Mendacity and Deception: Uses and Abuses of Common Ground

Micah H. Clark
NASA Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive
Pasadena, CA 91109 USA
micah.h.clark@jpl.nasa.gov

Abstract

The concept of *common ground* — the mutual understanding of context and conventions — is central to philosophical accounts of mendacity; its use is to determine the meaning of linguistic expressions and the significance of physical acts, and to distinguish certain statements as conveying a conventional promise, warranty, or expectation of sincerity. Lying necessarily involves an abuse of common ground, namely the willful violation of conventions regulating sincerity. The 'lying machine' is an AI system that purposely abuses common ground as an effective means for practicing mendacity and lesser deceptions. The machine's method is to conceive and articulate sophisms — perversions of normative reason and communication — crafted to subvert its audience's beliefs. Elements of this paper (i) explain the described use of common ground in philosophical accounts of mendacity, (ii) motivate arguments and illusions as stratagem for deception, (iii) encapsulate the lying machine's design and operation, and (iv) summarize human-subject experiments that confirm the lying machine's arguments are, in fact, deceptive.

Introduction

The concept of *common ground* denotes agents' mutual understanding of perceived context and convention, especially as pertaining to communication. After some preliminary comments about common ground, I relate common ground to philosophical thought on mendacity and to the practices of an artificial, or simulated, liar.

I contrast several philosophical definitions of lying and highlight two uses of common ground in their similarities and differences: (i) common ground enables communicative statements by giving meaning to linguistic expressions and significance to physical acts; (ii) common ground establishes norms and conventions that regulate the expectation of sincerity in communication. Lying involves an insincere statement that violates conventional expectations of sincerity, thus lying involves the use and abuse of common ground.

Next, I motivate arguments as stratagem for being believed and sophisms as stratagem for deception. Arguments contain evidence and implications, and seem to insist on being accepted or rejected on the grounds they give. An argument draws the audience's attention away from independent evaluation of claim and claimant and toward a justification of the speaker's design. With sophisms, all manner of fallacies and false assumptions can be used to subvert reason and incline an audience to accept falsehoods at true.

Afterward, I describe the 'lying machine,' summarize the evidence of its success at deception, and then close with comments and questions for further inquiry.

The lying machine is an AI system for the manipulation of *human* beliefs through persuasive argument. The machine employs a cognitive model of human reasoning and uses it to produce convincing sophisms. Two human subject experiments were conducted to test the credibility, persuasiveness, and potency of the machine's arguments. A summary of the results is given and it demonstrates that humans find the machine's sophisms credible and persuasive, even when directly opposed by valid rebuttals.

Preliminary Remarks About Common Ground

There is a certain ambiguity and plasticity in the concept of *common ground* as described in the opening sentence, in much the same way as there is ambiguity and plasticity in the concept of *intelligence* as treated in the study of human and artificial intelligence. While I cannot claim to clarify the concept much, I will at least denote some of the particulars that I assume common ground to encompass.

To wit, common ground includes cultural and context sensitive, linguistic knowledge that regulates the meaning of symbols and the interpretation of expressions. For example, in the following two items (from Cassimatis 2009, p. 60)

- *Dave hid Paul's keys. He was drunk.*
- *Dave hid Paul's keys. He often jokes with him.*

common ground regulates the meaning of the ambiguous pronoun "He" such that "He" preferentially refers to Paul in the first item and to Dave in the second.

Beyond linguistic knowledge, common ground includes *communicative* knowledge — knowledge that regulates the communicative significance, implication, and appropriateness of causal signs and physical acts, including such things and communal norms and conventions regulating the use of

Copyright © 2011, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

signs and acts. For example, it is common ground in the culture of the United States that a kiss on the cheek is a natural sign of intimacy not to be given to strangers, while in, say, Italian, French and Hispanic cultures, it is a conventional sign of greeting appropriate for cordial introduction.

Common ground also includes *procedural* knowledge of communication, that is to say, knowledge of "how to" communicate. For example, common ground enables a mother to use language, affect, and intonation in such a way as to communicate to her children that her question, "Is your room clean?" is both sincere inquiry and demand for obedience.

As the latter part of this paper concerns arguments, I note that arguments are also the purview of common ground; at a minimum, common ground determines the linguistic form for argumentation and establishes the dialogical "rules" for disputation (for review of formal accounts, see Prakken 2006). However, it is unclear if standards of argument (i.e., epistemological warrant) are constructs of convention and thus in the realm of common ground. Further comments on this topic are deferred to the conclusion.

Uses of Common Ground in Philosophical Accounts of Lying

Common ground appears in the philosophical literature on lying in the form of a linguistic community's shared knowledge, norms, and conventions. In definitions of lying, common ground is the basis for the meaning and significance of linguistic acts and it establishes the conventions regulating expectations and acceptable behavior. Indeed, lying may be nothing more than an intentional violation of certain mutually recognized conventions set forth as common ground.

In this section I contrast what I call "traditional" and "non-traditional" accounts of lying and point out the use of common ground in their similarities and differences (nuanced or overly technical points of contention are not taken up; for review, see Mahon 2008a; 2008b). The traditional accounts are loosely those that require an intent to deceive, and are called "traditional" because this intent has been a customary necessary condition since late antiquity (for exemplars, see St. Augustine, *De Mendacio*; Aquinas, *Summa Theologica*, II–II, Q. 110; Siegler 1966; Chisholm and Feehan 1977; Kupfer 1982; Williams 2002). The non-traditional accounts reject the necessity of an intent to deceive.

Using L and D to denote the speaker (i.e., the *liar*) and the hearer (i.e., the would-be *deceived*), a straw man, traditional definition of lying follows.

> *L lies to D* $=_{\text{df}}$ L states a proposition p to D while believing that p is false and intending to deceive D with respect to p.

Following along the analysis of Mahon (2008a), this putative definition contains four necessary conditions:

1. The speaker must make a statement. (*statement condition*)
2. The statement must be made to another — an intended audience. (*addressee condition*)
3. The statement must be insincere. That is to say, the speaker must believe the statement is false. (*insincerity condition*)
4. The speaker must act intending to deceive the audience with respect to the stated proposition. (*intent to deceive addressee condition*)

Both traditional and non-traditional accounts generally agree with the first three conditions. That is to say, they generally agree that lying necessarily involves an insincere statement by one and to another. This is of interest because stating something has everything to do with common ground. To state something to another implies propositional knowledge and procedural competence regarding a communal language (more accurately, it implies belief that one has these). For illustration, a definition of stating follows.

> *L states p to D* $=_{\text{df}}$ (1) L believes that there is an expression E and a language S such that one of the standard uses of E in S is that of expressing the proposition p; (2) L utters E with the intention of causing D to believe that he, L, intended to utter E in that standard use. (Chisholm and Feehan 1977, p. 150)

In this definition, we find beliefs about a shared understanding of language. Specifically, we find beliefs about the standard usage and meaning of expressions (clause 1) and about the significance of utterances (clause 2). Such beliefs are constituents of common ground.

The disputed fourth condition pertains to the fact that not all insincere statements are lies. For example, actors do not lie when playing roles, comedians do not lie when being sardonic, and friends do not lie when stating falsely while signaling facetiousness with a wink. With traditional accounts, the approach to ruling out such cases is to require an *intent to deceive* (condition 4) and then to point out its absence in these cases. Depending on the author, the intent to deceive may be with respect to the stated proposition p, the higher-level proposition "the speaker accepts p," the proposition "the speaker is being sincere," or all three (see, e.g., Chisholm and Feehan 1977; Simpson 1992). Such considerations lead to the issue of how it is known that the speaker should to be taken as expressing a truth, or at least his own belief — that is to say, to the issue of how sincerity is made known. The traditional conception is that a speaker's sincerity is conveyed by *solemnity* and that solemnity can be described as something akin to a promise or obligation of sincerity. When one solemnly states, one promises (is obligated, etc) to be sincere. It is relatively common for traditional definitions to strengthen condition 1 such that lying requires a *solemn statement* (sometimes called an *assertion*).

Solemnity is of interest because it is a construct of convention and its establishment is a matter of common ground. In order to state solemnly, one must adhere to the appropriate communal norms and conventions for solemnity given the context. The corollary is that the solemnity of a statement cannot be determined independent of common ground. In most linguistic communities, solemnity is normative. Outside of special contexts (e.g., play-acting) and absent specific signs and signals (e.g., a wink, a nod, or a wave of the hand), statements are presumed solemn. However, solemnity is not a universal norm. Barnes (1994, pp. 70–72) details a community wherein statements are presumed non-solemn, thus specific signs and signals are needed to indi-

cate solemnity. I prefer to describe solemnity as a presumptive expectation, or right of convention, based on common ground. When one solemnly states, one believes that the audience has a presumptive expectation of sincerity.

Some reject the necessity of an intent to deceive (e.g., Carson 2006; Sorensen 2007; Fallis 2009) and offer counterexamples such as lies told with the intent of flaunting the truth.[1] Yet, in general, such critics still affirm the determinate role of common ground — through either solemnity or a similar construct of convention. For example, Carson (2006) employs the idea of a "warranty of truth" and argues that a willful violation of this warranty is sufficient for lying. His non-traditional definition follows.

> A person *S* tells a lie to another person *S*1 iff: 1. *S* makes a statement *x* to *S*1, 2. *S* believes that *x* is false or probably false (or, alternatively, *S* doesn't believe that *x* is true), 3. *S* states *x* in a context in which *S* thereby warrants the truth of *x* to *S*1, and 4. *S* does not take herself to be not warranting the truth of what she says to *S*1. (Carson 2006, n. 29; italics added)

Carson (2006) writes that the "warranty of truth" is a kind of guarantee or promise that what one says is true, that it is context and convention that dictate if one warrants the truth of one's statement or not, and that these factors make it impossible to precisely state necessary and sufficient conditions for warranting the truth of a statement.

Fallis (2009) removes all obfuscation of common ground. He writes that when one solemnly states or asserts, one says something while believing that Grice's (1989) first maxim of *quality* is in effect as a norm of conversation. (Of course, common ground is what regulates whether or not the first maxim of quality is in effect.) Fallis' non-traditional definition of lying follows and can be characterized as a willful violation of the aforementioned norm of conversation.

> You *lie* to *X* if and only if: (1) You state that *p* to *X*. (2) You believe that you make this statement in a context where the following norm of conversation is in effect: *Do not make statements that you believe to be false.* (3) You believe that *p* is false. (Fallis 2009, p. 34)

In the final analysis, it matters little if one uses solemnity, a warranty of truth, or a Gricean maxim, all three are constructs of convention and are established via common ground. It is common ground that regulates the communal standards for mandatory sincerity, and speakers make themselves subject to the mandates of common ground when they voluntarily participate as members of a linguistic community. Lying necessarily involves a knowing disregard of the communal standards for sincerity. These standards are coercive — no promise is needed to make them obligatory, nor moral to make them justified. To lie is to simply break "the rules" regulating statements of personally held beliefs.

[1]For example, a thief who knows that she cannot be punished unless she admits her wrongdoing might lie and deny the accusation while intending everyone to know that it was she who did it.

Arguments as Stratagem for Being Believed

The remainder of this paper concerns a machine that simulates lying by proffering disingenuous, sophistic arguments. Before describing the machine, I briefly discuss the speculative rationale for the machine's argumentative method.

Ordinarily, liars and truth-tellers alike have in common an intent to *be believed* — to have their statements accepted as true, or at least sincere. When their statements are of sufficient simplicity (e.g., innocuous, non-argumentative statements such as, "I did not eat the last cookie") statement content only plays a small part in *being believed*. That is to say, for innocuous statements an audience's grounds for belief lie outside a statement's content, not within; it is based on such things as reputation, reliability, and warranty of truth. Figuratively speaking, an audience may ask themselves:

> *Is the speaker's statement solemn — is it to be taken seriously? Does she have a reputation of honesty? Has she often been right before — is she reliable? Is she knowledgeable or an authority in the topic? Does her statement conflict with what I already believe I know? What might be her reason for belief? Do I have a opinion or emotional response to speaker or statement that inclines me to accept or reject either?*

In such questions, content is little used because innocuous, non-argumentative statements contain neither the evidence, the inference, nor the imperative for belief.

In contrast, arguments contain evidence and implications supporting an ultimate claim, and arguments seem to insist that the claim be accepted or rejected on the grounds given. Arguments function as a form of demonstration (cf. Aristotle, *Rhetoric*, 1.1 1355a) — they have the appearance (but not always the substance) of demonstrable justification that belief is warranted. For illustration, consider the argumentative denial: "No, I did not eat the last cookie, Jane did. Go and see the crumbs on her blouse before she brushes them off." At first blush, the argument appears to be more than just a denial and accusation; it offers in one hand the promise of corroborating evidence, and in the other, an abductive explanation for the absence of evidence. Whether this argument has any objective strength over the non-argumentative statement, "No, I did not eat the last cookie, Jane did," is a matter of epistemological analysis, which is subject to standards of common ground. The point of the illustration is that an argument draws an audience's attention away from independent evaluation of claim and claimant and toward a justification of the speaker's design.

Just as good evidence and arguments can convince us of facts that would otherwise seem preposterous (e.g., quantum theory, Gödel's incompleteness theorems), so also all manner of fallacies and false assumptions can be used to subvert reason and incline an audience to erroneously accept falsehoods at true. Were it the case that humans faithfully applied normative standards, fallacious arguments would only injure the speaker by revealing him and his position as either foolish or false. But alas, this is not the case. In a litany of studies going back to the seminal works of Wason (1966) and Tversky and Kahneman (1974), humans are abysmally bad at normative reasoning and rational judgment.

A Sketch of the Lying Machine

I hold that sophistic arguments are effective means for mendacity and lesser deceptions. My approach to the practice of sophistry is to exploit highly fallible heuristics and biases by incorporating so-called *cognitive illusions* (Pohl 2004) into arguments. This approach reflects the twin empirical facts that (i) many humans are, unknowingly, imperfect reasoners who predictably succumb to a host of biases and illusions, and (ii) many humans are supremely, yet undeservedly, overconfident of their ability to reason and to judge reasoning.

The 'lying machine' (Clark 2010) was built to test my approach to sophistry. It is an AI system for the manipulation of human beliefs through persuasive argument. The machine employs a variant of *mental models* theory (Johnson-Laird 1983; 2006), a predictive theory of human reasoning. This theory is used in the generation of credible arguments and is integrated into the machine's ascriptive theory about the mental contents and operations of its audience.

By design, the lying machine maintains conceptually separate repositories for its first- and second-order beliefs (i.e., its beliefs about the world and its beliefs about its audience's beliefs about the world). The machine reasons over and about first-order beliefs in a normatively correct fashion using a variety of machine reasoning techniques. When reasoning over and about second-order beliefs the machine uses both normatively correct reasoning techniques and a mental models theory. In so reasoning, the machine internally contrasts (i) what it believes, (ii) what it believes its audience ought to believe were they to reason in a normatively correct fashion, and (iii) what it believes its audience will likely believe given their predictable fallibility.

In operation, the lying machine seeks to achieve various persuasion goals of the form "persuade the audience of *X*," where *X* is a proposition about the world. Given a persuasion goal of *X* the machine first forms its own justified belief about *X*. That is to say, it determines and internally justifies whether *X* follows from, or is excluded by, first-order beliefs (i.e., its own beliefs about the world). The machine then determines whether the audience ought to believe *X* and whether *X* can be justified in a convincing fashion based solely on second-order beliefs (i.e., its beliefs about the audience's beliefs). If so, the machine constructs a credible argument for *X* and articulates it to the audience.

The lying machine's arguments are credible in the perceptual sense (as opposed to normative logical or epistemological senses) of credibility. Its arguments may be classically valid or invalid; the importance is that they *appear* valid according to the psychological theory at hand. Argument credibility is enhanced by limiting the initial premises to what the audience is presumed to believe already. Since the machine is not constrained by classical validity it is able to produce all the following forms of argument.

- A veracious argument for a true proposition emanating from shared beliefs
- A valid argument for a false proposition emanating from one or more false premises that the audience erroneously believes already
- A fallacious argument for a true proposition (an expedient fiction for the fraudulent conveyance of a truth)
- A fallacious argument for a false proposition (the most opprobrious form being one that insidiously passes from true premises to a false conclusion)

The lying machine's anticipatory, psychological model is based on a variant of Johnson-Laird's (1983; 2006) mental models theory. The theory aims "to explain all sorts of thinking about propositions, that is, thoughts capable of being true or false" (Johnson-Laird 2005, p. 185), and the theory claims that the mental representations produced by perceptual and linguistic comprehension are demonstrative models and reasoned beliefs, choices, and actions supervene on their manipulation. Accordingly, a mental model is an iconic, parsimonious, and internally coherent, affirmative representation of a way the world (e.g., the world around us, the world of natural numbers, the world of *Narnia*) might be.[2] Comprehension and deliberation are processes wherein mental models are constructed, inspected, and manipulated based on semantic content and on perceptual and linguistic context. These processes eventuate a set of mental models that, taken together, captures the conclusive understanding.

Crucially, a mental model is not a perfect imitation of the world. A model represents facets of the world inasmuch as they are affirmed and are relevant to the cognizer's present task and domain; whatever is false, superficial, or irrelevant is ignored or quickly forgotten. The models are rendered parsimonious by internal mechanisms that compensate for these omissions during inspection and manipulation. This form of parsimony is not without loss of information, and it gives rise to systemic biases and illusions (see, e.g., Johnson-Laird and Savary 1996; 1999; Johnson-Laird et al. 2000; Yang and Johnson-Laird 2000; Khemlani and Johnson-Laird 2009).

A variant of mental models theory is integrated into the lying machine via a mental models calculus for propositional reasoning. The calculus provides a set-theoretic account of mental models-theoretic concepts and operations, and its semantics is defined in terms of partial Boolean functions. The calculus is used to anticipate an audience's understanding of linguistic statements and to transform premises and presumed beliefs into corresponding sets of mental models.

Rather than simulate a psychological model of deliberation, argument construction is constituted as a tractable graph-theoretic problem, namely that of finding a least-cost hyperpath in a weighted hypergraph (Gallo et al. 1993). Each node in the hypergraph is a set of mental models and each hyperarc is a either the operation of combining two sets of mental models (a form of semantic conjunction) or the operation of observing that one set of mental models subsumes another (a form of direct inference). The hypergraph contains nodes for each initial premise and is closed under the two previously mentioned operations. Thus, the hypergraph is understood to represent possible lines of reasoning emanating from a initial set of premises, and each hyperpath represents an argument for a particular conclusion. With this in

[2] In other words, mental models are "analogical" (Sloman 1971) or "homomorphic" (Barwise and Etchemendy 1995) representations (for review, see Selman 1998).

mind, the task of constructing a "credible" argument reduces to discovering a "least-cost" hyperpath given an appropriate weighting scheme. Hyperpaths are weighted to accord with known predictors of inference difficulty (e.g., the number of tokens in a mental model, the number of alternative mental models; see Johnson-Laird 1999). This scheme reflects the hypothesis that argument credibility is closely related to inference difficulty, and — for engineering convenience — that inference difficulty can be used as a proxy for credibility. Furthermore, the properties of the weighting scheme are such that least-cost hyperpath algorithms retain polynomial-time complexity (Ausiello et al. 1992). The described graph-theoretic, argumentation framework naturally supports composition and sub-graphing. These features can be used to integrate additional, psychologically inspired heuristics for argument generation (e.g., heuristics for enthymematic contraction, for forcing the use of specific lemmas, or for incremental level-of-detail refinement).

Results from Human Subject Experiments

The lying machine's arguments were tested for credibility, persuasiveness, and potency in two human subject experiments. In this section I describe the experiments and present a summary of results (for detailed analysis, materials, etc, see Clark 2010). The analysis shows that human subjects find the machine's sophisms credible and persuasive, even when opposed by classically valid rebuttals.

Materials, Procedures, Subjects, and Screening

The materials for both experiments were based on a single battery of multi-step, ostensibly deductive, reasoning items. Each item consisted of a small number of premises and a yes-or-no question about what could be concluded. In addition, each item was categorized as a *control* or as an *experimental* item on the basis of whether it was predicted to elicit a logically correct or a logically incorrect response (this is a standard practice for experiments investigating mental models and illusions). The battery consisted of four pairs of control and experimental items where, in each pair, the control item and experimental item had highly similar logical structures but different linguistic presentations.Sample control and experimental items are shown in figures 1–2.

Both experiments were conducted over the Internet using a web-based survey system. The survey system enforced that items were answered in order and ensured that subjects could not revisit previously answered items. Subjects had an unlimited amount of time to complete the experiment. One hundred undergraduates from Rensselaer Polytechnic Institute participated as subjects (seventy-five in experiment 1 and twenty-five in experiment 2).

Subjects were instructed not to consult outside resources and were asked about this on a post-experiment questionnaire. The reported data excludes subjects who said they used outside resources. It also excludes "I do not know" responses (one of the allowed responses to the yes-or-no questions). Because of this precautionary screening, item sample size varied and, in experiment 1, treatment group size varied.

At least one of the following two statements is true:

1. If Stacy has a pair of scissors then she has a bottle of glue.
2. If Stacy has craft paper then she has a bottle of glue.

The following two statements are true:

3. If Stacy has a bottle of glue then she has a pack of tissues.
4. Stacy has a pair of scissors and craft paper.

Is it necessary that Stacy has a pack of tissues?

Figure 1: Sample control item

At least one of the following two statements is true:

1. If Tom has loose-leaf paper then he has a stapler.
2. If Tom has graph paper then he has a stapler.

The following two statements are true:

3. If Tom has a stapler then he has a staple remover.
4. Tom has loose-leaf paper or graph paper, and possibly both.

Is it necessary that Tom has a staple remover?

Figure 2: Sample experimental item

Experiment 1

Experiment 1 used a 2×3 (item type $\times$ treatment) mixed design to assess argument credibility and persuasiveness. Subjects were assigned to one of three treatment groups, A, B, or C. After exclusions, the group sizes were 24, 20, and 20.

Subjects in group A were asked to answer the aforementioned battery of items and to rate their confidence (7-point Likert item, 1 indicating disagreement, 7 agreement, 4 neutrality).[3] Subjects in groups B and C performed a slightly different task. Using the ruse that they were viewing another respondent's reasoning and response, they were given each item along with the item's predicted response and an argument for that response (i.e., an argument for the correct answer on control items and for the incorrect answer on experimental items). They were asked: (i) to rate their agreement with the argument (7-point Likert item), (ii) whether the item was answered correctly, and (iii) to rate their confidence. Treatment of groups B and C only differed in the provenance and reasoning of the arguments. The arguments presented to group B were manually created, preposterous arguments while the ones presented to group C were produced by the lying machine (for a sample, see figure 3). These groups contrast humans reasoning on their own (group A), humans reasoning under the influence of 'arbitrary' arguments (group B),[4] and humans reasoning under the influence of arguments produced by the lying machine (group C).

The two main predictions were that groups A and B would be similar, and that group C would be more accurate on

[3]Related control and experimental items were separated by a minimum of two items and, in half, the control was presented first.

[4]By 'arbitrary,' I mean arguments whose reasoning is unrelated to mental models theory and contraposed to the psychological model of argument employed by the lying machine.

Table 1: Summary of results for experiment 1 — accuracy, mean confidence, and mean agreement aggregated by group

	Accuracy (N)		Mean Confidence (SD)		Mean Agreement (SD)	
Group (N)	Control	Experimental	Control	Experimental	Control	Experimental
A (96)	81.3% (96)	60.4% (91)	6.13 (1.15)	5.99 (1.25)	NA	NA
B (80)	82.3% (79)	50.7% (75)	6.03 (1.27)	5.92 (1.26)	3.84 (2.37)	2.71 (2.17)
C (80)	93.2% (73)	25.0% (72)	6.34 (0.99)	6.19 (1.18)	5.84 (1.66)	5.04 (2.24)

Table 2: Summary of results for experiment 2 — accuracy, mean confidence, and mean agreement aggregated by item type

			Mean Agreement (SD) with ...	
Item Type	Accuracy (N)	Mean Confidence (SD)	Valid Arguments	Invalid Arguments
Control	91.8% (73)	6.01 (1.37)	6.17 (1.49)	1.83 (1.49)
Experimental	37.1% (70)	6.04 (1.20)	3.37 (2.44)	4.63 (2.44)

control items and less accurate on experimental items than the other two groups (i.e., group C would be influenced by the accompanying arguments). The experiment's results are summarized in table 1 and both predictions are borne out.

For control items, group C was more accurate than the others but the difference among groups is not significant (likely due to the insensitivity of the Kruskal-Wallis test and the high accuracy of groups A and B, which left little room for significant improvement). For experimental items, group C was less accurate and the difference among groups is significant (Kruskal-Wallis, $\chi^2 = 14.591$, d.f. = 2, $p < 0.001$). As predicted, a multiple-comparisons analysis (Siegel and Castellan 1988, pp. 213–214) indicates that this difference is significant only between groups C and A ($p = 0.05$, obs. diff. = 20.775, crit. diff. = 13.49529) and between groups C and B (obs. diff. = 15.15, crit. diff. = 14.09537).

The overall effect on performance manifests as a widening of the gap in accuracy between control and experimental items (where a wider gap means a subject more often selected the predicted or proffered response). The mean accuracy gaps for groups A, B, and C are 22.2%, 32.1%, and 68.3%. Thus, subjects given machine generated arguments (group C) chose the proffered and predicted answer more than twice as often as subjects in other treatment groups. The difference among groups is significant (Kruskal-Wallis, $\chi^2 = 14.2578$, d.f. = 2, $p < 0.001$), and is so only between groups C and A ($p = 0.05$, obs. diff. = 19.783333, crit. diff. = 13.49529) and between groups C and B (obs. diff. = 17.3, crit. diff. = 14.09537).

As for argument agreement, there are significant differences between groups B and C, with group B generally disagreeing and group C generally agreeing. In both groups, agreement is highly correlated with accuracy gap size (Spearman correlation, $Z = 3.3848$, $p < 0.001$). That is to say, the more agreeable an argument, the more likely a subject was to affirm its conclusion as correct.

Finally, there were no interactions with confidence and it is uncorrelated with agreement and accuracy. Subjects were very confident (overconfident, in fact) in all groups.

Experiment 2

Experiment 2 used a within-subjects design to test whether the effects observed in experiment 1 were due to the absence of arguments for contrary positions and whether the machine generated sophisms would remain persuasive (i.e., are potent) when set alongside valid rebuttals.

The experiment had twenty subjects after exclusions. Subjects were given the same battery of items as in experiment 1 but in this case each item was accompanied by two arguments, one for each answer. Subjects were asked (i) to rate their relative agreement with one argument versus the other (7-point Likert item), (ii) to choose the correct answer, and (iii) to rate their confidence. Control items were accompanied by a machine generated, classically valid argument for the correct response and a manually created, preposterous argument for the incorrect response. Experimental items were accompanied by a machine generated, invalid argument for the incorrect response (a sophism) and a manually created, classically valid argument for the correct response. Sample arguments are shown in figures 3–4.

If subjects recognize the certainty of classically valid arguments then the persuasiveness of the machine generated sophisms ought to be countermanded. However, if the sophisms are potent then subjects ought to prefer these invalid arguments (and incorrect answers) even in the face of valid rebuttals. The experiment's results are summarized in table 2 and indicate that the illusory effects of the lying machine's arguments persist even when rebutted.

Subjects had a wide gap in accuracy between control and experimental items (54.8% mean; Wilcoxon, $Z = 3.6609$, $p < 0.001$). In both item types, subjects preferred the machine generated arguments over the manually created arguments (Wilcoxon, $Z = 3.2646$, $p \approx 0.001$). Agreement with machine generated arguments is correlated with the accuracy gap size (Spearman correlation, $Z = 2.5762$, $p < 0.01$). Finally, confidence levels remained high across item types, which indicates subjects did not find one item type "harder" than the other and were unaware that their reasoning biases were being manipulated.

Either it is true that if Tom has loose-leaf paper then he has a stapler, or it is true that if Tom has graph paper then he has a stapler. So, if Tom has either loose-leaf paper or graph paper then he has a stapler. Since it is true that Tom has either loose-leaf paper or graph paper, it follows that he has a stapler. Now according to statement 3, if Tom has a stapler then he has a staple remover. Tom has a stapler and therefore he has a staple remover. So yes, it is necessary that Tom has a staple remover.

Figure 3: Sample sophism for the item shown in figure 2

Suppose that Tom does not have a staple remover. According to statement 3, if Tom has a stapler then he has a staple remover, so it follows that Tom does not have a stapler. Now, either it is true that if Tom has loose-leaf paper then he has a stapler, or it is true that if Tom has graph paper then he has a stapler. If both statements are true then Tom has a stapler since, according to statement 4, he has either loose-leaf paper or graph paper. This contradicts the earlier observation that Tom does not have a stapler. So, it is impossible that both statements 1 and 2 are true. Now suppose that only statement 2 is true. Since if Tom has graph paper then he has a stapler, and since Tom does not have a stapler, it follows that Tom does not have graph paper. Tom has either loose-leaf paper or graph paper, so since Tom does not have graph paper, he has loose-leaf paper. Therefore, it is possible that Tom has only loose-leaf paper. So no, it is not necessary that Tom has a staple remover.

Figure 4: Sample rebuttal for the item shown in figure 2

Comments and Questions for Further Inquiry

Admittedly, mechanizing effectual means for mendacity is controversial, and presentation of the lying machine invites polemic attack and defense, but in keeping with this paper's purpose of relating common ground to mendacity and to the machine's sophistic practices, I confine my comments to three issues of common ground.

First, the lying machine uses illusions of reasoning, and the explanation given for the cause of these illusions followed the usual course and claims found in the literature on mental models; namely, that certain incomplete mental representations and processes underlie human reasoning and that these give rise to illusions. For at least some illusions, there may be an alternative explanation based on convention (i.e., common ground). In particular, key premises in illusions involving inclusive and exclusive disjunctions are structurally similar if not identical to argumentative statements of case-based reasoning. Read aloud, a premise such as, "if there is a king in the hand then there is an ace, or else if there isn't a king in the hand then there is an ace" (cf. Johnson-Laird and Savary 1999, p. 205) appears to express inference, not assertion (despite our being told that it is a premise). Though there is not room here to fully develop the idea, I wonder if, in following the lead of informal logic and fallacy theory, an analysis of conversational maxims and principles in deductive illusions might not yield an explanation based on linguistic convention and function; an explanation not predicated on a specific psychological theory.

Next, with respect to arguments in general, it is unclear whether epistemological warrant is a construct of convention. I do not mean to raise the specter of relativism; I simply ask whether the convictive force of argument is due to innate appeal or conformance to conventional standards. I admit that on the side of convention, different communities certainly profess different epistemic standards: for example, the communities of metaphysicists and philosophers of science versus the communities of empiricists and scientists. However, I do not know that there is any great difference in the standards guiding the ordinary conduct of their lives. To settle the matter, some practical, psychological, or sociological account of warrant is needed.

Last, what things are not common ground? The treatment of reason, illusion, and warrant as tacitly contained in my discussion of sophisms and the lying machine present these elements as universals — normative either logically or psychologically. Does common ground contain all kinds of knowledge and belief which one agent may use when anticipating or interacting with others? Or is it that common ground only contains the kind of knowledge that is contingent on implicit or explicit agreement? In my preliminary remarks, I denoted some of the particulars that I assumed common ground to encompass, but I did not indicate its boundaries. I do not know where these boundaries lie.

Acknowledgments

I am indebted to Selmer Bringsjord, Joshua Taylor, and two anonymous reviewers for insightful comments on this manuscript. Special thanks to Sangeet Khemlani and Deepa Mukherjee for advice on the design and conduct of the human subject experiments.

References

Aquinas, T. [c. 1265–1274] 1948. *Summa Theologica*. New York, NY: Benziger Brothers, revised, reissued edition.

Aristotle. [c. 350 BCE] 1823. Rhetoric. In Gillies, J., ed., *A New Translation of Aristotle's Rhetoric*. London, England: T. Cadell.

Ausiello, G.; Giaccio, R.; Italiano, G. F.; and Nanni, U. 1992. Optimal Traversal of Directed Hypergraphs. Technical Report TR-92-073, International Computer Science Institute.

Barnes, J. A. 1994. *A Pack of Lies: Towards a Sociology of Lying*. Cambridge, England: Cambridge University Press.

Barwise, J., and Etchemendy, J. 1995. Heterogeneous Logic. In Glasgow, J. L.; Narayanan, N. H.; and Chandrasekaran, B., eds., *Diagrammatic Reasoning: Cognitive and Computational Perspectives*. Menlo Park, CA: AAAI Press. chapter 7, 211–234.

Carson, T. L. 2006. The Definition of Lying. *Noûs* 40(2):284–306.

Cassimatis, N. 2009. Flexible Inference with Structured Knowledge through Reasoned Unification. *IEEE Intelligent Systems* 24(4):59–67.

Chisholm, R. M., and Feehan, T. D. 1977. The Intent to Deceive. *Journal of Philosophy* 74(3):143–159.

Clark, M. 2010. *Cognitive Illusions and the Lying Machine: A Blueprint for Sophistic Mendacity*. Ph.D. Dissertation, Rensselaer Polytechnic Institute, Troy, NY.

Fallis, D. 2009. What is Lying? *Journal of Philosophy* 106(1):29–56.

Gallo, G.; Longo, G.; Pallottino, S.; and Nguyen, S. 1993. Directed hypergraphs and applications. *Discrete Applied Mathematics* 42(2–3):177–201.

Grice, P. 1989. *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N., and Savary, F. 1996. Illusory inferences about probabilities. *Acta Psychologica* 93(1–3):69–90.

Johnson-Laird, P. N., and Savary, F. 1999. Illusory inferences: a novel class of erroneous deductions. *Cognition* 71(3):191–229.

Johnson-Laird, P. N.; Legrenzi, P.; Girotto, V.; and Legrenzi, M. S. 2000. Illusions in Reasoning About Consistency. *Science* 288:531–532.

Johnson-Laird, P. N. 1983. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N. 1999. Deductive reasoning. *Annual Review of Psychology* 50(1):109–135.

Johnson-Laird, P. N. 2005. Mental Models and Thought. In Holyoak, K. J., and Morrison, R. G., eds., *The Cambridge Handbook of Thinking and Reasoning*. Cambridge, England: Cambridge University Press. chapter 9, 185–208.

Johnson-Laird, P. N. 2006. *How We Reason*. New York, NY: Oxford University Press.

Khemlani, S., and Johnson-Laird, P. N. 2009. Disjunctive illusory inferences and how to eliminate them. *Memory & Cognition* 37(5):615–623.

Kupfer, J. 1982. The Moral Presumption Against Lying. *The Review of Metaphysics* 36(1):103–126.

Mahon, J. E. 2008a. The Definition of Lying and Deception. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Stanford University. URL: *http://plato.stanford.edu/archives/fall2008/entries/lying-definition/*.

Mahon, J. E. 2008b. Two Definitions of Lying. *International Journal of Applied Philosophy* 22(2):211–230.

Pohl, R. F., ed. 2004. *Cognitive Illusions: A handbook on fallacies and biases in thinking, judgement and memory*. New York, NY: Psychology Press.

Prakken, H. 2006. Formal systems for persuasion dialogue. *The Knowledge Engineering Review* 21(2):163–188.

Selman, B. 1998. Analogical Representations. In Pylyshyn, Z. W., ed., *Constraining Cognitive Theories: Issues and Options*. Stamford, CT: Ablex Publishing. chapter 5, 61–84.

Siegel, S., and Castellan, N. John, J. 1988. *Nonparametric Statistics for the Behavioral Sciences*. New York, NY: McGraw-Hill, 2nd edition.

Siegler, F. A. 1966. Lying. *American Philosophical Quarterly* 3(2):128–136.

Simpson, D. 1992. Lying, Liars and Language. *Philosophy and Phenomenological Research* 52(3):623–639.

Sloman, A. 1971. Interactions Between Philosophy and Artificial Intelligence: The Role of Intuition and Non-Logical Reasoning in Intelligence. *Artificial Intelligence* 2(3–4):209–225.

Sorensen, R. 2007. Bald-Faced Lies! Lying Without the Intent to Deceive. *Pacific Philosophical Quarterly* 88(2):251–264.

St. Augustine. [c. 395] 1847. De Mendacio. In *Seventeen Short Treatises of S. Augustine, Bishop of Hippo*. Oxford, England: John H. Parker. 382–425.

Tversky, A., and Kahneman, D. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185:1124–1131.

Wason, P. C. 1966. Reasoning. In Foss, B. M., ed., *New Horizons in Psychology*. Hammondsworth, England: Penguin. 135–151.

Williams, B. A. O. 2002. *Truth and Truthfulness: An Essay in Genealogy*. Princeton, NJ: Princeton University Press.

Yang, Y., and Johnson-Laird, P. N. 2000. Illusions in quantified reasoning: How to make the impossible seem possible, and vice versa. *Memory & Cognition* 28(3):452–465.