in \mathcal{C}$ then $y \in \Delta$.*

Definition 19 (Coherent sub-structure)**.** *Let $CAT = (\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C})$. A set $\Delta \subseteq \mathcal{R}^d$ is a coherent sub-structure with respect to $\mathcal{C}$, if Δ is a conflict closure under CAT and structural coherent under $CAT' = (\mathcal{L}, \mathcal{R}^s, \Delta, \mathcal{C})$.*

In order to represent an input sequence, we introduce an input function that labels each defeasible rule with a natural number. Different input functions over the same set of defeasible rules thus represent different input sequences.

Definition 20 (Input function)**.** *Let Δ be a finite set of distinct defeasible rules. Let $Q = (0, \ldots, |\Delta|)$. An input function f regarding (Δ, Q), is a bijection from the elements of Q other than 0 to Δ.*

Given an input function f, we denote $f(i)$ as r_i, signifying the input rule at stage $i \in Q$.

Theorem 4. *Let Δ be a finite set of distinct defeasible rules, and $\mathcal{C}_\Delta = \{c_x \mid x \in \Delta\}$ be the set of contexts appear in Δ. Let $Q = (0, \ldots, |\Delta|)$ be a sequence of construction, and $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d_i, \mathcal{C}_i, \trianglelefteq_i)$ be the theory constructed at stage $i \in Q$ according to Definition 17. For any input function f regarding (Δ, Q) and any coherent sub-structure $\Lambda \subseteq \Delta$ with respect to $\mathcal{C}_\Delta$,*

- $\Lambda \subseteq \mathcal{R}^{d+}_{|\Delta|}$
- $\trianglelefteq^{\mathtt{s}\text{-}\Lambda} \subseteq \trianglelefteq_{|\Delta|}$ *(where $\trianglelefteq^{\mathtt{s}\text{-}\Lambda}$ is the specificity ordering on Λ).*

Conclusion

This paper addresses the fundamental challenge of norm identification in open multi-agent systems, where agents cannot rely on pre-programmed knowledge. We have proposed a logic-based approach that frames norm identification as a problem of consistent moral theory construction. We adopt the method of reflective equilibrium, representing a moral theory as an argumentation theory. This framework inductively incorporates moral judgments as *prima facie* norms that are input from the environment and constructs a priority structure through specificity and firmness orderings. The resulting theory achieves a maximally consistent set of norms capable of justifying an agent's decisions through derived arguments. We have demonstrated that, whenever the firm part of this constructed theory is applicable, the agent can derive a definitive moral answer. Furthermore, given a set of uncontroversial moral norms in a community, our method guarantees the path-independent identification of them.

The primary contributions of this work are twofold. First, it provides a norm identification method that is robust to inconsistent data and inherently supplies justifications for an agent's norm-based decisions. Second, by enabling direct, real-time norm identification during execution—rather than requiring a separate training phase—our logic-based approach enhances the feasibility of deploying autonomous agents in truly open, dynamic, and unknown environments. This bridges a critical gap between the assumed awareness of norms in normative reasoning and the decentralized reality of multi-agent systems.

However, several promising directions remain for further investigation to enhance its applicability and robustness. For example, while our method requires agents to resolve normative conflicts once a *prime facie* is input to construction, we might allow agents to perform the conflict resolution only once every certain number of inputs. This "batch processing" approach would improve computational efficiency by amortizing the cost of reasoning over multiple observations, better simulating how humans accumulate experience before revising their moral judgments.

References

Akinkunmi, B. O.; and Babalola, F. M. 2020. A norm enforcement mechanism for a time-constrained conditional normative framework. *Autonomous Agents and Multi-Agent Systems*, 34(1): 20.

Awad, E.; Anderson, M.; Anderson, S. L.; and Liao, B. 2020. An approach for combining ethical principles with public opinion to guide public policy. *Artificial Intelligence*, 287: 103349.

Broersen, J.; Dastani, M.; Hulstijn, J.; Huang, Z.; and Van Der Torre, L. 2001. The BOID architecture: conflicts between beliefs, obligations, intentions and desires. In *Proceedings of the fifth international conference on Autonomous agents*, 9–16.

Caminada, M.; and Amgoud, L. 2007. On the evaluation of argumentation formalisms. *Artificial Intelligence*, 171(5-6): 286–310.

Cranefield, S.; and Dhiman, A. 2021. Identifying Norms from Observation Using MCMC Sampling. In *Proc. of the 30th International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence*.

Criado, N.; Argente, E.; Noriega, P.; and Botti, V. 2014. Reasoning about norms under uncertainty in dynamic environments. *International Journal of Approximate Reasoning*, 55(9): 2049–2070.

Dastani, M.; van der Torre, L.; and Yorke-Smith, N. 2017. Commitments and interaction norms in organisations. *Autonomous Agents and Multi-Agent Systems*, 31(2): 207–249.

Dell'Anna, D.; Dastani, M.; and Dalpiaz, F. 2020. Runtime revision of sanctions in normative multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 34(2): 43.

Dung, P. M. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2): 321–357.

Dworkin, R. 1986. *Law's Empire*. Cambridge, Mass. and London: The Belknap Press of Harvard University Press.

Johnston, B.; and Governatori, G. 2003. Induction of defeasible logic theories in the legal domain. In *Proceedings of the 9th international conference on Artificial intelligence and law*, 204–213.

Knobbout, M.; and Dastani, M. 2012. Reasoning under compliance assumptions in normative multiagent systems. In *AAMAS*, 331–340.

Knobbout, M.; Dastani, M.; and Meyer, J.-J. 2016. A dynamic logic of norm change. In *Proceedings of the Twenty-second European Conference on Artificial Intelligence*, 886–894.

Modgil, S.; and Prakken, H. 2013. A general account of argumentation with preferences. *Artificial Intelligence*, 195: 361–397.

Modgil, S.; and Prakken, H. 2014. The *ASPIC* $^+$ framework for structured argumentation: a tutorial. *Argument & Computation*, 5(1): 31–62.

Rawls, J. 1951. Outline of a Decision Procedure for Ethics. *The Philosophical Review*, 60(2): 177–197.

Rawls, J. 1974. The Independence of Moral Theory. *Proceedings and Addresses of the American Philosophical Association*, 48: 5–22.

Rawls, J. 1999. *A Theory of justice*. Cambridge, Mass.: The Belknap Press of Harvard University Press, revised edition edition. ISBN 0-674-00077-3.

Tolmeijer, S.; Kneer, M.; Sarasua, C.; Christen, M.; and Bernstein, A. 2020. Implementations in machine ethics: A survey. *ACM Computing Surveys (CSUR)*, 53(6): 1–38.

Woodgate, J.; Marshall, P.; and Ajmeri, N. 2025. Operationalising Rawlsian Ethics for Fairness in Norm Learning Agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 26382–26390.

Appendix

Proposition 1. *Let PCAT* $= (\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C}, \trianglelefteq)$. *Let SCAF*$^c = (\mathcal{A}^c, att, \preceq)$ *where* $c \in \mathcal{C}$ *and att is determined by* $\trianglelefteq$ *according to Definition 7. This produce the same argument ordering as those defined by weakest link or last link principle, using either elitist or democratic preference lifting in Modgil and Prakken (2014).*

Proof. Easy to verify by applying the weakest link or last lint principle coupled with either elitst or democratic preference lifting to determine $\preceq$. □

Proposition 2. *Argument* A *defeats argument* B *if and only if* A *attacks* B *and:*

- $\texttt{DefRule}(A) = \emptyset$*, or*
- $\texttt{DefRule}(A) \not\lhd \texttt{DefRule}(B)$.

Proof. Assume A attacks B. It then suffices to show that $A \not\prec B$ if and only if: $\texttt{DefRule}(A) = \emptyset$, or $\texttt{DefRule}(A) \not\lhd \texttt{DefRule}(B)$.

- $\Rightarrow$. Assume the contrary: $\texttt{DefRule}(A) =\neq \emptyset$ and $\texttt{DefRule}(A) \lhd \texttt{DefRule}(B)$. Then $A \preceq B$ and $B \not\preceq A$, namely $A \prec B$. Contradiction!
- $\Leftarrow$. We consider two cases separately. (1) Assume $\texttt{DefRule}(A) = \emptyset$. Then $A \not\preceq B$, and then $A \not\prec B$. (2) Assume $\texttt{DefRule}(A) \not\lhd \texttt{DefRule}(B)$. Then $\texttt{DefRule}(A) \not\trianglelefteq \texttt{DefRule}(B)$ or $\texttt{DefRule}(B) \trianglelefteq \texttt{DefRule}(A)$, and then $A \prec B$.

□

Theorem 1 (Satisfaction of rationality postulates)**.** *Let CAF*$^c = (\mathcal{A}^c, \mathcal{D}^c)$ *be a contextual argumentation framework based on Definition 9. For any complete extension* E *under the Dung's semantics,* E *satisfies the following rational postulates proposed by Caminada and Amgoud (2007).*

- *Subargument closure;*
- *Closure under strict rules;*
- *Direct consistency;*
- *Indirect consistency.*

Proof. According to (Modgil and Prakken 2014), the proof amounts to showing that *CAT* satisfies axiom consistency, transposition or contraposition, argument preference ordering is reasonable. The first is satisfied as the context is assumed to be consistent; transposition and contraposition is both satisfied given the propositional reasoning uses propositional calculus; the reasonableness of argument ordering is a direct result of Proposition 1. □

Proposition 3. *Let PCAT* $= (\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C}, \trianglelefteq)$ *be well-structured. Let* $c \in C$ *be a context and CAF*$^c = (\mathcal{A}^c, \mathcal{D}^c)$. *For any sequence of defeasible arguments* $A_1, \ldots, A_k, \ldots$ *from* $\mathcal{A}^c$*, if* A_{i+1} *defeats* A_i *for each* A_i *in the sequence* ($i \leq n-1$ *when the sequence only contains finite* n *arguments), then* $\texttt{DefRule}(A_i) \lhd \texttt{DefRule}(A_{i+1})$.

Proof. Assume A_{i+1} defeats A_i for each A_i in the sequence. Then $\texttt{Conc}(A_{i+1}) = -\psi$ for some $A'_i \in \texttt{Sub}(A_i)$ of the form $B_1, \ldots, B_n \Rightarrow \psi$, and $\texttt{DefRule}(A_{i+1}) \not\lhd \texttt{DefRule}(A'_i)$. It follows that $\texttt{DefRule}(A_{i+1}) \leftrightsquigarrow_c \texttt{DefRule}(A'_i)$, and furthermore $\texttt{DefRule}(A'_i) \lhd \texttt{DefRule}(A_{i+1})$ since *CAT* is well-structured. As each argument contains at most one defeasible rule, $\texttt{DefRule}(A'_i) = \texttt{DefRule}(A_i)$, and then $\texttt{DefRule}(A_i) \lhd \texttt{DefRule}(A_{i+1})$. □

Theorem 2. *Let PCAT* $= (\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C}, \trianglelefteq)$. *If PCAT is well-structured, then CAF*^{c} *is well-founded for any* $c \in \mathcal{C}$.

Proof. Assume there is some $c \in \mathcal{C}$, such that *CAF*^{c} is not well-founded. Then there is an infinite sequence of arguments $A_1, A_2, \ldots$, such that A_{i+1} defeats A_i. It is easy to see that such sequence contains no strict argument. There are two cases we need to consider.

- Case 1: No $r \in \mathcal{R}^d$ is used more than once in the infinite sequence (namely, $\texttt{DefRule}(A_j) \neq \texttt{DefRule}(A_k)$ for any A_j, A_k in the sequence with $j \neq k$). But then $\mathcal{R}^d$ must be infinite. Contradiction!
- Case 2: Some $r \in \mathcal{R}^d$ is used more than once in the infinite sequence. Then there are two argument A_j, A_k with $j < l < k$ or $k < l < j$ for some A_l between A_j, A_k in the sequence, and $\texttt{DefRule}(A_j) = \texttt{DefRule}(A_k) = r$. By proposition 3 and $\trianglelefteq$ being transitive, if $k < l < j$ then $\texttt{DefRule}(A_k) \lhd \texttt{DefRule}(A_j)$, if $j < l < k$ then $\texttt{DefRule}(A_j) \lhd \texttt{DefRule}(A_k)$. As $\texttt{DefRule}(A_j) = \texttt{DefRule}(A_k) = r$, in either case we have $r \lhd r$, namely $r \trianglelefteq$ and $r \not\trianglelefteq r$. Contradiction!

□

Proposition 4. *Let CAT $= (\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C})$. If $\mathcal{R}^d$is structurally coherent, then any $\Delta \subseteq \mathcal{R}^d$ is also structurally coherent.*

Proof. Straightforwardly by definition. □

Proposition 5. *Let PCAT $= (\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C}, \trianglelefteq)$. If $\mathcal{R}^d$ is structurally coherent and $\trianglelefteq$ is the specificity ordering on $\mathcal{R}^d$ according to Definition 11, then PCAT is well-structured.*

Proof. Straightforwardly by definition. □

Proposition 6. *Let CAT $= (\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C})$. The support relation $\gg$ on $\mathcal{R}^d$ is reflexive and transitive.*

Proof. Reflexivity is straightforward. For transitivity, assume $x \gg y$ and $y \gg z$, then: $\texttt{body}(y) \rightarrow \texttt{body}(x) \in \mathcal{R}^s$, $\texttt{body}(z) \rightarrow \texttt{body}(y) \in \mathcal{R}^s$, and therefore $\texttt{body}(z) \rightarrow \texttt{body}(y) \in \mathcal{R}^s$; $\texttt{head}(x) \rightarrow \texttt{head}(y) \in \mathcal{R}^s$, $\texttt{head}(y) \rightarrow \texttt{head}(z) \in \mathcal{R}^s$, and therefore $\texttt{head}(x) \rightarrow \texttt{head}(z) \in \mathcal{R}^s$. Together, $\texttt{body}(z) \rightarrow \texttt{body}(x) \in \mathcal{R}^s$ and $\texttt{head}(x) \rightarrow \texttt{head}(z) \in \mathcal{R}^s$, so $x \gg z$. □

Proposition 7. *Let $\mathcal{L}$ be a logical language, $\mathcal{R}^s$ the set of strict rules reflecting monotonic inferences over $\mathcal{L}$. Let $\mathcal{R}^d$ be a set of defeasible rules, and $\gg$ the support relation on $\mathcal{R}^d$ based on $\mathcal{R}^s$. For any $r \in \mathcal{R}^d$, the set $\{x \in \mathcal{R}^d \mid r \gg x\}$ contains no contextual conflict under $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d, \mathcal{C})$ for any set of contexts $\mathcal{C}$.*

Proof. Assume the contrary, then there is a $\mathcal{C}$ and some $x, y \in \{x \in \mathcal{R}^d \mid r \gg x\}$, such that $x \leftrightsquigarrow_c y$ for some $c \in C$. Then there are $A, A' \in \mathcal{A}^c$ with $\texttt{DefRule}(A) = \{x\}, \texttt{DefRule}(A') = \{y\}$ and $\texttt{Conc}(A) = -\texttt{Conc}(A')$. Therefore, $\texttt{head}(x) = -\texttt{head}(y)$, or there is a strict rule $\texttt{head}(x) \rightarrow -\texttt{head}(y) \in \mathcal{R}^s$. But given $r \gg x$, $\texttt{head}(r) \rightarrow \bot \in \mathcal{R}^s$, which contradicts the assumption that $\texttt{head}(r) \nvdash_{\mathcal{S}_\mathcal{L}} \bot$. □

Proposition 8. *Let $\mathcal{L}$ be a logical language, $\mathcal{R}^s$ the set of strict rules reflecting the monotonic inferences over $\mathcal{L}$. Let Δ, Λ be two defeasible rule sets, and r a defeasible rule, then:*

1. *For any set of contexts $\mathcal{C}$, $\texttt{ret}(r, \Delta, \Lambda, \mathcal{C})$ contains no contextual conflict under $(\mathcal{L}, \mathcal{R}^s, \texttt{ret}(r, \Delta, \Lambda, \mathcal{C}), \mathcal{C})$.*
2. *If Λ is structural coherent under $(\mathcal{L}, \mathcal{R}^s, \Lambda, \mathcal{C})$, then $\Lambda \cup \texttt{ret}(r, \Delta, \Lambda, \mathcal{C})$ is structural coherent under $(\mathcal{L}, \mathcal{R}^s, \Lambda \cup \texttt{ret}(r, \Delta, \Lambda, \mathcal{C}), \mathcal{C})$.*
3. *Let $\Gamma_0 = \texttt{ret}(r, \Delta_0, \Lambda_0, \mathcal{C})$, where $\Delta_0 = \Delta$ and $\Lambda_0 = \Lambda$. Let $\Gamma_{k+1} = \bigcup_{y \in \Gamma_k} \texttt{ret}(y, \Delta_{k+1}, \Lambda_{k+1}, \mathcal{C})$, where $\Delta_{k+1} = \Delta_k - \Gamma_k$ and $\Lambda_{k+1} = \Lambda_k \cup \Gamma_k$. For any $i \in \mathbb{N}$, $\Gamma_i \subseteq \texttt{ret}(r, \Delta, \Lambda, \mathcal{C})$.*

Proof. The first item is a corollary of proposition 7. The second item is straightforward by definition. The third item follows from the first item and the transitivity of $\gg$. □

Proposition 9. *Let $\mathcal{L}$ be a logical language, $\mathcal{R}^s$ the set of strict rules reflecting monotonic inferences over $\mathcal{L}$, $\mathcal{C}$ a set of contexts, and Δ a defeasible rule set, then:*

1. *For any $x, y \in \Delta$ that $x \leftrightsquigarrow_c y$ for some $c \in \mathcal{C}$ under $(\mathcal{L}, \mathcal{R}^s, \Delta, \mathcal{C})$, $x \in \texttt{rem}(\Delta, \mathcal{C})$ if and only if $y \in \texttt{rem}(\Delta, \mathcal{C})$.*
2. *If Δ is structurally coherent under $(\mathcal{L}, \mathcal{R}^s, \Delta, \mathcal{C})$, then $\texttt{rem}^\texttt{v}(\Delta, \mathcal{C}) = \emptyset$.*
3. *$\Delta - \texttt{rem}^\texttt{v}(\Delta, \mathcal{C})$ is structurally coherent under $(\mathcal{L}, \mathcal{R}^s, \Delta - \texttt{rem}^\texttt{v}(\Delta, \mathcal{C}), \mathcal{C})$.*

Proof. The first item is easy to verify, by observing that the $\leftrightsquigarrow_c$ is symmetric. The second item directly follows from the definition of structural coherence.

For the third item, assume the contrary. Then there are $x, y \in \Delta - \texttt{rem}^\texttt{v}(\Delta, \mathcal{C})$, such that $x \leftrightsquigarrow_c y$ under $(\mathcal{L}, \mathcal{R}^s, \Delta - \texttt{rem}^\texttt{v}(\Delta, \mathcal{C}), \mathcal{C})$ and not the case that either $\texttt{body}(x) \rightarrow \texttt{body}(y) \in \mathcal{R}^s$ or $\texttt{body}(y) \rightarrow \texttt{body}(x) \in \mathcal{R}^s$. Since $\Delta - \texttt{rem}^\texttt{v}(\Delta, \mathcal{C}) \subseteq \Delta$, it follows that the said $x, y \in \Delta$, with $x \leftrightsquigarrow_c y$ under $(\mathcal{L}, \mathcal{R}^s, \Delta, \mathcal{C})$ and not the case that either $\texttt{body}(x) \rightarrow \texttt{body}(y) \in \mathcal{R}^s$ or $\texttt{body}(y) \rightarrow \texttt{body}(x) \in \mathcal{R}^s$. Then $x, y \in \texttt{rem}^\texttt{v}(\Delta, \mathcal{C})$, furthermore $x, y \notin \Delta - \texttt{rem}^\texttt{v}(\Delta, \mathcal{C})$. Contradiction! □

Proposition 10. *Let Q be a sequence of constructions, and $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}^d_i, \mathcal{C}_i, \trianglelefteq_i)$ where $\mathcal{R}^{d+}_i = \mathcal{R}^{d+}_i \cup \mathcal{R}^{d-}_i$ be defined as in Definition 17. For each $i \in Q$:*

1. *$\mathcal{R}^{d+}_i = \bigcup\{r_i\}$*
2. *$\mathcal{R}^{d+}_i \cap \mathcal{R}^{d-}_i = \emptyset$.*
3. *$\mathcal{R}^{d+}_i$ is structurally coherent under $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}^{d+}_i, \mathcal{C}_i)$.*
4. *$\trianglelefteq_i$ is a preorder.*

Proof. The first and second items are easy to verify, by observing that at each $i \in Q$ the input rule r_i goes to either $\mathcal{R}^{d+}_i$ or $\mathcal{R}^{d-}_i$, and the retrieval /removal operations are moving a subset of $\mathcal{R}^{d-}_{i-1}$ to $\mathcal{R}^{d+}_i$ (respectively $\mathcal{R}^{d+}_{i-1}$ to $\mathcal{R}^{d-}_i$).

For the third item, there are two cases we need to consider.

- $i = 0$. In such case $\mathcal{R}^{d+}_0 = \emptyset$, whose structural coherence holds trivially.
- $i = k + 1 (0 \leq k < n)$. There are three sub-cases that we need to consider.
  - Case 1: $\mathcal{R}^{d+}_k \cup \{r_{k+1}\}$ is structurally coherent under $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}^{d+}_k \cup \{r_{k+1}\}, \mathcal{C}_{k+1})$. In this case, $\mathcal{R}^{d+}_{k+1} = \mathcal{R}^{d+}_k \cup \{r_{k+1}\} \cup \texttt{ret}(r_{k+1}, \mathcal{R}^{d-}_k, \mathcal{R}^{d+}_k, \mathcal{C}_{k+1})$. Assume that $\mathcal{R}^{d+}_{k+1}$ is not structurally coherent. Then there are $x, y \in \mathcal{R}^{d+}_{k+1}$, such that $x \leftrightsquigarrow_c y$ for some $c \in \mathcal{C}$ and it is not the case that either $\texttt{body}(x) \rightarrow \texttt{body}(y) \in \mathcal{R}^s$ or $\texttt{body}(y) \rightarrow \texttt{body}(x) \in \mathcal{R}^s$. If $x, y \in \mathcal{R}^{d+}_k \cup \{r_{k+1}\}$, then $x, y \in \mathcal{R}^{d+}_k \cup \{r_{k+1}\}$ is not structurally coherent, contradiction! If $x, y \in \texttt{ret}(r_{k+1}, \mathcal{R}^{d-}_k, \mathcal{R}^{d+}_k, \mathcal{C}_{k+1})$, then $\texttt{ret}(r_{k+1}, \mathcal{R}^{d-}_k, \mathcal{R}^{d+}_k, \mathcal{C}_{k+1})$ contains contextual conflict with respect to $\mathcal{C}_{k+1}$, which contradicts Proposition 8! If $x \in \mathcal{R}^{d+}_k \cup \{r_{k+1}\}$ and $y \in \texttt{ret}(r_{k+1}, \mathcal{R}^{d-}_k, \mathcal{R}^{d+}_k, \mathcal{C}_{k+1})$, by Proposition 7, $x \neq r_{k+1}$, therefore $x \in \mathcal{R}^{d+}_k$, which contradicts the definition of $\texttt{ret}(r_{k+1}, \mathcal{R}^{d-}_k, \mathcal{R}^{d+}_k, \mathcal{C}_{k+1})$!

– Case 2: $\mathcal{R}_k^{d+} \cup \{r_{k+1}\}$ is not structurally coherent under $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}_k^{d+} \cup \{r_{k+1}\}, \mathcal{C}_{k+1})$ and $r_{k+1} \in \texttt{rem}^{\texttt{v}}(r_{k+1} \cup \mathcal{R}_k^{d+}, \mathcal{C}_{k+1})$. In this case, $\mathcal{R}_{k+1}^{d+} = (\mathcal{R}_k^{d+} - \texttt{rem}^{\texttt{v}}(r_{k+1} \cup \mathcal{R}_k^{d+}, \mathcal{C}_{k+1})) - \texttt{rem}^{\texttt{g}}(\mathcal{R}_k^{d+}, Q)$. By Proposition 9, $\mathcal{R}_k^{d+} - \texttt{rem}^{\texttt{v}}(r_{k+1} \cup \mathcal{R}_k^{d+}, \mathcal{C}_{k+1})$ is coherent under $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}_k^{d+} - \texttt{rem}^{\texttt{v}}(r_{k+1} \cup \mathcal{R}_k^{d+}, \mathcal{C}_{k+1}), \mathcal{C}_{k+1})$. Since $\mathcal{R}_k^{d+} - \texttt{rem}^{\texttt{v}}(r_{k+1} \cup \mathcal{R}_k^{d+}, \mathcal{C}_{k+1})) - \texttt{rem}^{\texttt{g}}(\mathcal{R}_k^{d+}, Q) \subseteq \mathcal{R}_k^{d+} - \texttt{rem}^{\texttt{v}}(r_{k+1} \cup \mathcal{R}_k^{d+}, \mathcal{C}_{k+1})$, it follows that $\mathcal{R}_{k+1}^{d+}$ is structural coherent under $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}_{k+1}^{d+}, \mathcal{C}_{k+1})$.
– Case 3: $\mathcal{R}_k^{d+} \cup \{r_{k+1}\}$ is not structurally coherent under $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}_k^{d+} \cup \{r_{k+1}\}, \mathcal{C}_{k+1})$ and $r_{k+1} \notin \texttt{rem}^{\texttt{v}}(r_{k+1} \cup \mathcal{R}_k^{d+}, \mathcal{C}_{k+1})$. In this case, $\mathcal{R}_{k+1}^{d+} = \mathcal{R}_k^{d+'} \cup \{r_{k+1}\} \cup \texttt{ret}(r_{k+1}, \mathcal{R}_k^{d-}, \mathcal{R}_k^{d+'}, \mathcal{C}_{k+1})$ where $\mathcal{R}_k^{d+'} = (\mathcal{R}_k^{d+} - \texttt{rem}^{\texttt{v}}(r_{k+1} \cup \mathcal{R}_k^{d+}, \mathcal{C}_{k+1})) - \texttt{rem}^{\texttt{g}}(\mathcal{R}_k^{d+}, Q)$. By a proof similar to Case 2, $\mathcal{R}_k^{d+'}$ is structural coherent under $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}_k^{d+'}, \mathcal{C}_{k+1})$, furthermore, $\mathcal{R}_k^{d+'} \cup \{r_{k+1}\}$ is also structural coherent under $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}_k^{d+'} \cup \{r_{k+1}\}, \mathcal{C}_{k+1})$. By another proof similar to Case 1, $\mathcal{R}_{k+1}^{d+}$ is structurally coherent under $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}_{k+1}^{d+}, \mathcal{C}_{k+1})$.

For the fourth item. Regarding reflexivity: by $\mathcal{R}_i^{d+} \cap \mathcal{R}_i^{d-} = \emptyset$ it follows that for any $x \in \mathcal{R}_i^d$, $x \trianglelefteq_i^{\texttt{s}} x$ and not $x \trianglelefteq_i^{\texttt{f}} x$, therefore $x \trianglelefteq_i x$. Regarding transitivity: let $x, y, z \in \mathcal{R}_i^d$ such that $x \trianglelefteq_i y$ and $y \trianglelefteq_i z$, then: in the case that $x, y, z \in \mathcal{R}_i^{d+}$ (or $x, y, z \in \mathcal{R}_i^{d-}$), since $\mathcal{R}_i^{d+} \cap \mathcal{R}_i^{d-} = \emptyset$, it follows $x \trianglelefteq_i^{\texttt{s}} y$ and $y \trianglelefteq_i^{\texttt{s}} z$, and then $x \trianglelefteq_i^{\texttt{s}} z$ but not $z \trianglelefteq_i^{\texttt{f}} x$, therefore $x \trianglelefteq_i z$; in the case that some of x, y, z is in $\mathcal{R}_i^{d-}$ and some in $\mathcal{R}_i^{d+}$, by $\mathcal{R}_i^{d+} \cap \mathcal{R}_i^{d-} = \emptyset$ it follows $z \notin \mathcal{R}_i^{d-}$, then: if $x \in \mathcal{R}_i^{d-}$ and $y, z \in \mathcal{R}_i^{d+}$, then $x \trianglelefteq_i^{\texttt{f}} y$ and $x \trianglelefteq_i^{\texttt{f}} z$, therefore $x \trianglelefteq_i z$; if $x, y \in \mathcal{R}_i^{d-}$ and $z \in \mathcal{R}_i^{d+}$, then $x \trianglelefteq_i^{\texttt{f}} z$, therefore $x \trianglelefteq_i z$. □

Theorem 3. *Let $Q = (0, \ldots, n)$ be a construction sequence, and $PCAT_i = (\mathcal{L}, \mathcal{R}^s, \mathcal{R}_i^d, \mathcal{C}_i, \trianglelefteq_i)$ be constructed at stage i according to Definition 17. $CAF_i'^c = (\mathcal{A}^c, \mathcal{D}^c)$ is well-founded, for any set of contexts $\mathcal{C} \subseteq \mathcal{C}_i$ satisfying:*

for any $c \in \mathcal{C}$ and any $A, A' \in \mathcal{A}^c$, if A attack A' then $\texttt{DefRule}(A) \in \mathcal{R}_i^{d+}$ *or* $\texttt{DefRule}(A') \in \mathcal{R}_i^{d+}$

where $CAF_i'^c = (\mathcal{A}^c, \mathcal{D}^c)$ is defined by $PCAT_i' = (\mathcal{L}, \mathcal{R}^s, \mathcal{R}_i^d, C, \trianglelefteq_i)$ with respect to a $c \in \mathcal{C}$, according to Definition 6 and 9

Proof. Let $\mathcal{C}$ be a set of contexts that satisfy the said condition. We first show that $PCAT_i' = (\mathcal{L}, \mathcal{R}^s, \mathcal{R}_i^d, \mathcal{C}, \trianglelefteq_i)$ is well-structured. By Proposition 10, $\trianglelefteq_i$ is a preorder. Now assume two $x, y \in \mathcal{R}_i^d$, such that $x \leftrightsquigarrow_c y$ for some $c \in \mathcal{C}$. According to the said condition satisfied by $\mathcal{C}$, there are two cases we need to consider.

- Case 1: $x, y \in \mathcal{R}_i^{d+}$. By Proposition 10, $\mathcal{R}_i^{d+}$ is structural coherent under $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}_i^{d+}, \mathcal{C}_i)$, with $\mathcal{C} \subseteq \mathcal{C}_i$ this yields either $\texttt{body}(x) \rightarrow \texttt{body}(y)$ or $\texttt{body}(y) to \texttt{body}(x)$, then either $x \trianglelefteq_i^{\texttt{s}} y$ or $y \trianglelefteq_i^{\texttt{s}} x$, and then either $x \trianglelefteq_i y$ or $y \trianglelefteq_i x$.
- Case 2: $x \in \mathcal{R}_i^{d+}$ and $y \in \mathcal{R}_i^{d-}$ (or vice versa). Then $y \trianglelefteq_i^{\texttt{f}} x$, and then $y \trianglelefteq_i x$ (respectively $x \trianglelefteq_i y$ in the case that $y \in \mathcal{R}_i^{d+}$ and $x \in \mathcal{R}_i^{d-}$).

All in all, we have either $x \trianglelefteq_i y$ or $y \trianglelefteq_i x$. Therefore, $PCAT_i' = (\mathcal{L}, \mathcal{R}^s, \mathcal{R}_i^d, \mathcal{C}, \trianglelefteq_i)$ is well-structured. Then by Proposition 2, $CAF_i'^c = (\mathcal{A}^c, \mathcal{D}^c)$ is well-founded. □

Theorem 4. *Let Δ be a finite set of distinct defeasible rules, and $\mathcal{C}_\Delta = \{c_x \mid x \in \Delta\}$ be the set of contexts appear in Δ. Let $Q = (0, \ldots, |\Delta|)$ be a sequence of construction, and $(\mathcal{L}, \mathcal{R}^s, \mathcal{R}_i^d, \mathcal{C}_i, \trianglelefteq_i)$ be the theory constructed at stage $i \in Q$ according to Definition 17. For any input function f regarding (Δ, Q) and any coherent sub-structure $\Lambda \subseteq \Delta$ with respect to $\mathcal{C}_\Delta$, $\Lambda \subseteq \mathcal{R}_{|\Delta|}^{d+}$ and $\trianglelefteq^{\texttt{s-}\Lambda} \subseteq \trianglelefteq_{|\Delta|}$ (where $\trianglelefteq^{\texttt{s-}\Lambda}$ is the specificity ordering on Λ).*

Proof. Fix an input function f. First note that for any $r \in \Lambda$, there is an $i \in Q$ $(i > 0)$ such that $f(i) = r$. To show that $\Lambda \subseteq \mathcal{R}_{|\Delta|}^{d+}$, it suffices to that for any i and any $r \in \Lambda$ such that $f(i) = r$, $f(i) \in \mathcal{R}_{[i,|\Delta|]}^{d+}$ (i.e., $f(i) \in \mathcal{R}_i^{d+}$ and stay till $\mathcal{R}_{|\Delta|}^{d+}$).

Assume the contrary, then there is a $k \in Q$ such that $i \leq k$ and $f(i) \in \texttt{rem}^{\texttt{v}}(\{r_k\} \cup \mathcal{R}_{k-1}^{d+}, \mathcal{C}_k)$. Then there is a $r' \in \mathcal{R}_{k-1}^{d+}$, such that $f(i) \leftrightsquigarrow_c r'$ for some $c \in \mathcal{C}_k$ under $(\mathcal{L}, \mathcal{R}^s, \{r_k\} \cup \mathcal{R}_{k-1}^{d+}, \mathcal{C}_k)$ and it is not the case that either $\texttt{body}(f(i)) \rightarrow \texttt{body}(r') \in \mathcal{R}^s$ or $\texttt{body}(r') \rightarrow \texttt{body}(f(i)) \in \mathcal{R}^s$. Since $\mathcal{R}_{k-1}^{d+} \subset \Delta$ and $\mathcal{C}_k \subseteq \mathcal{C}_\Delta$, it follows $f(i) \leftrightsquigarrow_c r'$ for some $c \in \mathcal{C}_\Delta$ under $(\mathcal{L}, \mathcal{R}^s, \Delta, \mathcal{C}_\Delta)$. As Λ is a conflict closure with respect to $\mathcal{C}_\Delta$, $f(i), r' \in \Lambda$. But then Λ is not structural coherent under $(\mathcal{L}, \mathcal{R}^s, \Delta, \mathcal{C}_\Delta)$. Contradiction!

$\trianglelefteq^{\texttt{s-}\Lambda} \subseteq \trianglelefteq_{|\Delta|}$ immediately follows from $\Lambda \subseteq \mathcal{R}_{|\Delta|}^{d+}$, based on the way we define $\trianglelefteq_{|\Delta|}$ according to Definition 17. □