A Formalization of the Slippery Slope Argument

Zhe Yu(✉)

Institute of Logic and Cognition, Department of Philosophy,
Sun Yat-sen University, Guangzhou 510275, China
yuzh28@mail.sysu.edu.cn

Abstract. To bridge the gap between human reasoning and machine reasoning, one of the key problems in argumentation research is how to model natural language arguments by formal argumentation. The slippery slope argument (SSA) is a commonly used type of argument in the context of deliberation, with the intent of persuading people not to take a particular action. In this paper, an argumentation theory for the basic form of SSA is given based on the formal argumentation framework $ASPIC^+$ and argumentation schemes of SSA. Then, an SSA occurrence in a popular blog post about gene editing is taken as an example. By analyzing the case, this paper tries to model these arguments based on our argumentation theory and evaluates the arguments using abstract argumentation frameworks. The paper then points out that since whether an SSA is persuasive rests on whether its ultimate consequence is really unacceptable to the audience, value judgement should play an important role in the deliberation.

Keywords: Formal argumentation · Argumentation schemes · Slippery slope argument · Structured argumentation

1 Introduction

Argumentation is a cross-disciplinary topic involving multiple subjects such as philosophy, cognitive science, logic, linguistics and computer science. There are several research directions in the field of artificial intelligence, such as natural language processing and argumentation mining, that can be combined with argumentation and benefits from it [5]. As an approach for non-monotonic reasoning, formal argumentation is promising to bridge the gap between human reasoning and machine reasoning. To achieve this goal, a key problem is how to model natural language arguments by formal argumentation.

I would like to thank the anonymous reviewers for their valuable comments. This work is supported by the China Postdoctoral Science Foundation [No. 2019M663353] and the MOE Project of Key Research Institute of Humanities and Social Sciences in Universities [No. 18JJD720005].

M. Dastani et al. (Eds.): CLAR 2020, LNAI 12061, pp. 346–361, 2020.
https://doi.org/10.1007/978-3-030-44638-3_21

Based on this concern, argumentation schemes can be seen as a "semi-formal" generalization of arguments [18]. Many researchers have shown their interests in the formalization of argumentation schemes, such as the concerns for the argumentation scheme of argument from expert opinion [1,9,16].

In [20], Walton mentioned that the slippery slope argument (SSA), as a subclass of argument from negative consequences, is commonly used in the context of deliberation, with the intent of persuading people not to take an action that is under consideration. Here is an interpretation of the possible applications of SSA, taken from a book on informal logic [19].

Example 1. "You may hear such arguments in court. For example, the prosecuting attorney may encourage you (the jury) to be stern, severe, and courageous and not to shrink from your duty of demanding severe punishment for this guilty defendant; otherwise, this crime will be unpunished, criminals will run amok, and the social fabric of society will be threatened."

Though has been introduced in many logic textbooks as a sort of fallacy, there is also a lot of researchers hold the opinion that slippery slope arguments can be legitimate if good reasons are given for deeming that the first action will lead to catastrophic consequences [10,11,19,20]. Typically, SSA can be found in the discussions about legal, biomedical, and ethical issues. For instance, the topics of abortion, gay marriage, euthanasia, human gene therapy, etc.

This paper aims to formalise slippery slope arguments based on formal argumentation theory, and discuss if we can evaluate a slippery slope argument using formal methods. Firstly, by consulting the argumentation schemes for slippery slope argument presented by Walton [21,22], we give a formal model of slippery slope argument based on the structured argumentation framework $ASPIC^+$ [14,15]. Afterwards, we attempt to give a formal definition of the Critical Questions for slippery slope argument schemes, thus bring the informal way for evaluating a slippery slope argument into our theory. Meanwhile, we point out that the value judgement is an important factor in the evaluation of a slippery slope argument. For illustration, this paper models an application of the slippery slope argument found in a popular blog post using our argumentation theory.

The rest of this paper is structured as follows. In Sect. 2, we first summarize the basic features of SSA according to Walton's basic argumentation scheme for SSA. Then an argumentation theory for SSA (called SSAT) based on a formal argumentation system is constructed. After that, we try to define the Critical Questions for evaluation of SSA. In Sect. 3, we analyze an SSA from nature language text, and model it by SSAT. In Sect. 4, we briefly discuss some key ideas of this paper and list several related works, while in Sect. 5 we summarize this paper.

2 Argumentation Theory for SSA

In this section, we model the slippery slope argument based on Walton's basic scheme for this kind of argument and the structured argumentation framework $ASPIC^+$.

2.1 Basic Components of SSA

Several kinds of SSA as well as their schemes have been mentioned in [10,20,23], such as the Causal Slippery Slope Argument, the Sorites Slippery Slope Argument, etc. In [21], Walton gives a basic scheme for SSA, intending to capture the basic features of SSA. He also emphasized that "there are factors that help to propel the argument and series of consequences along the sequence, making it progressively harder for the agent to resist continuing to move ahead". These factors have been called "Drivers" [21].

Based on Walton's interpretation, in this paper we use 'a_0' to denote an action under consideration, 'a_n' to denote a catastrophic outcome; 'a_1, a_2, ..., a_x, ..., a_y' denotes a sequence of action or events between 'a_0' and 'a_n', each causes the next one, and 'd_i' ($i = 1, 2, 3, \ldots$) denotes the drivers. Then we can set out that an SSA has the following 8 basic components:

1. An initial event/action a_0.
2. A sequence of events/actions: a_0, a_1, a_2, ..., a_x, ..., a_y, ..., a_n. As the sequence proceeds, the consequences tend to become more serious.
3. Drivers: d_i. Catalyst that helps to propel the argument along the sequence in the argument. Drivers could be factors like precedent, public acceptance, vagueness, climate of social opinion, public acceptance, etc. [21]
4. Gray area: the area that starts at an undetermined point x (denoted by a_x), and end at another undetermined point y (denoted by a_y). In this area a slippery slope argument is turning form controllable to uncontrollable.
5. Controllable area: the area between the initial event/action and the gray area.
6. Uncontrollable area: the area between the gray area and the catastrophic consequence.
7. Catastrophic consequence: a_n, which should be avoided if possible.
8. Conclusion: not to take the initial step a_0.

According to this summarization, the developing process of an SSA can be illustrated by Fig. 1.[1]

2.2 SSAT

Our current work is mainly based on the structured argumentation framework $ASPIC^+$, which is proposed by Prakken et al. in [14]. $ASPIC^+$ is not a system but a framework, so that people can specify or extend it as an instantiation, as long as meeting some specific requirements.

Based on the above analysis of SSA, we can define an argumentation theory for SSA. First of all, an argumentation theory starts with a logical language $\mathcal{L}$. Since an SSA always leads to a negative consequence, we add a symbol "$\perp$" into the language of the argumentation theory, which denotes "bad/unwanted (consequence)".

[1] In a proper SSA, Drivers should always exist within every step. Here we write d_1, d_2 and d_3 as an example.

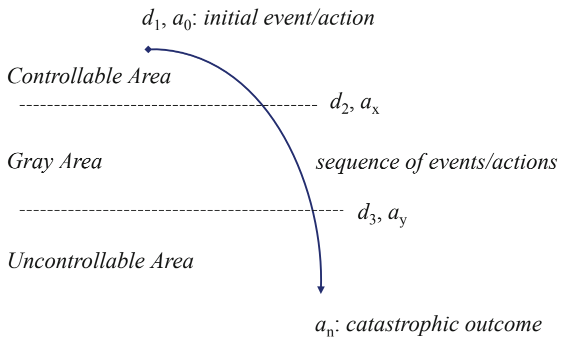
d_1, a_0: initial event/action
Controllable Area
d_2, a_x
Gray Area
sequence of events/actions
d_3, a_y
Uncontrollable Area
a_n: catastrophic outcome

Fig. 1. Process of an SSA

What's more, we divide the rules used in the SSA into two kinds: slippery slope rules and consequence judgements rules, denote as $\mathcal{R}_{sl}$ and $\mathcal{R}_j$ respectively. The slippery slope rules are always defeasible, and that's the reason an SSA is "sloping"; on the contrary, since an SSA must include a bad/unwanted consequence, the consequence judgements rules are always strict.

Then a knowledge base $\mathcal{K}$ is needed, which contains the premise sets of an argumentation theory, and from which we can proceed to build arguments. We put "$\neg\bot$" into the premise set, because if something is bad/unwanted, people are supposed to resist it instinctively. As for the premises of the SSA, the initial step is more like a presumption, or something that is still under consideration, so that we use $\mathcal{K}_0 = \{a_0, b_0, c_0, \ldots\}$ to denote the set of initial actions/events.[2] Since an argument $A = a_0$ represents a pending event or action, if A attacks other arguments without any supporter, it seems counterintuitive. Conversely, if A is not attacked by any other argument, it should be acceptable. Meanwhile, there is no reason not to accept any argument that depend on A, otherwise the entire SSA and its sub-arguments would be unacceptable.

Last but not least, we use C to denote the set of actions/events, and D to denote the set of drivers.

An argumentation theory for SSA can be defined as follows.

Definition 1 (SSAT). *A slippery slope argumentation theory (SSAT) is a tuple $SSAT = (\mathcal{L}, ^{-}, \mathcal{R}, n, \mathcal{K}, C, D)$, where:*

[2] The idea of $\mathcal{K}_0$ was inspired by Prakken [15], where the knowledge base $\mathcal{K}$ consists of 4 disjoint subsets: $\mathcal{K}_n$, $\mathcal{K}_p$, $\mathcal{K}_a$, $\mathcal{K}_i$, which are respectively the sets of axioms, ordinary premises, assumptions, and issues. The definitions of the axioms and the ordinary premises are the same as in this paper, while attacks on the assumptions are always succeed, and an issue must always be backed with a further argument. However, since the initial premise in the SSA is not only an event/action under consideration, but also the premise of the SSA and its sub-arguments, none of the above premise sets is particularly suitable as the set of initial premises of the SSA.

- *$\mathcal{L}$ is a logical language; $\bot \in \mathcal{L}$.*
- *$^-$ is a function from $\mathcal{L}$ to $2^{\mathcal{L}}$, such that*
  1. *φ is a contrary of ψ if $\varphi \in \overline{\psi}$, $\psi \notin \overline{\varphi}$;*
  2. *φ is a contradictory of ψ, if $\varphi \in \overline{\psi}$, $\psi \in \overline{\varphi}$ (denoted by '$\varphi = -\psi$');* [3]
  3. *each $\varphi \in \mathcal{L}$ has at least one contradictory.*
- *n is a partial function such that n: $\mathcal{R}_d \rightarrow \mathcal{L}$.*
- *$\mathcal{R} = \mathcal{R}_s \cup \mathcal{R}_d$ is a set of strict ($\mathcal{R}_s$) and defeasible ($\mathcal{R}_d$) inference rules of the form $\varphi_1, \ldots, \varphi_n \rightarrow \varphi$ and $\varphi_1, \ldots, \varphi_n \Rightarrow \varphi$ respectively (φ_i, φ are elements in $\mathcal{L}$), and $\mathcal{R}_s \cap \mathcal{R}_d = \emptyset$. $r_s l \in \mathcal{R}_{sl} \subseteq \mathcal{R}_d$ is slippery slope rule of the form $\varphi_1, \ldots, \varphi_n \Rightarrow_{sl} \varphi$, $\mathcal{R}_{sl} \neq \emptyset$; $r_j \in \mathcal{R}_j \subseteq \mathcal{R}_s$ is consequence judging rule of the form $r_j = \varphi_1, \ldots, \varphi_n \rightarrow_j \bot$, $\mathcal{R}_j \neq \emptyset$.*
- *$\mathcal{K} \subseteq \mathcal{L}$ is a knowledge base in an argumentation system, consisting of three disjoint subsets $\mathcal{K}_n$, $\mathcal{K}_p$ and $\mathcal{K}_0$ (i.e. $\mathcal{K} = \mathcal{K}_n \cup \mathcal{K}_p \cup \mathcal{K}_0$), where:*
  1. *$\mathcal{K}_n$ is a set of axioms;*
  2. *$\mathcal{K}_p$ is a set of the ordinary premises, such that $\neg\bot \in \mathcal{K}_n \cup \mathcal{K}_p$;*
  3. *$\mathcal{K}_0$ is a set of initial steps in a slippery slope argument of the form $\mathcal{K}_0 = \{a_0, b_0, c_0, \ldots\}$, where a_0, b_0, c_0 are initial actions or events.*
- *C is a set of actions or events in a slippery slope argument of the form $C = \{a_0, \ldots, a_n, b_0, \ldots, b_m, c_0, \ldots, c_q, \ldots\} \subseteq \mathcal{L}$, where a_i, b_j, c_k are actions or events; $\mathcal{K}_0 \subseteq C$.*
- *D is a set of drivers, $D = \{d_1, \ldots, d_n\} \subseteq \mathcal{K}_p$, where d_i is a driver.*

We use $Prem(A)$ to denote all the formulas of $\mathcal{K}$ that used to build an argument A, $Conc(A)$ to denote the conclusion of A, $Sub(A)$ to denote all the sub-arguments of A, $DefRule(A)$ to denote all the defeasible rules of A, and $TopRule(A)$ to denote the last rule of A. Depending on $ASPIC^+$, an argument in $SSAT$ can be defined as follows.

Definition 2 (Arguments). *An argument A on the basis of an $SSAT = (\mathcal{L}, ^-, \mathcal{R}, n, \mathcal{K}, C, D)$ is defined as:*

1. *φ if $\varphi \in \mathcal{K}$ with: $Prem(A) = \{\varphi\}$, $Conc(A) = \varphi$, $Sub(A) = \{\varphi\}$, $DefRules(A) = \emptyset$, $TopRule(A) = undefined$.*
2. *$A_1, \ldots, A_n \rightarrow \psi$ if $A_1, \ldots, A_n$ $(n \geq 1)$ are arguments such that there exists a strict rule $Conc(A_1), \ldots, Conc(A_n) \rightarrow \psi$ in $\mathcal{R}_s$ with: $Prem(A) = Prem(A_1) \cup \ldots \cup Prem(A_n)$; $Conc(A) = \psi$; $Sub(A) = Sub(A_1) \cup \ldots \cup Sub(A_n) \cup \{A\}$; $DefRules(A) = DefRules(A_1) \cup \ldots \cup DefRules(A_n)$; $TopRule(A) = Conc(A_1) \ldots Conc(A_n) \rightarrow \psi$.*
3. *$A_1, \ldots, A_n \Rightarrow \psi$ if $A_1, \ldots, A_n$ $(n \geq 1)$ are arguments such that there exists a defeasible rule $Conc(A_1), \ldots, Conc(A_n) \Rightarrow \psi$ in $\mathcal{R}_d$ with: $Prem(A) = Prem(A_1) \cup \ldots \cup Prem(A_n)$; $Conc(A) = \psi$; $Sub(A) = Sub(A_1) \cup \ldots \cup Sub(A_n) \cup \{A\}$; $DefRules(A) = DefRules(A_1) \cup \ldots \cup DefRules(\alpha_n) \cup \{Conc(A_1), \ldots, Conc(A_n) \Rightarrow \psi\}$; $TopRule(A) = Conc(A_1) \ldots Conc(A_n) \Rightarrow \psi$.*

[3] For all $\varphi \in \mathcal{L}$, we have $\neg - \varphi \in \overline{\varphi}$ and for all $\neg\varphi \in \mathcal{L}$, we have $\varphi \in \overline{\neg\varphi}$.

According to Walton [20–22], an integrated SSA should consists of two main lines, one from the initial action a_0 to a catastrophic consequence, and the other from the undesirability of the catastrophic consequence to the final conclusion ($\neg a_0$). However, from Example 1 we can see that in practical applications, the proponent of an SSA may only state the first line explicitly. If the SSA is used properly, the audiences will automatically infer the second line through rational intuition, and draw a conclusion $\neg a_0$. Since the main focus of this paper is on SSAs expressed in natural language, we consider an arguments containing the components in the first line as an SSA. Therefore, we define an SSA in the SSAT as follows.

Definition 3 (SSA). *If an argument A in $SSAT = (\mathcal{L}, ^-, \mathcal{R}, n, \mathcal{K}, C, D)$, such that: $Prem(A) \cap \mathcal{K}_0 \neq \emptyset$, $Prem(A) \cap D \neq \emptyset$, $SlRule(A) \neq \emptyset$, $JRule(A) \neq \emptyset$, $Conc(A) = \perp$, for every $A' \in Sub(A)$ and $A' \neq A$, $Conc(A') \in C \cup D$, then A is a slippery slope argument (SSA).*

Note that Definition 3 is not strictly corresponding to Walton's basic scheme of SSA, for it does not include the final conclusion. We have two reasons for this. On the one hand, with this definition, we can better identify an SSA, for the conclusion of an SSA is omitted in many cases. On the other hand, based on the current argumentation theory, an argument with the conclusion $\neg a_0$ would otherwise attack its sub-argument with the conclusion a_0 (see Definition 5). As a result, the SSA will cause inconsistency and cannot be accepted.

By claiming that the bad outcome is unacceptable, the slippery slope argument always attempt to draw a conclusion that the initial step should not be taken. To capture this feature, in addition to transposition under strict rules required by $ASPIC^+$, we define a "weak transposition" for the slippery slope rule used in the SSA.

Definition 4 (Transposition and Weak Transposition). *Let $SSAT = (\mathcal{L}, ^-, \mathcal{R}, n, \mathcal{K}, C, D)$ be an SSAT, SSAT is closed under transposition and weak transposition, iff the following two conditions hold:*

1. *if $\varphi_1, \ldots, \varphi_n \rightarrow \psi \in \mathcal{R}_s$, then for each $i = 1 \ldots n$, there is $\varphi_1, \ldots, \varphi_{i-1}, -\psi, \varphi_{i+1}, \ldots \varphi_n \rightarrow -\varphi_i \in \mathcal{R}_s$;*
2. *if $\varphi_1, \ldots, \varphi_n \Rightarrow_{sl} \psi \in \mathcal{R}_{sl}$, then for each $i = 1, \ldots, n$, such that $\varphi_i \in C$, there is $\varphi_1, \ldots, \varphi_{i-1}, -\psi, \varphi_{i+1}, \ldots \varphi_n \Rightarrow_{slt} -\varphi_i \in \mathcal{R}_d$. The set of transposed rules is denoted as $\mathcal{R}_{slt} \subseteq \mathcal{R}_d$; the transposed rule of a slippery slope rule $r_i \in \mathcal{R}_{sl}$ ($i = 1 \ldots n$) is denoted as $r_{it} \in \mathcal{R}_{slt}$.*

Weak transposition enables us to achieve the second main line of reasoning of the SSA. According to this definition, the weak transposition can only apply on the sequence of action/events that are linked by slippery slope rules. It is possible that one of the drivers is in fact refutable. However, on the one hand, the attack on drivers can be achieved by means other than weak transposition; on the other hand, we realize that applying transposition to all defeasible rules is dangerous because it can lead to counter-intuitive results. The transposition of slippery slope rules may still cause some disagreement, we will discuss this in Sect. 4.

In $ASPIC^+$, arguments could be attacked in three ways: (1) undermining attack on the ordinary premises; (2) rebutting attack on the conclusions (only when the last rule is defeasible); (3) undercutting attack on the defeasible rules. In this paper we add a special set of premises $\mathcal{K}_0$, whose elements are more like presumptions. So that we define the undermining attack slightly different from in $ASPIC^+$. Besides, since we have defined the weak transposition, the undercutting attack should also become different. Thus the attack relation in SSAT is defined as follows.

Definition 5 (Attack). *Let* A, B *and* X *be arguments in* $SSAT = (\mathcal{L}, ^-, \mathcal{R}, n, \mathcal{K}, C, D)$, $\varphi, \psi \in \mathcal{L}$. *$A$ attacks B (and X), iff A undercuts, rebuts or undermines B, where:*

- *A undercuts B on B' iff:*
  1. *$B' \in Sub(B)$ such that $TopRule(B') = r$ and $r \in \mathcal{R}_d$, $Conc(A) \in \overline{n(r)}$[4];*
  2. *$\exists X$, $X' \in Sub(X)$, $TopRule(X') = r_i (i = 1, \ldots, n)$, $r_i \in \mathcal{R}_{sl}$, and $\exists r_{it} \in \mathcal{R}_{slt}$[5], $Conc(A) \in \overline{n(r_i)}$ (i.e. A undercuts X on X'), while $B' \in Sub(B)$, such that $TopRule(B') = r_{it}$.*
- *A rebuts B on B', iff $Conc(A) \in \overline{\varphi}$ for some $B' \in Sub(B)$ of the form $B''_1, \ldots, B''_n \Rightarrow \varphi$, and if $A = \psi$, then $\psi \notin \mathcal{K}_0$; A contrary-rebuts B iff $Conc(A)$ is a contrary of φ.*
- *A undermines B on B', iff:*
  1. *$B' = \varphi$ and $\varphi \in Prem(B) \cap \mathcal{K}_p$, such that $Conc(A) \in \overline{\varphi}$ and if $A = \psi$, then $\psi \notin \mathcal{K}_0$;*
  2. *$B' = \varphi$ and $\varphi \in Prem(B) \cap \mathcal{K}_0$, such that $Conc(A) \in \overline{\varphi}$.*

  A contrary-undermines B iff $Conc(A)$ is a contrary of φ.

Based on this definition, if an argument undercuts an SSA on one of its slippery slope rules r_i, it will also undercut another argument that contains the defeasible rule r_{it}, which is obtained by applying weak transposition on r_i. Besides, a presumption can only attack another presumption, which means that an argument consisting only of element in $\mathcal{K}_0$ can merely undermine another argument that is also consisted only of element in $\mathcal{K}_0$.

In $ASPIC^+$, whether an attack from A to B (on its sub-argument B') succeeds as a defeat depends on the relative strength of A and B'. In [14], this is determined by a binary ordering $\preceq$ on the set of all arguments. With arguments and the defeat relations, we can evaluate the status of arguments using Dung style abstract argumentation frameworks [8] and decide the set of arguments that jointly acceptable (called an extension) under particular argumentation semantics. Due to limitation of space, we omit the formal introduction of defeat relation in $ASPIC^+$ and the abstract argumentation framework here, the readers are referred to paper [8] and [14] to find more details.

[4] '$n(r)$' means that rule r is applicable.

[5] r_{it} is the transposed rule of r_i.

2.3 Evaluation of SSA

According to the set-up of argumentation schemes, each scheme is corresponded with a specific sequence of critical questions. Basically, there are two ways to evaluate a given argument: (1) use relevant schemes to check the form of the argument; (2) ask the corresponding critical questions, to see if the questions can be answered satisfactorily.

In this section, we try to give some way to evaluate a slippery slope argument based on formal argumentation. The main idea is to formalize the critical questions of the argumentation scheme for SSA, thus we can involve the critical questions into an argumentation framework and evaluate all the arguments together.

Critical Questions. In [22], the author gives 5 critical questions for the basic scheme of SSA, as described below.

- **CQ1** What intervening links in the sequence of events a_1, a_2, ..., a_i needed to drive the slope forward from a_0 to a_n are explicitly stated?
- **CQ2** What missing steps are required as links to fill in the sequence of events from a_0 to a_n, to make the transition forward from a_0 to a_n plausible?
- **CQ3** What are the weakest links in the sequence, where additional evidence needs to be given on whether one event will really lead to another?
- **CQ4** Is the sequence of argumentation meant to be deductive, so that if the first step is taken, it is claimed that the final outcome a_n must necessarily come about?
- **CQ5** Is the final outcome a_n shown to be catastrophic by the value-based reasoning needed to support this claim?

Suppose that a proposed SSA fails to answer CQ1, CQ2 or CQ4 properly, it means that (at least one of) the links from the initial step a_0 to the bad outcome a_n is too weak. In other words, the slippery slope rules between premises to the conclusion is too weak to apply (then we have $\overline{n(r_{sl})}$). And if a proposed SSA fails to answer CQ3 properly, it perhaps that there lacks a driver to back up the 'sloping', or the given driver is not good enough. For the first situation, it could also be seen as that the related link is too weak; for the second situation, it means at least one of the given drivers has been attacked (then we have $\overline{d_i}$). At last, if a proposed slippery slope argument cannot answer CQ5, it means that the final outcome of this argument is not really unacceptable or cause resistance as it has been claimed to (then we have $\overline{\neg\perp}$).

Thus we define the critical questions for slippery slope argument as following.

Definition 6 (Critical Question). *Let argument A, B be arguments in $SSAT = (\mathcal{L}, ^{-}, \mathcal{R}, n, \mathcal{K}, C, D)$, $\varphi, \psi \in \mathcal{L}$. Let A be an SSA, such that $d_i \in Prem(A)$, $r_{sli} \in SlRule(A)$, $Conc(A) = \perp$. B is an argument of critical question for A (denoted by CQA) iff $TopRule(B) = \varphi_1, \ldots, \varphi_n \rightarrow / \Rightarrow \psi$, while $\psi = \overline{n(r_{sli})}$, $\psi = \overline{d_i}$ or $\psi = \overline{\neg\perp}$.*

Here the CQ5 make us aware that the persuasive powers of an SSA should be rested on the fact that the ultimate consequence is catastrophic and really unacceptable to its audiences. Which indicates that the value judgement of the audience may need to be taken into account. Through the case analysis in Sect. 3, the readers should be able to see this point clearer, then we could look back upon this issue and further discuss about it.

3 A Case Analysis

In this section, we apply our argumentation theory for SSA on a slippery slope argument observed in natural language text. The argument came from a Chinese biologist's comments on the Chinese gene editing baby experiment exposed in November 2018.

3.1 The Gene Editing Baby Case: A Practical Application of SSA

On November 26, 2018, Chinese researcher Jiankui He claims that his lab had been editing embryos' genetic codes for seven couples undergoing in-vitro fertilization. Twin girls had been born with DNA altered to make them resistant to HIV, which is the virus that causes AIDS.[6] He used a tool known as CRISPR-cas9 to disable a gene called CCR5, which could form a protein doorway that allows HIV to enter a cell. By doing this, as He claimed, the twin babies are immune to HIV.

Editing the genes of embryos intended for pregnancy is banned in many countries, while in some other countries, editing of embryos may be permitted for research purposes with strict regulatory approval. Jiankui He's experiment is the world's first case of germline gene therapy that performed on humans, which is likely to spark significant ethical questions around gene editing and so-called designer babies. This action shocked and outraged scientists around the world.

Liming Wang, a professor of Zhejiang University who is familiar with genetic technology, released a blog post online to announce his attitude to this event immediately after the news was announced. In which he clearly explained his opinion from several perspectives. In short, there are already many ways to control the genetics of AIDS and reduce the impact of it on patients' lives, therefore the benefit of this action to the newborn children is actually negligible. In turn, the risk of gene editing, including CRISPR-cas9 technique, is still unpredictable and uncontrollable. Furthermore, Wang says, "In addition to the scientific considerations, I have deeper concerns: concerns about the future fate of human beings." In the following text, we can clearly find an application of slippery slope argument. From the following excerpts, we can see more distinctly (translate from Chinese):

[6] The news can be find at the following websites: https://edition.cnn.com/2018/11/26/health/china-crispr-gene-editing-twin-babies-first-intl/index.html, https://www.theguardian.com/science/2018/nov/26/worlds-first-gene-edited-babies-created-in-china-claims-scientist, etc.

Example 2. "... from "treatment" to "prevention" greatly extends the application of gene editing technology. An apparent question is: **where is the boundary** of this technology? You will find it's very difficult to **draw a line**."

"Since editing CCR5 for treating AIDS is reasonable, then isn't it nature to modify CCR5 gene in advance for protection? In this case, is it wrong that an ordinary person also want to protect his children from AIDS? **Take one more step**, if a person has 1% higher risk of getting a genetic disease, isn't it reasonable that he asks for gene editing to reduce the risk? If it is reasonable, can one in ten thousand of the risks be genetically edited? How about one in a million? If it is unreasonable, how much risk can make us allow the gene editing?"

"What more terrible is that once the boundaries of 'treatment' and 'prevention' are broken, it will be much easier to **break the line** between 'prevention' and 'improvement'! What if people want their children to get more muscle, get taller, have blonde hair, double eyelids, or high nose bridges? Even further, what if they want their children to be smarter, have greater abilities on language, analysis and leadership? "

"Though so far, our knowledge about human genes may not achieve these goals, I believe that one day in the future we can figure out all of these things. At that time, will the development of gene technology **bring human beings into the abyss**? Will gene editing destroy the diversity of human gene pool? Will it make human beings monotonous and uncharacteristic? Most seriously, will it cause eternal inequalities? ...... If some people's children get genetically improved, they may have competitive advantages not only in appearance but also in intelligence. What even worse is that these advantages are written into the genome and can be inherited. Thus the other children may never catch up with them!"

The words like "draw a line", "boundary", "one more step", "even further", "break the line", "bring ... to the abyss" appearing in these statements indicate that the SSA is applied.

3.2 Modeling of SSA

The SSA in Example 2 contains arguments from precedent and causal arguments. Apart from some analogies and metaphors in the detail, the author's main idea is as follows:

Firstly, because the boundary of gene editing application is difficult to delimit, if using gene editing to prevent AIDS is approved, then we can hardly stop people to use gene editing on the prevention of other genetic diseases, even if the possibility of getting these diseases are very small but the risks are unpredictable;

Next, since it's much easier to break the line between 'prevention' and 'improvement', then from appearance, physique to intelligence, gradually people will use gene editing techniques to achieve human enhancement.

Then the author gives several negative consequences that may occur. Apparently, he believes that the public will think the most unwanted consequence is "causing eternal inequalities", which is because those people who cannot get genetic improvement, including their offspring, will never be able to catch up with those who have adopted genetic improvement.

In this process, the substantial changing is from *approving the gene-edited HIV-immune babies* (a presumption, the initial step), to *the abuse of gene editing techniques on genetic diseases prevention* (the first step), then to *use gene editing techniques for human enhancement* (the second step), and ultimately lead to *eternal inequalities of human society* (disastrous consequence) and other bad consequences.

The first and second steps can be seen as indications for the beginning and ending of the "gray area" in this SSA respectively. The author gives three reasons to support his statements: (1) *it's very difficult to draw a line*; (2) *it will be much easier to break the line between 'prevention' and 'improvement'*; and (3) *the other children may never catch up with them.*

We use a_0 to denote "approving the gene-edited HIV-immune babies", a_x to denote "abuse of gene editing techniques on genetic diseases prevention", a_y to denote "use gene editing techniques for human enhancement", a_n to denote "eternal inequalities of human society"; [7] d_1, d_2, and d_3 denote the three reasons (drivers) respectively. According to the definition of SSAT in Sect. 2.2, we can get the following argumentation theory.

Example 3 (Example 2 continued). $\mathcal{L} = \{a_0, a_x, a_y, a_n, d_1, d_2, d_3, \bot, \neg\bot\}$;
$\mathcal{K} = \{a_0, d_1, d_2, d_3, \neg\bot\}$; $\mathcal{K}_0 = \{t_0\}$;
$\mathcal{K}_n = \{\}$; $\mathcal{K}_p = \{d_1, d_2, d_3, \neg\bot\}$;
$\mathcal{R}_d = \mathcal{R}_{sl} \cup \mathcal{R}_{slt} = \{a_0, d_1 \Rightarrow_{sl} a_x; a_x, d_2 \Rightarrow_{sl} a_y; a_y, d_3 \Rightarrow_{sl} a_n\}$
$\cup\{\neg a_x, d_1 \Rightarrow_{slt} \neg a_0; \neg a_y, d_2 \Rightarrow_{slt} \neg a_x; \neg a_n, d_3 \Rightarrow_{slt} \neg a_y\}$;
$\mathcal{R}_s = \{a_n \rightarrow_j \bot\} \cup \{\neg\bot \rightarrow \neg a_n\}$.

Arguments are:

$A_1 : a_0$ $A_2 : d_1$ $A_3 : d_2$
$A_4 : d_3$ $A_5 : A_1, A_2 \Rightarrow_{sl} a_x$ $A_6 : A_3, A_5 \Rightarrow a_y$
$A_7 : A_4, A_6 \Rightarrow a_n$ $A_8 : A_7 \rightarrow_j \bot$ $A_9 : \neg\bot$
$A_{10} : A_9 \rightarrow \neg a_n$ $A_{11} : A_4, A_{10} \Rightarrow_{slt} \neg a_y$ $A_{12} : A_3, A_{11} \Rightarrow_{slt} \neg a_x$
$A_{13} : A_2, A_{12} \Rightarrow_{slt} \neg a_0$

According to Definition 5, assuming that all the attack relations we get are success as defeats, we have the following set $\mathcal{D}$ of defeat relations:

$\mathcal{D} = \{(A_5, A_{12}), (A_5, A_{13}), (A_6, A_{11}), (A_6, A_{12}), (A_6, A_{13}), (A_8, A_9), (A_8, A_{10}), (A_8, A_{11}), (A_8, A_{12}), (A_8, A_{13}), (A_{10}, A_7), (A_{10}, A_8), (A_{11}, A_6), (A_{11}, A_7), (A_{11}, A_8), (A_{12}, A_5), (A_{12}, A_6), (A_{12}, A_7), (A_{12}, A_8), (A_{13}, A_1), (A_{13}, A_5), (A_{13}, A_6), (A_{13}, A_7), (A_{13}, A_8)\}$.[8]

[7] We use x, y and n instead of $1, 2$ and 3 because actually between these steps, many intervening small steps are omitted.

[8] Due to the restricted rebutting applied in $ASPIC^+$, A_9 does not directly attack A_8 because the last rule of A_8 is strict. Instead, A_{10} obtained by the transposition of rule '$a_n \rightarrow_j \bot$' rebuts A_8's sub-argument A_7, and thus also attacks A_8.

Now we can get an abstract argumentation framework based on [8] as shown in Fig. 2.

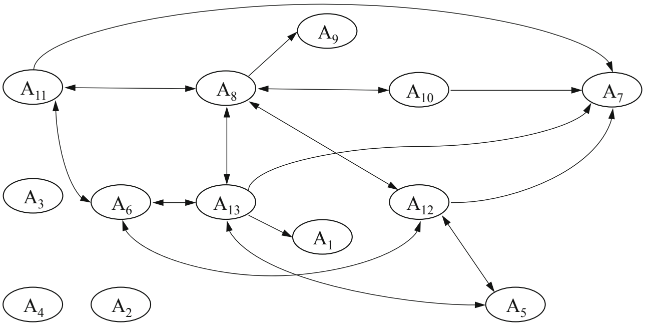

Fig. 2. An abstract argumentation framework

Applying the argumentation semantics [8], we can get four extensions under preferred semantics: $E_{\mathcal{P}1} = \{A_2, A_3, A_4, A_9, A_{10}, A_{11}, A_{12}, A_{13}\}$, $E_{\mathcal{P}2} = \{A_1, A_2, A_3, A_4, A_5, A_9, A_{10}, A_{11}\}$, $E_{\mathcal{P}3} = \{A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8\}$, $E_{\mathcal{P}4} = \{A_1, A_2, A_3, A_4, A_5, A_6, A_9, A_{10}\}$. Compared with extensions under other semantics, preferred extensions can reflect a more credulous attitude. If the arguments has equal priorities and there is no additional information, a credulous agent may accept one of the above extensions.[9] Besides, we can get one extension under grounded semantics: $E_{\mathcal{G}1} = \{A_2, A_3, A_4\}$. The grounded extension reflects the most skeptical attitude of agents. In the argumentation framework of Fig. 2, only argument A_2, A_3 and A_4 (whose conclusions are d_1, d_2 and d_3 respectively) are not attacked, thus a very skeptical agent will only accept these three arguments. There are other argumentation semantics introduced in [8], [2], etc. Here we only take two of them for instance.

In Example 2, the blogger mentions that the key reason he disagree with the CCR5 gene-edited babies experiment is that the benefit it will bring is far less than the risk. Apparently, in addition to the unpredictable accidents such as "off-target effects" during operations, the catastrophic consequence (i.e. *cause eternal inequalities in human society*) mentioned in his SSA is also one of the risks - perhaps the worst one. This kind of statements reflects his value judgement:

[9] The proponent of an SSA will expect the audience to accept $E_{\mathcal{P}1}$. However, the persuasiveness of an SSA depends on audiences, and many factors will affect their final decision. For example, whether the audience is worried enough about the catastrophic consequence. Since we has assumed that all the attacks are successful (while the proponent won't consider that a_8 will defeats a_9), from the perspective of the audience, we believe that the current result is in line with human intuition, i.e. some audiences are successfully persuaded by the SSA (thus accepting $\neg a_0$), whereas some audiences are not.

The value of avoiding the catastrophic consequence is much higher than the value of enjoying the benefit of CCR5 gene editing. And he believes that the public will agree with this opinion.

In fact, many other experts also expressed their opposition to this experiment in the mass media.[10] In popular social media platforms in China, such as Weibo, people almost unanimously criticized the experiment of He's team. Liming Wang's blog post has also been widely reposted by users of various social media platforms. These phenomena reveal that Wang's point of view and value judgment are generally approved, and his arguments are convincing to the public.

In argumentation theory, the statement "the babies are immune to HIV, which is good; good thing should not be resisted and we should approve the gene-edited HIV-immune babies" can be modeled as:

Example 4 (Example 3 continued). ('imH' and 'G' denote 'immune to HIV' and 'good' respectively, $\overline{\neg\bot}$ denotes 'not be resisted'[11], here we add them into $\mathcal{L}$. Three more rules are obtained: $a_0 \Rightarrow imH$, $imH \Rightarrow G$ and $G \rightarrow \overline{\neg\bot}$. We add them into $\mathcal{R}_d$ and $\mathcal{R}_s$ respectively.)

$A_{14} : A_1 \Rightarrow imH$ $A_{15} : A_{14} \Rightarrow G$ $A_{16} : A_{15} \Rightarrow \overline{\neg\bot}$
$A_{17} : A_{15} \Rightarrow a_0$

According to Definition 5, A_{16} conflicts with argument A_9 and arguments A_{10}, A_{11}, A_{12}, A_{13} (because A_9 is their sub-argument), A_{17} conflicts with argument A_{13}. Suppose that based on argument $A_1 - A_{17}$, there is an audience who prefer a_0 than $\neg a_0$, $\overline{\neg\bot}$ than $\neg\bot$, then the only preferred extension and grounded extension of the updated argumentation framework will be $\{A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_{14}, A_{15}, A_{16}, A_{17}\}$. So that the initial action a_0 is acceptable to this particular audience. On the contrary, if we obtain an ordering on arguments according to the value judgement of most people in this case, it's more likely that A_{13} will has higher priority than A_{17}, A_9 will has higher priority than A_{16}, thus both the attack from A_{16} to A_{13} and from A_{17} to A_{9-13} will not be successful.

4 Discussion and Related Work

In this section we discuss some basic ideas and important issues in this paper, and introduce some related works.

4.1 Discussion About the Weak Transposition

Firstly, in addition to defining the transposition of strict rules, we also give a definition of 'weak transposition' of defeasible rules in Sect. 2.2. The reason lies in the operating mechanism of the slippery slope argument: unacceptable outcomes

[10] Refer to the news https://edition.cnn.com/2018/11/26/health/china-crispr-gene-editing-twin-babies-first-intl/index.html and [24,25].

[11] Remember that $\neg\bot$ in $\mathcal{K}$ represents "resistance to something bad/unwanted".

indicate that its premise is unacceptable. Though going through a long chain, it still implies a backward reasoning. What's more, without the weak transposition, in order to come up with a final conclusion (which is "not to take the first step"), the slippery slope argument is self-attacking based on the formal argumentation theory.

However, although the application of this transposition may raise criticism and controversy, the current paper is neither the unique nor the first to propose the contraposition/transposition of defeasible rules. In [7], Caminada examined Socrates's elenchus, which always leads the audience to make an inference that discredits his own reasoning (thus called "hang yourself argument"), and put forward the issue of contraposition and defeasible reasoning. Then in [6], he distinguished between epistemical reasoning and constitutive reasoning, and concluded that whether there should be contraposition of defeasible rules depends on which type of reasoning one is considering. In many aspects, the slippery slope argument is comparable to the "hang yourself argument", thus analysis in [6,7] are considerable references.

4.2 Discussion About Value Judgement

Through the case analysis in Sect. 3, it is not difficult to find that: if there is a counter-argument which asserts that the benefits may outweigh the harm claimed by an SSA, whether the attacks will succeed depends on the value judgement of the audience.

As Walton mentioned in [22], SSA is a subspecies of argument from negative consequence, and could also be seen as an approach to achieve practical reasoning. A slippery slope argument works by claiming that take the first step will lead to a highly undesirable consequence, which means that the consequence strongly contravenes values held by the audience [22].

In the current work, we model the negative value by adding a symbol "$\perp$" into the language $\mathcal{L}$ of an argumentation system. Correspondingly, we add a symbol "$\neg\perp$" into the knowledge base $\mathcal{K}$ to represent the intrinsic unacceptability of something bad. Then a slippery slope argument can be attacked by statements like "the final outcome is not as bad as it has been claimed", i.e., based on Definition 6, a CQA with the conclusion $\overline{\neg\perp}$. In $ASPIC^+$, conflicts can be resolved by comparing arguments based on preference, thus when we consider the preference in an SSAT, value judgement deserved to be taken into account.

Several systems that consider values based on formal argumentation have already been proposed. In [12], Liao and Oren et al. introduced a hierarchical abstract theory of normative system (called HANS) to resolve conflicts amongst norms. In simpler terms, this system associated numbers that indicating priorities of norms to an abstract theory of normative system defined by Tosatto et al. [17]. When conflicts arise between norms, HANS resolve it by the priorities assigned to them, and derive extensions according to different detachment procedure. In [13], Liao and Slavkovik et al. consider moral values and present an approach based on formal argumentation and normative systems to reach moral agreements. In [3], Bench-Capon clarified the role of persuasion in practical argumentation, and extends the abstract argumentation frameworks to a value-based argumentation framework (VAFs). In [4], Bench-Capon and Atkinson et al. focus on legal reasoning and discusses how to instantiate a VAF.

5 Summary

On the basis of the basic scheme of an SSA given in [21] and the formal argumentation framework $ASPIC^+$ [14,15], the present paper gives an argumentation theory for SSA (called SSAT). In addition, we give a definition of critical questions. Accordingly, based on the SSAT, we can model the basic form of SSA and evaluate it by formal argumentation system.

We apply this argumentation theory to model an SSA found in a popular blog post, which criticized the gene-edited babies experiment. The blog post has got a lot of attention since the news released in last November. The blogger, a Chinese biologist, used an SSA to back up the opinion that the benefits of adopting such an experiment are far less than the risk. By argumentation evaluation based on an abstract argumentation framework, we get extensions of arguments (and thus we can get the corresponding extensions of conclusions). It shows that our SSAT is able to model SSAs found in natural language text, and get reasonable results.

Furthermore, we point out that value judgement plays an important role in the evaluation of effectiveness of an SSA. How to lift preference on arguments through the value assignment or ranking in SSA, is a topic for future studies.

References

1. Atkinson, K., Bench-Capon, T.: Abstract argumentation scheme frameworks. In: Dochev, D., Pistore, M., Traverso, P. (eds.) AIMSA 2008. LNCS (LNAI), vol. 5253, pp. 220–234. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85776-1_19
2. Baroni, P., Caminada, M., Giacomin, M.: An introduction to argumentation semantics. Knowl. Eng. Rev. **26**(4), 365–410 (2011)
3. Bench-Capon, T.: Persuasion in practical argument using value-based argumentation frameworks. J. Logic Comput. **13**(3), 429–448 (2003)
4. Bench-Capon, T., Atkinson, K., Chorley, A.: Persuasion and value in legal argument. J. Logic Comput. **15**(6), 1075–1097 (2005)
5. Cabrio, E., Hirst, G., Villata, S., Wyner, A.: Natural language argumentation: mining, processing, and reasoning over textual arguments (Dagstuhl seminar 16161). Technical report, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany (2016)
6. Caminada, M.: On the issue of contraposition of defeasible rules. In: Computational Models of Argument: Proceedings of COMMA 2008, Toulouse, France, 28–30 May 2008, vol. 172, pp. 109–115, January 2008
7. Caminada., M.: For the sake of the argument; explorations into argument-based reasoning. Ph.D. thesis, Vrije University Amsterdam, Amsterdam (2004)

8. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artif. Intell. **77**(2), 321–357 (1995)
9. Gabbay, D.M., Thiruvasagam, P.K.: Reasoning schemes, expert opinion and critical questions. Sex offenders case study. In: FLAP, vol. 4 (2017)
10. Govier, T.: What's wrong with slippery slope arguments? Can. J. Philos. **12**(2), 303–316 (1982)
11. Johnson, R.H., Blair, J.A.: Logical Self-defense, 2nd edn. McGraw-Hill Ryerson, Toronto (1983)
12. Liao, B., Oren, N., van der Torre, L., Villata, S.: Prioritized norms and defaults in formal argumentation. In: Proceedings of the 13th International Conference on Deontic Logic and Normative Systems (DEON 2016), pp. 139–154 (2016)
13. Liao, B., Slavkovik, M., van der Torre, L.W.N.: Building Jiminy cricket: an architecture for moral agreements among stakeholders. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, 27–28 January 2019, pp. 147–153 (2019)
14. Modgil, S., Prakken, H.: A general account of argumentation with preferences. Artif. Intell. **195**, 361–397 (2013)
15. Prakken, H.: An abstract framework for argumentation with structured arguments. Argum. Comput. **1**(2), 93–124 (2010)
16. Prakken, H., Wyner, A., Bench-Capon, T., Atkinson, K.: A formalization of argumentation schemes for legal case-based reasoning in ASPIC+. J. Logic Comput. **25**(5), 1141–1166 (2015)
17. Tosatto, S.C., Boella, G., van der Torre, L., Villata, S.: Abstract normative systems: semantics and proof theory. In: Brewka, G., Eiter, T., McIlraith, S.A. (eds.) Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning, KR 2012, Rome, Italy, pp. 358–368 (2012)
18. Verheij, B.: The toulmin argument model in artificial intelligence. In: Simari, G., Rahwan, I. (eds.) Argumentation in Artificial Intelligence, pp. 219–238. Springer, Boston (2009). https://doi.org/10.1007/978-0-387-98197-0_11
19. Waller, B.N.: Critical Thinking: Consider the Verdict, 6th edn. Prentice Hall, Upper Saddle River (1998)
20. Walton, D.: Slippery Slope Arguments. Oxford UniversityPress, Oxford (1992)
21. Walton, D.: The basic slippery slope argument. Informal Logic **35**(3), 273–311 (2015)
22. Walton, D.: The slippery slope argument in the ethical debate on genetic engineering of humans. Sci. Eng. Ethics **23**(6), 1507–1528 (2017)
23. Walton, D., Reed, C., Macagno, F.: Argumentation Schemes. Cambridge University Press, Cambridge (2008)
24. Wang, C., Zhai, X., Zhang, X., Li, L., Wang, J., Liu, D.P.: Gene-edited babies: Chinese academy of medical sciences' response and action. Lancet **393**(10166), 25–26 (2019)
25. Zhang, L., Zhong, P., Zhai, X., Shao, Y., Lu, S.: Open letter from Chinese HIV professionals on human genome editing. Lancet **393**(10166), 26–27 (2019)