Persuasion Technology Through Mechanical Sophistry

Micah Clark and **Selmer Bringsjord**[1]

Abstract. We introduce an approach to digital persuasive technology. In this approach, computing machines are given the power to persuade by generating what we call *illusory arguments*. These arguments derive their power in significant part by exploiting the empirical fact that humans succumb to *cognitive illusions*. The persuasion technology we present can be used for weal or woe, but at least for now, we approach the subject purely as AI engineers, not ethicists.

1 INTRODUCTION

AI has developed systems involving argumentation in two ways: modeling *with* argument, and modeling *of* argument [26]. Our work is in neither of these categories, but it is inspired by the distinction. We aim to engineer a machine that generates sophistic arguments in order to persuade. Such a machine could be used to persuade humans of truth or falsehood, and our emphasis is on the latter: persuading humans of falsehoods. As an engineering endeavor, constructing a machine that *lies* via sophistic argument makes sense, because, if the engineering succeeds for falsehoods, it can be adapted to persuade humans of truth.

We introduce a psychology of reasoning-based approach to mechanical sophistry, namely an approach for generating *illusory arguments* — arguments that compel erroneous belief by exploiting the dichotomy of what humans *ought* to believe and what humans *will* likely believe. In illusory arguments, the gap between *ought* and *will* is significant and predictable; these arguments incorporate, and prey upon, reliable systematic errors in human reasoning. In short, illusory arguments are psychologically persuasive yet fallacious arguments that lead humans into erroneous beliefs while simultaneously confirming their naïve intuitions about what is true.

In the mechanical sophistry we propose, illusory arguments emerge from an ecumenical theory of human reasoning. This paper roughly circumscribes our approach to sophistry; an approach novel in its application of empirically supported psychological theories to the practical challenge of persuading humans through argument.

2 SOPHISTIC PERSUASION

Some have viewed persuasion in argument as what force or influence a justification should have (surveyed, e.g., in [5, 23]), using as their lens a continuum of strength factors (e.g., Chisholm's [6]) and frameworks for defeasible reasoning in the tradition of Pollock [21] and Toulmin [31]. Others have treated persuasion in argument as a way to resolve a difference of opinion, or to establish a common belief or intent [32, 17], studying persuasion's game-theoretic properties in multi-agent dialog systems (surveyed in [22]). There seem to be few studying persuasion (e.g., [10, 27]) who implement machines capable of persuading humans.

Our work appeals to an older, decidedly philosophical conception of persuasion: sophistry. The essence of sophistry *is* persuasion, the intentional manipulation of others' beliefs through the force of argument.[2] The skilled practitioners of sophistry are those whose rhetorical methods successfully manipulate.

There are two broad sorts of persuasive arguments, "the one producing belief without knowledge, the other knowledge" [19]. Likewise, there are two archetypal practitioners: the *liar* and the *truth-teller*. The truth-teller argues in good faith for what he/she believes with the intention of producing knowledge. The liar argues falsely, and sophistically, for what he/she does not believe with the intention of producing belief without knowledge.[3] Of the two, the successful liar demonstrates greater skill than the successful truth-teller: the truth-teller has in persuasion an ally, truth, while the liar must overcome, and conquer, truth as an adversary. It is this inequity of skill that calls us to study arguments that deceive rather than arguments that inform, and liars rather than truth-tellers.

3 COGNITIVE ILLUSIONS

Our theory of sophistry is driven by the difference between what humans *will* likely believe in light of arguments, and what ideally rational agents *ought* to believe in light of said arguments. According to the neo-Piagetian view of the development of bias-free human reasoning: Given sufficient training, neurobiologically normal humans can reason in a normatively correct, bias-free manner [3, 28]. A mounting body of empirical evidence of *cognitive illusions* (see, e.g., [18, 20, 1]) strongly suggests the inverse of the neo-Piagetian view is also true, that being: Naïve humans (i.e., those untrained or insufficiently trained in logic) often reason in an incorrect, bias-plagued manner, and they do so in regular, systematic fashion.

Cognitive illusions designate phenomena where the beliefs and behavior of humans are consistent and predictable, yet deviate substantially from what is ideally rational. Worse still, when succumbing to cognitive illusions humans believe, and report, that they are thinking and acting rationally. The allusion here to "illusions" is not accidental. It reflects the position that formal, normative rationality has primacy over base human intuition.[4] Even if one rejects formal rationality as the appropriate benchmark for human rationality, the fact remains that deviation from the maxims of formal rationality leads inexorably to the "Dutch book" [25, 8], i.e., to situations where one will believe falsehoods and act against one's material self-interest. Cognitive illusions are, in essence, empirically validated exemplars of human vulnerability to "Dutch book" situations.

[1] Rensselaer Polytechnic Institute, USA, email: {clarkm5, brings}@rpi.edu

[2] "[Sophistry] is the creator of persuasion, and that all its activity is concerned with this, and this is its sum and substance." [19]

[3] There are many variants of the two archetypes; e.g., one can speak a falsehood believing it to be true, and a fallacious argument can be given in support of a true conclusion.

[4] This position is dominate but not universal (see, e.g., [4], for discussion).

4 ILLUSORY INFERENCES

Illusory inference is a particular class of cognitive illusion wherein naïve humans, when presented with a collection Φ of declarative sentences, overwhelmingly conclude ψ, though ψ does not follow from Φ in any standard proof calculus c (i.e., $\Phi \nvdash_c \psi$). One example is:

> "All the Frenchmen in the room are wine-drinkers. Some of the wine-drinkers in the room are gourmets." [11]

When asked what, if anything, they can rightly conclude from the above sentences, the majority of human subjects erroneously infer that some of the Frenchmen in the room are gourmets, while in fact, no relation between Frenchmen and gourmets can be validly inferred. Another example (viz., the *king–ace* illusion) is:

> "If there is a king in the hand, then there is an ace, or if there isn't a king in the hand, then there is an ace, but not both of these if-thens are true." [11]

When asked about this sentence, almost every subject draws the conclusion that there is an ace in the hand, even though it follows deductively that there *cannot* be an ace in the hand. To see why there cannot be an ace, recognize that one of the two conditionals in the premise must be false, and a conditional is false only when its antecedent is true and its consequent is false. If "if there is a king in the hand, then there is an ace" is false, then there is not an ace, and if "if there isn't a king in the hand, then there is an ace" is false, then again, there is not an ace. So regardless of which conditional is false, there cannot be an ace in the hand. If the subjects had been offered wagers on the correctness of their answers, they would have been summarily fleeced.

How humans make inferences and draw conclusions (correct or otherwise) is, in psychology, a matter of controversy. Some hold that humans reason by application of inference schemata akin to a proof theory, i.e., a sort of "mental logic" [29, 2]. But explaining illusory inferences has challenged the proponents of mental logics.[5] In contrast, illusory inferences are consonant with the view that humans reason by constructing and inspecting iconic representations homomorphic to real or abstract situations, so called "mental models" [12, 30]. According to mental model theories, illusory inferences are the result of incomplete mental models; in response to limited cognitive resources, humans normally construct only partial mental representations. These partial representations are usually sufficient for correct reasoning but, under certain conditions, they lead to compelling yet erroneous conclusions, i.e., illusory inferences. Mental model theories have successfully predicted illusions in, e.g., modal, deontic, spacial, causal, and probabilistic reasoning domains.

Human reasoning is heterogeneous. Our ecumenical theory subsumes elements and ideas from mental logics and mental model theories. The bridging of mental logics and mental models is an aspiration our theory shares with *mental meta-logic* [33, 34]; however, their pursuit and ours are quite different. Mental meta-logic pursues a predictive, statistical theory of perceived difficulty and objective accuracy in reasoning tasks, while we pursue predictive simulation of human reasoning through a unified inferential calculus. Toward this end, we developed an inferential calculus modeling human sentential reasoning. By then varying the definitions of the calculus' basic operators we formed a lattice of calculi, each modeling a degree of reasoning competency distinguished according to completeness of mental representation, and ability to compensate for remaining incompleteness. The calculi share a decision procedure for assessing consistency, validity, and modal consequence (i.e., possible consequence, necessary consequence); the actual assessments (e.g., what sentences are validities), of course, differ between calculi.

Space constraints do not permit us to detail the calculi and decision procedure here, but their import is this: Through a short series of diagnostic questions we can identify the specific calculus that best approximates an individual subject's reasoning ability, and thus make subject-specific predictions of *perceived* consistency, validity and (modal) consequence in sentential reasoning.

5 ILLUSORY ARGUMENTS

Plato held that sophistry could only sway the uneducated and uninformed.[6] But cognitive illusions, and illusory inferences in particular, suggest that Plato was wrong. There might be a sophistic technique persuasive to domain experts and the ignorant alike; a technique based on illusory inferences to which few are immune, viz., *illusory arguments*.

To explain, theories of human reasoning (including our own) are generally extrapolations from the results of empirical studies wherein subjects are asked to perform some constructive reasoning task.[7] These tasks are "constructive" in the sense that subjects are tacitly expected to originate a line of reasoning, or justification, for their answers. Illusory inferences are also studied in this fashion. What has been largely ignored are evaluative tasks, where subjects are asked to agree or disagree with a determinate line of reasoning, i.e., to accept or reject an argument given to them.

The hypothesis of *illusory arguments* is this: If humans regularly err in concluding that ψ follows from a set Φ of declarative sentences due to their own imperfect reasoning, they should all the more affirm ψ when presented with an argument mirroring what would likely be their own flawed justification for ψ. For example, many may be persuaded of the presence of an ace in the hand (in the *king–ace* illusion) by the following illusory argument:

> *We know that either there is a king in the hand or there isn't. Suppose that there is a king. Now, we also know that if there is, then there is an ace. We know this from the conditional that "if there is a king in the hand, then there is an ace." Therefore, there is an ace in the hand.*
>
> *Now, let's suppose that there* isn't *a king in the hand. This connects to the other if-then: "if there isn't a king in the hand, then there is an ace." Therefore, we can infer again that there is an ace in the hand.*
>
> *So, while we don't know whether there's a king in the hand or not, it doesn't matter: There is an ace in the hand.*

We represent an argument $\mathcal{A}$ as a finite sequence of inferences much as one might articulate a natural deduction proof. The semantic

---

[5] The inference schemes and procedures associated with mental logics are typically valid yet incomplete — all drawn conclusions are valid but not all valid conclusions are reachable. Yet in illusory inferences, the conclusion drawn is invalid. Proponents of mental logics explain illusory inferences as inadvertent deviations from the prescribed process, e.g., mis-applying, or failing to apply, an inference schema. This style of explanation has largely precluded mental logics from predicting illusory inferences.

[6] "[Sophistic rhetoricians] merely discover a technique of persuasion, so as to appear among the ignorant to have more knowledge than the expert [. . .] for surely, among those who know, he will not be more convincing than the [expert]." [19]

[7] E.g., to decide what can be concluded from premises, to decide whether one premise is possible given the truth of another, to describe what would make a premise false, etc.

structure of an argument is "natural" insofar as it includes the basic format and structures of Fitch-style [9] natural deduction extended to informal arguments. In a proof, each sentence s_i must follow from one or more of the preceding $i-1$ sentences in accordance with the specific proof theory at hand. The proof theory also determines the "size" of an inferential leap. In illusory arguments, the ith sentence need only appear to follow from what precedes it, and the size of an inferential leap may be arbitrary so long as the veneer of logical consequence is maintained. With regard to the inferential calculi mentioned in §4, an inference is *perceived* as valid if its conclusion is perceived to be a necessary consequence of its premises, and an argument $\mathcal{A}$ is *perceived* as valid if all of its inferences are perceived as valid. Thus, the calculi's decision procedure for (modal) consequence can be extended to decide the validity of inferences and arguments.

In illusory arguments and sophistry, veracity is not our concern, instead credibility (i.e., believability, persuasive force) is. We use as the measure of an argument's credibility, its probability of acceptance by a naïve human or humans. Based on prior published results (and our own ongoing experiments) we are developing a parametric model of credibility for our theory of human sentential reasoning. In its current embryonic form, our model equates the frequency at which particular inferences are made in constructive tasks to the probability that those inferences will be accepted in evaluative tasks. For example, the probability of acceptance (i.e., credibility) for *affirming the consequent* is assumed to be the observed frequency at which subjects commit this fallacy in constructive tasks. The credibility of an argument $\mathcal{A}$ is taken to be the probability that all necessary inferences, leading from initial premises to final conclusion, are accepted.

This model of credibility is quite impoverished; we do not expect it to predict actual argument acceptance rates. Fortunately, in sophistry objective accuracy is not required. All that is needed is a relative ordering of arguments, i.e., a determination of whether one argument is more, or less, credible than another — though obviously, the more accurate one's prediction of credibility, the better one's discrimination between arguments. We are working toward a more sophisticated model of credibility, one that makes reasonably accurate predictions of argument acceptance rates. However, there are significant difficulties in extracting such a model from extant studies of performance in constructive tasks; we briefly mention two of them.

1. We are interested in making single-subject predictions of argument credibility, yet most of the relevant studies report only aggregate subject data.
2. Core competence may subvene the reasoning in both constructive and evaluative tasks, but it seems unlikely that the strategies used to direct reason would be common across these tasks. For example, subjects on their own rarely infer a disjunctive statement from a categorical assertion (i.e., disjunction-introduction), but when evaluating the reasoning of another, it is doubtful that they would so rarely recognize the validity of this inference.

These two difficulties (as well as others) point out that further study of human reasoning in evaluative tasks is needed if we are to achieve the goal of a truly predictive model of credibility.

6 MECHANICAL SOPHISTRY

Our approach to sophistry is made concrete in the machine $\mathcal{M}$. $\mathcal{M}$ is a normatively correct reasoner, i.e., $\mathcal{M}$'s reasoning is valid; it employs a standard inferencing scheme for many-sorted logic [15] (we use $\vdash_{MSL}$). $\mathcal{M}$ reasons over its own beliefs, and its beliefs about the beliefs of $\mathcal{D}$, whom $\mathcal{M}$ intends to deceive or persuade. Using capital Greek letters Γ and Σ to indicate sets of declarative sentences, we say $\mathcal{M}$ believes Γ, and $\mathcal{M}$ believes that $\mathcal{D}$ believes Σ. Further, $\mathcal{M}$ believes that $\mathcal{D}$'s reasoning capacity and process accords with our ecumenical theory of human reasoning, i.e., $\mathcal{M}$ has a *theory of mind* (ToM) [24] relative to $\mathcal{D}$ (we use $\vdash_{ToM}$). $\mathcal{M}$ does not assume that $\mathcal{D}$ is logically omniscient with respect to the ToM. Rather, the ToM describes what $\mathcal{D}$ is likely to believe upon sufficient reflection; that is to say, the ToM describes $\mathcal{D}$'s *implicit beliefs* (see, e.g., [14, 13]).

In this framework, $\mathcal{M}$ can deceive $\mathcal{D}$ about the truth of a sentence s when $\mathcal{M}$ believes that s is not necessarily true (i.e., $\Gamma \nvdash_{MSL} s$) or that s is false (i.e., $\Gamma \vdash_{MSL} \neg s$), and yet $\mathcal{M}$ believes $\mathcal{D}$ to be predisposed to believe that s is true (i.e., $\Sigma \vdash_{ToM} s$). To deceive $\mathcal{D}$ about the truth of s, $\mathcal{M}$ constructs an argument concluding s, and does so in a way that maximizes the argument's credibility (i.e., believability, persuasive force). $\mathcal{M}$'s process of argument construction is similar to that of proof-search in automated reasoning. $\mathcal{M}$ consummates the intended deception by articulating to $\mathcal{D}$ the most credible argument found for the truth of s. (Note that the most credible argument for s may well be normatively valid; this can occur when $\mathcal{M}$ believes that $\mathcal{D}$ already holds an erroneous belief, and s saliently follows, in part, from the erroneous belief in a perceptually, and logically, valid way. When $\mathcal{M}$ believes that $\mathcal{D}$'s beliefs are correct, $\mathcal{M}$'s argument for the truth of s is necessarily illusory.)

7 GENERATING ILLUSORY ARGUMENTS

There is no shortage of techniques available for generating illusory arguments. Here we quickly encapsulate three algorithms for doing so. For the first algorithm consider the following valid rule (R) of deductive inference:

$$\frac{(\phi \rightarrow \psi) \wedge (\neg\phi \rightarrow \psi)}{\psi}$$

Generally speaking, the validity of R can be grasped by a college-educated subject S. In light of this, an algorithm-sketch for generating a corresponding sophistic argument, assuming R (and the like) as input, is to first modify this rule to produce a variant R' that inherits the air of plausibility of the original, but which is nonetheless invalid; and to then generate, in natural language, an argument that conforms to R'. This is how the illusory English argument we presented above (§5) "showing" that there is an ace in the hand can be automatically generated. R' in the case of this "demonstration" is

$$\frac{(\phi \rightarrow \psi) \vee (\neg\phi \rightarrow \psi)}{\psi}$$

which is an invalid rule of inference.

For the second algorithm we use a set of inference rules (including some formal fallacies), and a rule-application procedure for enumerating sequences of inferences ordered by length. (Note that some applications of these rules may be perceived as invalid.) To produce an argument apparently demonstrating that ψ follows from Φ, we iterate over inferential sequences that conclude ψ from Φ, testing each for perceived validity. For those sequences perceived as valid, our measure of credibility is computed, and the most credible sequence is retained. The algorithm terminates when a threshold for credibility is surpassed, or when the time alloted for search is exhausted.

The third algorithm recasts argument construction as a graph-theoretic path-finding problem. Nodes in the graph represent states of a knowledge-base, edges characterize perceptually valid inferences, and paths represent perceptually valid inferential sequences. By suitably defining path-length in terms of our valuation of credibility, constructing a most-credible argument that ψ follows from Φ reduces to finding a shortest path from the state where the knowledge-base contains only Φ to any state where the knowledge-base contains ψ.

8 PURELY PERSUASION

Our emphasis so far has been on persuading humans of falsehoods, i.e., the production of belief without knowledge via illusory arguments. While our approach to mechanical sophistry facilitates this type of deception, it is not limited to it. Our approach can equally persuade humans of truth, i.e., produce belief that is knowledge. This is so because the difference between a corruptive, illusory argument and an intuitive (i.e., persuasive) explanation of truth is simply whether (i) the argument's premises are true, and (ii) the reasoning articulated is logically valid. When our machine $\mathcal{M}$ is charged with arguing for a false conclusion, it attempts the corruption of another's mind (e.g., to instill erroneous belief).[8] But when $\mathcal{M}$ is charged with arguing for a true conclusion using only logically valid reasoning, it attempts the education or restitution of another's mind (e.g., to instill knowledge, to disabuse erroneous belief).[9] That is to say, by maximizing the credibility of valid reasoning, $\mathcal{M}$ tries to make truth comprehensible — to persuade humans of truths that they would otherwise likely not apprehend.

We do not aspire to a new philosophic theory of rhetoric. Instead, we are answering the charge that theoreticians and compositionists in Western rhetoric "have effectively halted the development of non-philosophic, sophistic theories of rhetoric and the realization of their practical possibilities" [16]. Our aim, ultimately, is an empirically-grounded, falsifiable theory of the relative persuasiveness of arguments to humans, and computational methods for generating persuasive arguments. The two combined are useful in the bending persuasion toward whatever end one desires.

ACKNOWLEDGEMENTS

We are indebted to Joshua Taylor, Michel Ingham, and three anonymous referees for insightful criticisms and comments (responses to which are woven into the present version of the paper).

REFERENCES

[1] D. Ariely, *Predictably Irrational*, Harper Collins, New York, NY, 2008.
[2] *Mental Logic*, eds., M. D. S. Braine and D. P. O'Brien, Lawrence Erlbaum, Mahwah, NJ, 1998.
[3] S. Bringsjord, E. Bringsjord, and R. Noel, 'In Defense of Logical Minds', in *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, eds., M. A. Gernsbacher and S. J. Derry, pp. 173–178, Mahwah, NJ, (1998). Lawrence Erlbaum.
[4] N. Chater and M. Oaksford, 'Rationality, rational analysis, and human reasoning', in *Psychology of Reasoning*, eds., K. Manktelow and M. C. Chung, 43–74, Psychology Press, New York, NY, (2004).
[5] C. I. Chesñevar, A. G. Maguitman, and R. P. Loui, 'Logical Models of Argument', *ACM Computing Surveys*, **32**(4), 337–383, (2000).
[6] R. M. Chisholm, *Theory of Knowledge*, Prentice Hall, Englewood Cliffs, NJ, 1966.
[7] R. M. Chisholm and T. D. Feehan, 'The Intent to Deceive', *Journal of Philosophy*, **74**(3), 143–159, (1977).
[8] B. de Finetti, 'La prévision: Ses lois logiques, ses sources subjectives', *Annales de l'Institute Henri Poincaré*, **7**, 1–68, (1937).
[9] F. Fitch, *Symbolic Logic*, Ronald Press, New York, NY, 1952.
[10] F. Grasso, A. Cawsey, and R. Jones, 'Dialectical argumentation to solve conflicts in advise giving: a case study in the promotion of healthy nutrition', *International Journal of Human-Computer Studies*, **53**(6), 1077–1115, (2000).
[11] P. N. Johnson-Laird, 'Rules and Illusions: A Critical Study of Rips's *The Psychology of Proof*', *Minds and Machines*, **7**(3), 387–407, (1997).
[12] P. N. Johnson-Laird, *How We Reason*, Oxford University Press, New York, NY, 2006.
[13] G. Lakemeyer, 'Steps Towards a First-Order Logic of Explicit and Implicit Belief', in *Proceedings of the 1986 Conference on Theoretical Aspects of Reasoning about Knowledge*, ed., J. Y. Halpern, pp. 325–340, San Francisco, CA, (1986). Morgan Kaufmann.
[14] H. J. Levesque, 'A Logic of Implicit and Explicit Belief', in *Proceedings of the Fourth National Conference on Artificial Intelligence*, pp. 198–202, Menlo Park, CA, (1984).
[15] M. Manzano, *Extensions of First Order Logic*, Cambridge University Press, Cambridge, UK, 1996.
[16] R. Marback, *Plato's Dream of Sophistry*, University of South Carolina Press, Columbia, SC, 1999.
[17] P. McBurney, S. Parsons, and M. Wooldridge, 'Desiderata for Agent Argumentation Protocols', in *Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pp. 402–409, New York, NY, (2002). ACM Press.
[18] M. Piattelli-Palmarini, *Inevitable Illusions*, John Wiley & Sons, 1994.
[19] Plato, 'Gorgias', in *The Collected Dialogues of Plato*, eds., E. Hamilton and H. Cairns, 229–307, Princeton University Press, Princeton, NJ, (1971/380 BC). tr., W. D. Woodhead.
[20] *Cognitive Illusions*, ed., R. Pohl, Psychology Press, New York, NY, 2004.
[21] J. L. Pollock, *Cognitive Carpentry*, MIT Press, Cambridge, MA, 1995.
[22] H. Prakken, 'Formal systems for persuasion dialogue', *The Knowledge Engineering Review*, **21**(2), 163–188, (2006).
[23] H. Prakken and G. A. W. Vreeswijk, 'Logics for Defeasible Argumentation', in *Handbook of Philosophical Logic*, eds., D. M. Gabbay and F. Guenthner, volume 2, 219–318, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2nd edn., (2002).
[24] D. Premack and G. Woodruff, 'Does the chimpanzee have a theory of mind?', *Behavioral and Brain Sciences*, **1**(4), 515–526, (1978).
[25] F. P. Ramsey, *The foundations of mathematics and other logical essays*, Routledge & Kegan Paul, London, UK, 1931.
[26] C. Reed and F. Grasso, 'Recent Advances in Computational Models of Natural Argument', *International Journal of Intelligent Systems*, **22**(1), 1–15, (2007).
[27] E. Reiter, R. Robertson, and L. M. Osman, 'Lessons from a failure: Generating tailored smoking cessation letters', *Artificial Intelligence*, **144**(1–2), 41–58, (2003).
[28] K. J. Rinella, S. Bringsjord, and Y. Yang, 'Efficacious Logic Instruction: People are not Irremediably Poor Deductive Reasoners', in *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, eds., J. D. Moore and K. Stenning, pp. 851–856, Mahwah, NJ, (2001). Lawrence Erlbaum.
[29] L. J. Rips, *The Psychology of Proof*, MIT Press, Cambridge, MA, 1994.
[30] *The Mental Models Theory of Reasoning*, eds., W. Schaeken, A. Vandierendonck, W. Schroyens, and G. d'Ydewalle, Lawrence Erlbaum, Mahwah, NJ, 2006.
[31] S. E. Toulmin, *The Uses of Argument*, Cambridge University Press, Cambridge, UK, 1958.
[32] D. N. Walton and E. C. W. Krabbe, *Commitment in Dialogue*, State University of New York Press, Albany, NY, 1995.
[33] Y. Yang and S. Bringsjord, 'Mental Metalogic: A New Paradigm for Psychology of Reasoning', in *Proceedings of the Third International Conference of Cognitive Science*, eds., L. Chen and Y. Zhuo, pp. 199–204, Hefei, China, (2001). USTC Press.
[34] Y. Yang and S. Bringsjord, 'Mental metalogic and its empirical justifications', in *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, eds., R. Alterman and D. Kirsh, pp. 1275–1280, Boston, MA, (2003). Cognitive Science Society.

[8] Namely, $\mathcal{M}$ attempts one or more of the following (adapted from [7]): (i) contribute to another acquiring the belief in a falsehood; (ii) contribute to another continuing in the belief in a falsehood; (iii) contribute to another ceasing to believe in a truth; (iv) contribute to preventing another from acquiring the belief in a truth.

[9] Namely, $\mathcal{M}$ attempts one or more of the following: (i) contribute to another acquiring the belief in a truth; (ii) contribute to another continuing in the belief in a truth; (iii) contribute to another ceasing to believe in a falsehood; (iv) contribute to preventing another from acquiring the belief in a falsehood.