

Computational Deception and Noncooperation

Anton Nijholt, University of Twente

We don't always mean what we say. We don't always fully cooperate when interacting with others. Sometimes, it's convenient to lie or omit the truth. We might not be fully honest to a conversational partner, for example, because we don't want to hurt his or her feelings. Many daily life interactions are negotiations. We have certain goals or preferences, and we want to realize them, often without being explicit. Written communication doesn't really differ from these aspects of face-to-face interactions. All this happens many times each day.¹

Human-computer interaction examines how humans interact with computers and looks at new technology that enables such interactions to be more efficient, more convenient, or more entertaining. New interaction technology uses all kinds of sensors; rather than being attached to PCs, such sensors disappear in the environment, into walls, furniture, clothes, toys, personal mobile devices, and mobile devices in the environment, such as social robots that help humans. Hence, human-computer interaction has become human-media or human-environment interaction, where the environment supports a human inhabitant in his or her activities. Providing the best possible support requires understanding and anticipating natural human behavior and not just the explicit commands that are issued to get a certain task done. When multiple users inhabit the environment, it requires understanding human-human or multiparty interaction. Members in such a party can be social robots or autonomous agents that can sometimes take the form of virtual humans that display verbal and nonverbal behavior.

In such an environment, looking at natural interaction behavior will then include looking at non-cooperative and misleading behavior and deciding how to deal with it. Obviously, this is particularly true if we must understand human-human

interaction in such environments and if we try to model natural interaction between a human and a social robot or virtual human. But, of course, we're interested in understanding, modeling, designing, and displaying misleading and noncooperative behavior in many more situations and, in particular, applications.

In a mixed-reality training and simulation environment equipped with cameras, microphones, position sensors, and maybe sensors that gather physiological information from a trainee, we might want to enter misleading information, enable miscommunication, and include virtual agents' noncooperating behavior to make it more realistic.² In a sports simulation environment, we might want to train a player on a virtual opponent's misleading actions (feints) or vice versa — that is, try to mislead the virtual opponent.³ In a more traditional educational environment, a virtual teacher can express disappointment or satisfaction when a need exists to stimulate a student, although these might not be the emotions the teacher feels.⁴ Therapy robots for treating autism have been introduced that simulate emotions to help develop communication skills.⁵

Such tools for developing social interaction skills need to know about desired behavior as well as undesired behavior. In multi-agent negotiation systems, we can send our agent away to negotiate, and, by definition, the agent shouldn't start negotiating by revealing its strategy and what its final offer will be. This must be kept secret. Agents in simulation and negotiation situations can have secrets. And virtual humans or socially intelligent robots that act in social situations must be careful in revealing information from their interaction partners to gain and maintain their trust.

Modeling and Detecting Deception

Many applications requiring knowledge about how to deceive are related to safety, security, and warfare. Speech and text analysis can help

detect deception in speech and text. Cameras, microphones, physiological sensors, and intelligent software can detect behavioral cues that identify misleading and suspicious activity in public spaces.⁶ Models of deception and noncooperation can make a virtual or mixed-reality training environment more realistic, improve immersion, and thus make it more suitable for training military or security personnel. In more serious applications, we can have robots that operate in physical and nontraining environments where they must perform military activity, including misleading the enemy (such as enemy robots) or where they're involved in rescue operations. In the latter situation, omitting the truth could help prevent a panic situation that would lead to more victims.⁷

Deception in digital games and entertainment applications is natural, whether we're talking about chess or a role-playing video game. In particular, when agents in game and entertainment situations become more intelligent, more autonomous, and more emotional, misleading the opponent becomes part of the game. When the human player has no more secrets from his or her computer opponent,⁸ the empathetic computer must determine how interesting or humiliating the game will be for its human opponent.

Deception isn't a new issue in computer science or in human-computer interaction. In fact, the Turing test is about deception. It works both ways: the computer tries to hide not being human, while its human conversational partner attempts to detect and exploit weak parts of the computer's intelligence by misleading it with trick questions; the human certainly isn't cooperative. Similarly, Joseph Weizenbaum's well-known Eliza program is about deception.

Finally, I should mention the behavior of HAL in Stanley Kubrick's *2001: A Space Odyssey* (1968). In a critical situation during a space mission, the astronauts have doubts about the decisions the spaceship's computer (HAL) has made. They try to hide their intentions to overrule HAL, but the computer has sensors that let it monitor their conversations and anticipate their actions. However, in the end, the astronauts are more clever in misleading HAL than the other way around. What happens when the computer doesn't obey us, or we don't obey it?⁹

Here, I've surveyed many aspects of computational deception and noncooperative behavior. Some aspects are still missing. For example, what is the relation between this deception and noncooperative behavior research and the many attempts to model humor in a computational way? What role do emotions (affect) play as regards the topics we'll discuss? Considerable literature is also available on lying and lie detection,¹⁰ and on noncooperative interaction behavior¹¹ and deceptive agents.¹²

In This Issue

In the six contributions to this installment of Trends & Controversies, we find state-of-the-art research approaches to the analysis and generation of noncooperative and deceptive behavior in virtual humans, agents, and robots; the analysis of multiparty interaction in the context of deceptive behavior; and methods to detect misleading information in texts and computer-mediated communication.

Ronald Arkin discusses modeling deception within a group of robotic agents. He looks at models from social psychology and cognition (partner modeling), but also introduces biologically inspired models of deception, looking at squirrels' food-protecting

strategies and bird mobbing behavior. The research from Sébastien Brault and his colleagues examines sports and deception. It doesn't really consider models of deception, but rather reports on experiments with novices and experts about presenting bodily deceptive cues in virtual reality sports training environments. Players can be trained to detect and understand deceptive movements and to perform them. The third contribution, by David Traum, covers deceptive and noncooperative behavior in virtual reality environments, where embodied conversational agents interact and negotiate with human partners about certain tasks and interests. He discusses creating such agents and application areas. Hayley Hung's contribution returns to the detection of deception's nonverbal signals (from speech and movements), but in a multiparty context where more than one deceiver might be present. Eugene Santos and his colleagues emphasize the role and the modeling of deceptive intent when modeling deceptive behavior and detecting it via human reasoning. Finally, Lina Zhou and Dongsong Zhang discuss online deception—that is, deception in computer-mediated text-based communication and how we can identify from texts using natural language processing and machine learning techniques.

This installment of Trends & Controversies provides different viewpoints on modeling deception. It also makes clear that deception researchers with different viewpoints, different theoretical and empirical approaches, and having different applications in mind can learn from each other.

References

1. B.M. DePaulo et al., "Lying in Everyday Life," *J. Personality and Social*

- Psychology*, vol. 70, no. 5, 1996, pp. 979–995.
2. D. Traum et al., “Fight, Flight, or Negotiate: Believable Strategies for Conversing Under Crisis,” *Proc. Intelligent Virtual Agents (IVA 05)*, LNCS 3661, Springer, 2005, pp. 52–64.
 3. T. Komura, A. Kuroda, and Y. Shinagawa, “NiceMeetVR: Facing Professional Baseball Pitchers in the Virtual Batting Cage,” *ACM Symp. Applied Computing*, ACM, 2002, pp. 1060–1065.
 4. D. Heylen, A. Nijholt, and R. op den Akker, “Affect in Tutoring Dialogues,” *J. Applied Artificial Intelligence*, vol. 19, nos. 3–4, 2005, pp. 287–311.
 5. K. Dautenhahn and I. Werry, “Towards Interactive Robots in Autism Therapy: Background, Motivation, and Challenges,” *Pragmatics & Cognition*, vol. 12, no. 1, 2004, pp. 1–35.
 6. J.E. Driskell, E. Salas, and T. Driskell, “Social Indicators of Deception,” *Human Factors: J. Human Factors and Ergonomics Society*, SAGE Publications, vol. 54, 2012, pp. 577–588.
 7. A.R. Wagner and R.C. Arkin, “Acting Deceptively: Providing Robots with the Capacity for Deception,” *Int’l J. Social Robotics*, vol. 3, no. 1, 2011, pp. 5–26.
 8. J. Reissland and T.O. Zander, “Automated Detection of Bluffing in a Game—Revealing a Complex Covert User State with a Passive BCI,” *Proc. Human Factors and Ergonomics Soc., Europe Chapter Ann. Meeting 2009*, D. de Waard et al., eds., Shaker Publishing, 2009, pp. 435–443.
 9. L. Takayama, V. Groom, and C. Nass, “I’m Sorry, Dave: I’m Afraid I Won’t Do That: Social Aspects of Human-Agent Conflict,” *Proc. ACM SIGCHI Conf. Human Factors in Computing Systems (CHI 09)*, ACM, 2009, pp. 2099–2107.
 10. A. Vrij, *Detecting Lies and Deceit: Pitfalls and Opportunities*, Wiley Series in Psychology of Crime, Policing, and Law, Wiley, 2008.
 11. D. Traum, “Computational Models of Non-Cooperative Dialogue,” *Proc. LONDIAL 2008 Workshop Semantics and Pragmatics of Dialogue*, abstract of invited talk, J. Ginzburg, P. Healey, and Y. Sato, eds., Queen Mary Univ. of London, 2008, pp. 11–14; www.dcs.qmul.ac.uk/tech_reports/RR-08-02.pdf.
 12. C. Castelfranchi, “Artificial Liars: Why Computers Will (Necessarily) Deceive Us and Each Other,” *Ethics and Information Technology 2*, Kluwer Academic Publishers, 2000, pp. 113–119.

Anton Nijholt is a professor of computer science in the human-media interaction department at the University of Twente. Contact him at a.nijholt@utwente.nl.

Robots that Need to Mislead: Biologically-Inspired Machine Deception

Ronald C. Arkin, *Georgia Institute of Technology*

The Georgia Tech Mobile Robot Laboratory has conducted considerable research on deception and its application within robotic systems for the US Navy. Here, I review three areas: using psychological interdependence theory as the basis for producing deception in robotic systems to evade capture; studying deception in squirrel hoarding as a means for misleading a predator regarding hidden cached resources; and mimicking bird mobbing behaviors as they apply to deceptive activity to assess the value and risks associated in feigning strength when none exists.

Deception has many definitions. The one I use to frame the rest of this discussion is “a false communication that tends to benefit the communicator.”¹ Robotics research is slowly progressing in this space, with some of

the earliest work focusing on the evolutionary edge that deceit can provide within a group of robotic agents.²

Deception and Interdependence Theory

As an outgrowth of our laboratory’s research in robot-human trust, where robots were concerned as to whether to trust a human partner rather than the other way around, we considered the dual of trust: deception. As any good con artist knows, trust is a precursor for deception,³ so the transition to this domain seemed natural. We applied the same models of interdependence theory⁴ used in our trust research and game theory to create a framework whereby a robot could make decisions regarding both when and how to deceive.⁵ This involves using *partner modeling*, a simplistic view of theory of mind that lets the robot assess a situation; recognize whether conflict and dependence exist in that situation between the deceiver and the mark, which indicates the value of deception; probe the partner (mark) to develop an understanding of its potential actions and perceptions; and choose an action that induces an incorrect outcome assessment in the partner. We implemented these results for a simple pursuit-evasion task (hide and seek) both in simulations and in successfully tested robotic systems (see Figure 1a).

Changing Strategies to Mislead

Biologists uncovered an interesting example in nature regarding deception’s possible role that involves the patrolling strategies squirrels use to protect their food caches from other predators.⁶ Normally, squirrels spend time between caches that are well stocked. Researchers observed, however, that when a predator is present (typically, conspecifics

that are interested in raiding a cache), a squirrel will change its patrolling behavior to visit empty cache sites, with the apparent intent to mislead the raider into believing those sources are where the valuables are located, a diversionary tactic of sorts. This is a form of misdirection in which communication occurs implicitly through a behavioral change in the deceiver. We've implemented this strategy in simulation⁷ and, unsurprisingly, found that these deceptive behaviors worked effectively, letting robots perform better with deception than without with respect to delaying the time for cache discovery. Figure 1b shows the experimental layout for real robots.

Deception and the Handicap Principle

Steered by our discussions with biologists, we then investigated the handicap principle⁸ as a means for understanding honest and dishonest signaling in animal species. While the original formulation of the handicap principle stated that all signaling in biology must be honest when a sufficiently high cost exists, Rufus Johnstone and Alan Grafen argued that a certain level of dishonesty (bluffing) could be introduced while preserving the system's overall stability in the presence of such deceit.⁹ This requires a delicate balance of knowing when generating such a false signal is important and its costs relative to the potential success's value. We explored this phenomenon¹⁰ in the context of

bird mobbing behavior, which served as an original case study for the *handicap principle*. This model assesses the value of a less-than-fit bird (that would be prone to capture if set upon) joining a mob where group harassment, if sufficiently strong, can lead to a predator abandoning an attack.

Our simulation studies showed that deception is the best strategy when the addition of deceitful agents pushes the mob size to the minimum level required to produce enough frustration in the predator for it to flee. In this case, the predator is driven away, and no mob member is attacked. For smaller mob sizes, complete honesty yields the lowest mortality rate because the punishment

for bluffing is high. If the cost of bluffing is reduced, adding deception can result in a reduced mortality rate when the predator attacks. Our quantitative results appear elsewhere.¹⁰ We're now importing these simulation results into our robotic platforms for further evaluation.

We've successfully demonstrated the value of biologically inspired deception in three separate cases as applied to robotic systems: pursuit evasion using interdependence theory when hiding from an enemy; misdirection based on behavioral changes; and feigning strength when it doesn't exist. Robotic deception is still in its infancy, and considerable further study is required to make definitive assertions about its overall value. This is with particular regard to situations that aren't simple one-shot deception scenarios, but rather require far more sophisticated mental models of the mark to sustain deceptive activity for longer time periods.

Serious ethical questions arise regarding deception's role in intelligent artifacts that could deceive humans.¹¹ Sun Tzu is quoted as saying that "All warfare is based on deception," and Machiavelli in *The Discourses* states that "Although deceit is detestable in all other things, yet in the conduct of war it is laudable and honorable," so it appears a valuable role exists for this capability in robotic warfare. Indeed, an entire US Army Field Manual exists on the subject of deception in the battlefield.¹² Nonetheless, when

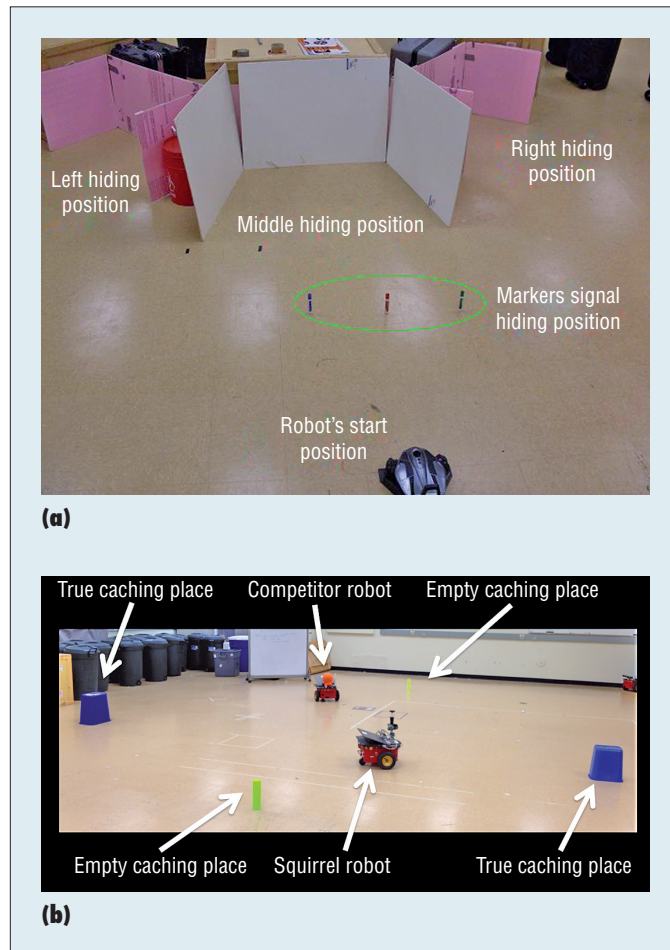


Figure 1. Machine and biological deception. We created experiments for (a) robot deception based on interdependence theory and (b) misleading competitors based on squirrel patrolling strategies.

these research ideas and results leak outside the military domain, significant ethical concerns can arise. We strongly encourage further discussion regarding the pursuit and application of research in deception as applied to intelligent machines to assess its risks and benefits.

Acknowledgments

This research was supported by the US Office of Naval Research under MURI grant number N00014-08-1-0696. I thank Alan Wagner, Jaeun Shim-Lee, and Justin Davis for their contributions.

References

1. C.F. Bond and M. Robinson, "The Evolution of Deception," *J. Nonverbal Behavior*, vol. 12, no. 4, 1988, pp. 295–307.
2. D. Floreano et al., "Evolutionary Conditions for the Emergence of Communication in Robots," *Current Biology*, vol. 17, no. 6, 2007, pp. 514–519.
3. A. Salehi-Abari and T. White, "Trust Models and Con-Man Agents: From Mathematical to Empirical Analysis," *Proc. 24th AAAI Conf. Artificial Intelligence*, AAAI Press, 2010, pp. 842–847.
4. H.H. Kelley and J.W. Thibaut, *Interpersonal Relations: A Theory of Interdependence*, John Wiley & Sons, 1978.
5. A.R. Wagner and R.C. Arkin, "Acting Deceptively: Providing Robots with the Capacity for Deception," *Int'l J. Social Robotics*, vol. 3, no. 1, 2011, pp. 5–26.
6. M.A. Steele et al., "Cache Protection Strategies of a Scatter-Hoarding Rodent: Do Tree Squirrels Engage in Behavioural Deception?" *Animal Behavior*, vol. 75, no. 2, 2008, pp. 705–714.
7. J. Shim and R.C. Arkin, "Biologically-Inspired Deceptive Behavior for a Robot," *Proc. 12th Int'l Conf. Simulation of Adaptive Behavior (SAB 12)*, Springer, 2012, pp. 401–411.
8. A. Zahavi and A. Zahavi, *The Handicap Principle: A Missing Piece of Darwin's Puzzle*, Oxford Univ. Press, 1997.
9. R. Johnstone and A. Grafen, "Dishonesty and the Handicap Principle," *Animal Behavior*, vol. 46, 1993, pp. 759–764.
10. J. Davis and R.C. Arkin, "Mobbing Behavior and Deceit and Its Role in Bio-Inspired Autonomous Robotic Agents," *Proc. 8th Int'l Conf. Swarm Intelligence (ANTS 12)*, Springer, 2012, pp. 276–283.
11. R.C. Arkin, P. Ulam, and A.R. Wagner, "Moral Decision-Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception," *Proc. IEEE*, vol. 100, no. 3, 2012, pp. 571–589.
12. *Battlefield Deception*, US Army Field Manual 90-2, 1988; www.enlisted.info/field-manuals/fm-90-2-battlefield-deception.shtml.

Ronald C. Arkin is a Regents' Professor at the Georgia Institute of Technology. Contact him at arkin@gatech.edu.

Deception in Sports Using Immersive Environments

Sébastien Brault, Richard Kulpa, Franck Multon, and Benoit Bideau, *University Rennes 2*

Why deceive in sports? Basically, to exploit an opponent in order to win a point—for example, score a goal. In sports, various research has reported that players frequently use deceptive bodily actions, and opponents' ability to detect them is a key factor for anticipation skill and sports performance.^{1,2} Additionally, Robin C. Jackson and his colleagues distinguish between a player's attempt to disguise his or her potentially informative bodily cues by minimizing them and an attempt to mislead observers via false information (deception).¹

We've explored precisely these suggestions in recent work,³ showing that

players minimized the center of mass displacement in the medio-lateral plane and lower trunk yaw (rotation) during deceptive movements in rugby. Conversely, they exaggerated outfoot displacement in the medio-lateral plane along with head and upper trunk yaw. This suggests that rugby players who want to win a one-on-one duel must take two actions. First, a player should use exaggerated body-related information to consciously deceive defenders into thinking he or she will run in a given direction. Second, the player must minimize other postural control parameters to disguise the sudden change in posture necessary to modify final running direction.³ With such an attacker's motor strategy, what might be the expert defender's perceptual strategy? Is this defender really influenced by the exaggerated parameters? How can we explore these skills in a relevant manner?

Detecting Deception Using Virtual Reality

Jackson and his colleagues didn't make a perceptual analysis of anticipation skills in deceptive movement in isolation. Their experiments tested rugby players' ability to predict an attacker's correct running direction as early as possible. Results suggested that professional players show significantly better performance than novices in terms of anticipation. To explain experts' high performance, we must determine which deceitful or disguised information is relevant. Virtual reality lets us control the displayed deceitful or disguised information, such as the virtual character's movement (we discuss virtual reality's advantages over video playback elsewhere³). We can then isolate this information to determine its influence on performance.

We proposed using virtual reality for rugby to better understand how

people explore visual information to predict an opponent's final running direction and decide on a motor response strategy. To this end, we stopped the displayed attacking actions at key moments and looked at the correlations between the displayed information and the immersed players' decision-making (see Figure 2).

We demonstrated that novices are more influenced by exaggerated parameters than experts, whereas experts are more attuned to honest signals. Regarding the defender's action response, we also showed that experts wait significantly longer than novices before initiating a displacement to "intercept" the virtual attacker. Experts can consequently initiate smaller movement in the wrong direction. Another study exposes similar results as regards exploring the influence of the ball effect on the goalkeeper's hand movement during a free kick in soccer.⁴ Results showed that hand movements were biased in the direction of initial ball heading. Consequently, during a curved free kick, the ball's initial biased direction implies larger bias displacement for novice goalkeepers compared to experts. We can thus view initial ball direction as a kind of deceptive movement.

These studies' main advantage relates to the technology used to immerse participants. Contrary to previous studies using video stimuli to explore perception-action in sports, more and more projects are based on virtual reality. Several sports, including soccer,^{4,5} rugby,⁶⁻⁸ and handball,^{9,10} have thus used it.

Training Deception Skills Using Virtual Reality

Trainers and players have already used virtual reality, mainly to improve motor skills such as coordination in rowing.¹¹ In a one-on-one duel

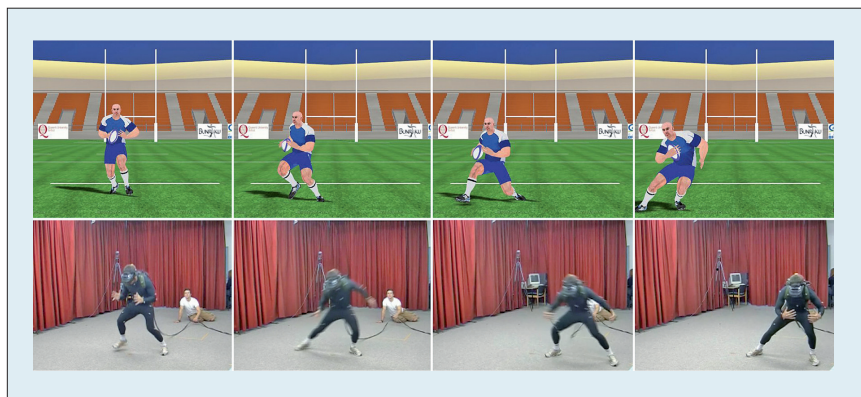


Figure 2. Duel in a virtual environment. We stopped the displayed attacking actions at key moments and looked at the correlations between the displayed information and the immersed players' decision-making.

in rugby involving deceptive movements, we can use it to train both the attacker and the defender. The defender must pick up relevant information on an opponent's movements to react accordingly. Previous experiments show that we can highlight the relevant visual information that the defender must look at to make an appropriate decision. To this end, virtual reality offers numerous possibilities for focusing the player's attention on the right information at the right time. To achieve this goal, we must evaluate which type of multisensory feedback is most appropriate in speeding up the learning process. For example, using higher- or lower-fidelity animated human figures can influence the goalkeeper in his or her information uptake when trying to intercept throws.¹⁰

To perform an efficient deception as an attacker, the player must disguise his or her potential informative bodily cues. Virtual reality can give the attacker real-time multisensory feedback about what bodily cues the defender might be able to view. The idea is to compute a model based on the logistic regression obtained in the aforementioned perceptual analysis.⁸ Each logistic regression informs us about a predictive percentage of the defender's decision to go to the right or the left as regards a given value of an attacker's visual parameter. We can

then create a model of a defender's decision-making from these real data, one for each expertise level. Experts' higher sensitivity, for example, involves a higher percentage of good answers with the same quantity and quality of information. We can then design a virtual defender and use it in virtual reality to inform the attacker about the visual body cues that the defender can use to detect deception. However, to ensure that the attacker's behavior is realistic, we must place him or her in an interactive situation with the defender. The virtual defender will then act in real time by animating a virtual human accordingly to intercept the user.

Virtual reality is an interesting tool for both understanding the detection of deceptive movements and training players to perform this detection. It nevertheless needs special care during setup and use. We must validate at all stages—for example, with regard to the perception of distance that we can alter in virtual environments or when modifying the virtual player's movement. For training, we should also check whether the skills learned in the virtual environment will really transfer to the field, and whether they last more than a few days.

References

1. R. Jackson, S. Warren, and B. Abernethy, "Anticipation Skill and Susceptibility

- to Deceptive Movement,” *Acta Psychologica*, vol. 123, no. 3, 2006, pp. 355–371.
2. R. Cañal-Bruland, J. Van der Kamp, and J. Van Kesteren, “An Examination of Motor and Perceptual Contributions to the Recognition of Deception from Others’ Actions,” *Human Movement Science*, vol. 29, no. 1, 2010, pp. 94–102.
 3. S. Brault et al., “Balancing Deceit and Disguise: How to Successfully Fool the Defender in a 1 vs. 1 Situation in Rugby,” *Human Movement Science*, vol. 29, no. 3, 2010, pp. 412–425.
 4. J. Dessing and C. Craig, “Bending It Like Beckham: How to Visually Fool the Goalkeeper,” *PLoS One*, vol. 5, no. 10, 2010, e13161.
 5. C.M. Craig et al., “Judging Where a Ball Will Go: The Case of Curved Free Kicks in Football,” *Naturwissenschaften*, vol. 93, no. 2, 2006, pp. 97–101.
 6. G. Watson et al., “Judging the ‘Passability’ of Dynamic Gaps in a Virtual Rugby Environment,” *Human Movement Science*, vol. 30, no. 5, 2011, pp. 942–956.
 7. V. Correia et al., “Perceiving and Acting Upon Spaces in a VR Rugby Task,” *J. Sport and Exercise Psychology*, vol. 34, no. 3, 2012, pp. 305–321.
 8. S. Brault et al., “Detecting Deception in Movement: The Case of the Side-Step in Rugby,” *PLoS One*, vol. 7, no. 6, 2012, e37494.
 9. B. Bideau et al., “Using Virtual Reality to Analyze Sports Performance,” *IEEE Computer Graphics & Applications*, vol. 30, no. 2, 2010, pp. 14–21.
 10. N. Vignais et al., “Influence of the Graphical Levels of Detail of a Virtual Thrower on the Perception of the Movement,” *Presence: Teleoperators and Virtual Environments*, vol. 19, no. 3, 2010, pp. 243–252.
 11. E. Ruffaldi et al., “Feedback, Affordances, and Accelerators for Training Sports in Virtual Environments,” *Presence: Teleoperators and Virtual*

Environments, vol. 20, no. 1, 2011, pp. 33–46.

Sébastien Brault is a temporary assistant professor in M2S at the University Rennes 2, France. Contact him at sb.brault@gmail.com.

Richard Kulpa is an assistant professor in M2S at the University Rennes 2, France. Contact him at richard.kulpa@univ-rennes2.fr.

Franck Multon is a professor in M2S at the University Rennes 2, France. Contact him at franck.multon@univ-rennes2.fr.

Benoit Bideau is an assistant professor in M2S at the University Rennes 2, France. Contact him at benoit.bideau@univ-rennes2.fr.

Non-Cooperative and Deceptive Virtual Agents

David Traum, *University of Southern California*

Virtual agents that engage in dialogue with people can be used for a variety of purposes, including as service and information providers, tutors, confederates in psychology experiments, and role players in social training exercises. It seems reasonable that agents acting as service and information providers, and arguably as tutors, would be truthful and cooperative. For other applications, however, such as role-playing opponents, competitors, or more neutral characters in a training exercise, total honesty and cooperativeness would defeat the purpose of the exercise and fail to train people in coping with deception. The Institute for Creative Technologies at the University of Southern California has created several role-playing characters, using different models of dialogue and uncooperative and deceptive behavior.

This article briefly describes these models, as used in two different genres of dialogue agent: interviewing and negotiation. The models are presented in order from least to most sophisticated reasoning about deception.

Most accounts of pragmatic reasoning in dialogue use versions of Grice’s cooperative principles and maxims¹ to derive utterance meanings (which might be indirect in their expression). However, these maxims, such as “be truthful,” don’t cover situations in which conversationalists are deceptive or otherwise uncooperative, even though much human dialogue contains aspects of uncooperative behavior. Gricean accounts alone don’t adequately cover cases in which conversational participants aren’t cooperative—for example, why do they ever answer at all? The notion of *discourse obligations*² differentiates the obligation to respond from the mechanism of response generation, which could be either cooperative, neutral, or deceptive.

Creating Deceptive Characters

The simplest way to create deceptive characters is for the scenario author to program the deceptive answers directly into the dialogue policy. In this way, the deception comes only from the designer, not from the character itself, which can’t distinguish deceptive from sincere utterances. We’ve used this technique for some simple interviewee characters (including C3IT³), where the trainee’s goal was to uncover which of several suspects was guilty, and deceptive answers aided in the diagnosis.

To engage in more flexible behavior, characters must know the difference between truthful and deceptive or evasive information and be able to decide on an honest, evasive,



Figure 3. Simulations for tactical questioning and negotiation. We can see human and character interactions with the (a) SASO-ST, (b) Tactical Questioning Amani, and (c) SASO4 systems.

or deceptive dialogue-management and response-generation policies. A more advanced question-answering architecture includes three levels of compliance (compliant, reticent, and adversarial), and the designer creates a different response set for each of these when creating a character.⁴ When characters are *compliant*, they provide information when asked, but fall short of Gricean cooperativity because they don't provide helpful information that was implicated rather than explicitly solicited. When characters are *reticent*, they provide neutral information, but will evade any questions about important or sensitive information. When characters are *adversarial*, they provide deceptive or untruthful answers. The characters maintain a set of social and emotional variables throughout the dialogue to determine their compliance level. These include *respect* (trainee for character and character for trainee), *social bonding*, and *fear*. When the variables change significantly, the characters change their compliance levels, letting the trainee experiment with different interview strategies, including empathy, threatening, or bargaining. Although these characters can reason about when to employ deception, the models operate globally, depending on the character's mood rather than on the specifics of the information itself. When in an antagonistic mood, the character will

lie about everything it can. The scenario author must still decide explicitly which content can have deceptive answers.

Tactical Questioning and Negotiation

The SASO-ST system provides a virtual reality environment for a trainee to practice negotiation (for example, convincing a doctor to move his clinic).⁵ It includes characters who have goals that might either align or conflict with those of a human dialogue participant. When engaged in negotiation, the characters will dynamically adopt a negotiation strategy, based on a similar, but more advanced, calculation of social and emotional variables, including trust (with subcomponents of solidarity, credibility, and familiarity), expected utility, and control. Several of these strategies are uncooperative, in that they try to achieve different outcomes than the trainee is trying to achieve. In the *avoidance* strategy, the characters will refuse to answer questions germane to the negotiation topic. However, they will answer questions that are considered irrelevant (meaning questions about issues that aren't related causally to the topic through a plan structure). Characters also use trust to decide when to believe or doubt another character. Moreover, characters use the task model structure to infer implications of what has

been said, both for Gricean cooperative purposes and to recognize hidden motivations that a speaker might like to keep private. Figure 3a shows negotiations between the character "Dr. Perez" and a human playing a captain who wants to move the doctor's clinic.

The third-generation Tactical Questioning architecture,^{6,7} allows authors much more fine-grained control in crafting sophisticated policies for uncooperative and deceptive behavior than the aforementioned models. Using the domain-editor software, an author constructs a domain-specific ontology of information (including false situations) and can create specific policies for conditions under which the character should be compliant, evasive, or deceptive about any information in the ontology. These conditions can include aspects of the emotional variables, as before, but also arbitrary aspects of the information state, including whether specific topics had previously been discussed, or whether specific incentives have been offered. Strategies for several kinds of responses have been constructed in the form of finite-state subdialogue networks, which the character can use to meet obligations while answering (truthfully or deceptively), eliciting an offer, or refusing to answer. This architecture has been used by students and other scenario authors to

construct more than a dozen different characters for purposes such as training tactical questioning (for example, the Amani character in Figure 3b), negotiation, or for use as a virtual confederate in psychology experiments.⁷ All of these characters engage in uncooperative behavior at times, and most include some deceptive aspects. Two characters in particular are used for training deception detection,⁸ where characters' verbal and nonverbal behavior can differ along various dimensions. Students can learn to follow reliable cues and discount unreliable ones.

Finally, in more recent work on the SASO architecture, we have created an ability for characters to reason more thoroughly about secrecy. In the Tactical Questioning architecture just described, authors must indicate each sensitive piece of knowledge separately. However, knowledge is often related, such that talking about one topic might reveal another closely related one. For the SASO4 scenario (see Figure 3c),⁹ each virtual character has a shameful secret and wants to avoid revealing it to the other virtual character. To achieve a successful result, the human participants must separate the virtual characters and only then will the characters bring up the sensitive issues, so that the humans can address them. When the participants are all together, the virtual characters must appear to be seriously negotiating while keeping secret the shameful reason that they're against the proposal. To achieve this with a minimum of authoring overhead, we have designed and implemented a secrecy inference scheme in which the author designates only the secret concept and who it must be kept secret from. The inference scheme then automatically marks secret related actions and states that would reveal the secret. Secret items

won't be discussed—as proposals, arguments in the negotiation, answers to questions, or justifications of other claims—if the character that the secret is to be kept from is in contact. The preliminary set of inference rules is as follows:

- The sole precondition for a secret action is secret.
- A task with a secret precondition is secret.
- A state that can be achieved only as the effect of a secret task is secret.
- The only task that can establish or remove a secret effect is secret.

These rules are currently at the level of heuristics rather than sound and complete guarantees of secrecy. In particular, the first rule might be overly conservative, because other reasons might exist for discussing the precondition. Moreover, revealing the secret assumes that other participants have similar task knowledge and inference ability. We have performed a preliminary evaluation within the SASO4 scenario. We asked eight participants to read a brief description of the scenario and the main item to be kept secret, and then rate all elements of the domain as to whether it would reveal the secret (76 total concepts, 15 derived secrets). In all but two cases, the majority view agrees with the inferences that the system makes using the set of aforementioned rules. In the two exceptions, there was a 50 percent split among the participants as to whether the concepts were secret (the inference rules mark them secret).

We have briefly presented several different architectures for creating characters that can engage in noncooperative or deceptive dialogue behavior. These vary in the authoring burden, the characters' ability to dynamically decide on and change their

behavior, and their ability to perform inference about how sensitive material is related. Generally, a trade-off exists between simplicity of authoring, the amount of authoring needed, and the inference ability that the character can perform.

References

1. J.P. Grice, "Logic and Conversation," *Syntax and Semantics*, vol. 3, Speech Acts, P. Cole and J.L. Morgan, eds., Academic Press, 1975, pp. 41–58.
2. D.R. Traum and J.F. Allen, "Discourse Obligations in Dialogue Processing," *Proc. 32nd Ann. Meeting Assoc. Computational Linguistics*, 1994, pp. 1–8.
3. P. Kenny et al., "Building Interactive Virtual Humans for Training Environments," *Proc. Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*, 2007, paper no. 7105.
4. A. Roque and D. Traum, "A Model of Compliance and Emotion for Potentially Adversarial Dialogue Agents," *Proc. 8th SIGDIAL Workshop on Discourse and Dialogue*, SIGDIAL, 2007, pp. 35–38.
5. D. Traum et al., "Virtual Humans for Non-Team Interaction Training," *Proc. AAMAS Workshop on Creating Bonds with Embodied Conversational Agents*, ACM, 2005, pp. 70–75.
6. S. Gandhe et al., "From Domain Specification to Virtual Humans: An Integrated Approach to Authoring Tactical Questioning Characters," *Proc. 9th Ann. Conf. Int'l Speech Communication Assoc. (InterSpeech 08)*, 2008, pp. 2486–2489.
7. S. Gandhe et al., "Evaluation of an Integrated Authoring Tool for Building Advanced Question-Answering Characters," *Proc. 12th Ann. Conf. Int'l Speech Communication Assoc. (InterSpeech 11)*, 2011, pp. 1296–1299.
8. H.C. Lane et al., "Virtual Humans with Secrets: Learning to Detect Verbal Cues to Deception," *Intelligent Tutoring*

Systems, LNCS 6095, V. Alevan, J. Kay, and J. Mostow, eds., Springer, 2010, pp. 144–154.

9. B. Plüss, D. DeVault, and D. Traum, “Toward Rapid Development of Multiparty Virtual Human Negotiation Scenarios,” *Proc. 15th Workshop Semantics and Pragmatics of Dialogue (SemDial 11)*, 2011, pp. 63–72.

David Traum is a principal scientist at the University of Southern California’s Institute for Creative Technologies. Contact him at traum@ict.usc.edu.

Deception Detection in Multiparty Contexts

Hayley Hung, *University of Amsterdam*

Typically, when we consider lie detection, we think about two people sitting in a room, where one of them is probably interrogating the other. One person is perhaps attached to some sort of device that can measure physiological changes, while the other asks questions and measures the physiological responses. This can be quite an intimidating and stressful situation.

However, lying or deceiving occurs much more often than we might think, and deception or lies might not always have a destructive intent. We can also view lying as a form of social lubricant¹ that people use to spare someone’s feelings—for example, not resolutely turning down a social engagement even when you have no intention of going. Under such circumstances, lying to prevent unnecessary conflict is socially acceptable. Because lying is such an embedded part of our social activity, we can easily imagine several people lying to a single person or vice versa over an extended period of time or just one meeting.

So, deception doesn’t exist just in hostile interrogation scenarios or between just one person and another. What about how we actually detect deception? Invasive methods that measure physiological signals aren’t so easy to apply in groups. However, researchers have suggested that the way we speak and move, which are known as nonverbal cues, can be indicative of deception.^{1,2} Deceiving others is a cognitively demanding task that involves trying to behave convincingly in one way, while maintaining an awareness of the actual truth of the situation. Sometimes, the stress of maintaining two different versions of reality (the true and the fictitious one) can manifest itself as nervousness or subtle differences in behavior. Then again, the stress of interrogation can be sufficient for a truth teller to behave equally nervously.¹

Challenges of Deception Detection

The problem with automated methods of deception detection is that stress can manifest itself in many different ways, through either physiological changes or delays in cognitive response. However, different people handle stress differently, and some people can train themselves to suppress such feelings. The level of stress we feel can also be related to the stakes involved in having to lie—if the benefit of lying is small, we might be less concerned about doing it.

Deception can occur in groups of more than two people and in relaxed social settings; our nonverbal behavior indicates whether we’re being deceptive or not, although no clearly generalizable set of cues exists. One aspect of multiparty deception rarely considered forms the basis of the work presented here: the hypothesis is that in multiparty settings, liars who collaborate to dupe others behave

differently in their nonverbal behavior, in when they choose to join a conversation, and even in how they behave while their “partners in crime” are speaking. In these group settings, is it more effective to detect deceptive behavior automatically using cues related to the situation’s social context?

The Werewolf Game and Idiap Wolf Corpus

To investigate this hypothesis, an experiment was devised where groups of eight to 12 people sat in a room and played a role-playing game.³ In the Werewolf Game, players are randomly cast as either villagers or werewolves, and a games master oversees the game, which game has two phases. In the night phase, all players close their eyes, and the wolves surreptitiously indicate to the games master which villagers should be killed. In the day phase, the players awake and find out which players have been knocked out. Then, the remaining “living” players discuss who could be a wolf and therefore who they should knock out of the game. The game continues with alternating day and night phases, with a player knocked out during each phase. The game is over when all the wolves are killed or the number of villagers is equal to the number of wolves.

In total, four different groups played the game, resulting in 15 games being played and 81 hours of audiovisual data for experiments. We used a role-playing game because it created a natural, nonthreatening environment in which people could lie in a group setting. We anticipated that in such a scenario, people would be more inclined to get involved and lie when necessary. We fitted each player with a headset microphone and recorded them with video cameras (see Figure 4). The data is available at



Figure 4. Video setup of the Werewolf Game. Three camera views capture the visual behavior of eight to 12 people. All views show the players facing the camera directly.

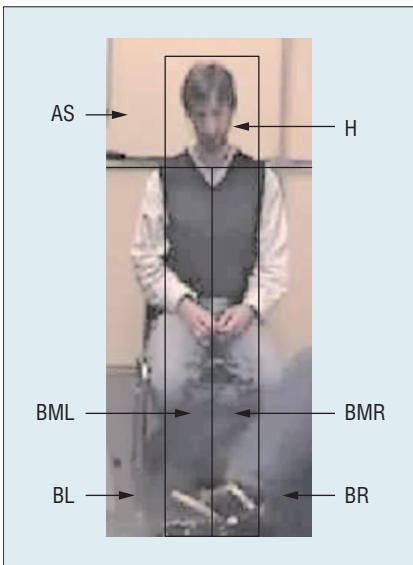


Figure 5. Extracting cues in speech and nonspeech. The example frame shows how we extracted regions of motion interest from around each person's body. H: head, AS: above shoulders, BML/R: body middle left/right, BL/R: body left/right.

www.idiap.ch/scientific-research/resources/wolf-corpus.

Automatically Extracting Nonverbal Cues

From the audio signal we recorded for each person, we extracted derived prosodic features such as power, pitch, and fundamental frequency.³ In addition, we used an existing speech/nonspeech detection system to automatically derive a binary vector representing each participant's speaking status. We also extracted cues derived from

each person's physical motion using the zones in each frame Figure 5 shows.⁴ The zones were normalized by the distance of each person's chair from the camera using the pixel width of the chair, and then using the frontal face Haar cascade classifier implementation from OpenCV to automatically determine the zones of interest. We extracted audio-visual cues by accumulating distributions of the visual features during periods of each person's speech or nonspeech.

Detecting Deceptive Roles

Within the game's framework, we attempted to identify the werewolves' roles that were explicitly designed to deceive the others. We used supervised learning methods such as a support vector machine and relevance vector machine, exploiting audio and visual features extracted from each individual's behavior to predict players' roles. The prosodic audio features weren't very discriminative, leading to a worse-than-average estimation performance. Using both audio and video cues, the performance was better, with an above-average chance of success.

When we used the speaking status feature and accumulated a feature that represented the amount of time two people were observed talking within a short time window, we observed much more striking differences in the werewolves' behavior compared to that of the villagers. Players tended to speak within the same time

window with different people when they were werewolves compared to when they were villagers.

Future Research

Using social context to detect deception in multiparty contexts is preliminary work, and further analysis is needed to investigate its influence on how deceivers behave. Recent analysis provides interesting food for thought. Figure 6 shows the distribution of visual activity for both wolves and villagers, separated by whether they were moving while a different wolf or villager spoke. Note the clear peak in the probability of low visual activity in a wolf when another wolf is speaking. Meanwhile, villagers' behavior distributions are identical regardless of whether a wolf or villager is speaking. This suggests a lot of potential for identifying subconscious influences that one deceptive speaker can have on another. By moving away from looking at individual behavior to the influence that other conversants can have on people's behavior, we'll likely be able to improve automated estimates of deception from nonverbal cues, performing significantly better than humans at the same task.

Acknowledgments

The European IST Program project FP6-033812 (AMIDA) partially supported this work. I thank Gokul Chittaranjan, Gwenn Englebienne, and Nimrod Raiman for their collaboration on the work presented.

References

1. A. Vrij, *Detecting Lies and Deceit: Pitfalls and Opportunities*, John Wiley & Sons, 2011.
2. P. Ekman, *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*, W.W. Norton, 2009.
3. H. Hung and G. Chittaranjan, "The Idiap Wolf Corpus: Exploring Group

Behavior in a Competitive Role-Playing Game,” *Proc. Int’l Conf. Multimedia*, ACM, 2010, pp. 879–882.

4. N. Raiman, H. Hung, and G. Englebienne, “Move, and I Will Tell You Who You Are: Detecting Deceptive Roles in Low-Quality Data,” *Proc. 13th Int’l Conf. Multimodal Interfaces*, ACM, 2011, pp. 201–204.

Hayley Hung is a Marie Curie postdoctoral research fellow at the University of Amsterdam, the Netherlands. Contact her at h.hung@uva.nl.

Deception Detection, Human Reasoning, and Deception Intent

Eugene Santos Jr., Deqing Li, and Fei Yu, *Dartmouth College*

Researchers have found human deception-detection skills to be no better than random guessing.¹ Although people can learn deception cues for specific individuals from past successful detections, learning opportunities for people to enhance their capabilities are still scarce.^{2,3} Possible ways to improve detection skills include having a detector focus on specific individuals and circumstances (for example, drug transportation through customs checkpoints) or on deceptions in a certain problem domain (such as accounting fraud).

Various methods concentrate on different communication channels, including facial expressions and observable physiological reactions to emotion. However, the prevalent form of human interaction has changed from the traditional face-to-face

interaction to computer-mediated communication, in which perceivers can rarely observe cues in body language and facial expressions. Faced with this challenge, researchers are investigating language-based cues based on the intuition that a deceiver selects different wording and phrasing patterns. In general, deceptive stories are often shorter and less compelling than true stories.⁴ Deceivers tend to use words that are detached from themselves, such as “you,” “others,” and “human,” whereas truth tellers choose words that are closely connected to themselves such as “I,” “friends,” and “self.”⁵ The main challenge for these methods is that linguistic cues induced in one problem domain might not transfer well to another. For instance, the deceptive reports submitted to a jury can be lengthier and even sound more compelling than typical reports.

In parallel with the development of (computational) deception-detection methods, considerable effort has also been spent on constructing benchmarking datasets for validation. Some datasets are built by collecting cases in real life. However, deceivers might deny their deceptive behavior because of real-life penalties from

and the risk of being detected. This all points to the need to account for intent in both the development and validation of detection methods.

The Role of Intent

Our group at Dartmouth College has been exploring the role of intent and its different uses in deception detection.^{7,8} We define deception as follows: In a deceptive communication, the information is false from the speaker’s viewpoint, the act is intentional, and the purpose is to take advantage of the listener. The intent of deceiving leads to deliberate manipulations and the presupposition of a goal. Deliberate manipulations refer to manipulating the argument or observations away from the deceiver’s true beliefs, while presupposing the goal indicates that all the deceiver’s behaviors aim at supporting the goal.

In our previous work,⁸ we successfully detected all malicious insider threats by detecting changes in the cognitive styles of attackers’ written reports. Changes in a person’s cognitive styles are reflected as changes in his or her underlying syntactic argument structures and semantic relation graphs. Our method’s promising performance implies that although

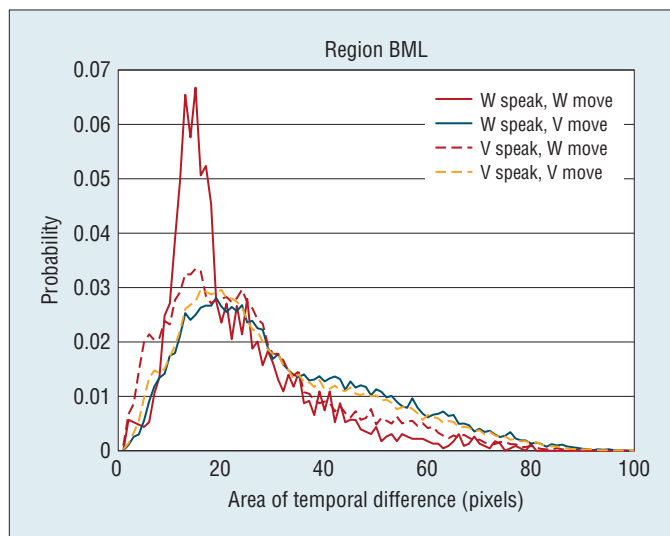


Figure 6. Detecting deception in multiparty contexts. We can see motion activity for wolves (W) and villagers (V) during the times when another W or V was speaking.

linguistic content (both syntax and semantics) exhibits wide variations, we can still quantitatively measure cognitive style. Based on this finding, we hypothesize that we might also be able to reconstruct the underlying reasoning process from the linguistic content.

Deception Detection through Human Reasoning

We investigated two major directions based on reconstructing the reasoning process from communication content.

Deception Deviates from Original Beliefs

The first direction is to detect deception by identifying self-inconsistencies in an individual's reasoning pattern. A deception deviates from the deceiver's original beliefs. Most existing studies focus on the search for general deception indicators that are applicable to any individual. Although such indicators are effective in evaluating deceivers in the same domain and might be easy to find and apply, they aren't reliable; they might be moderated by the environment, and outliers can frequently be classified as deceivers.⁷ Compared with general indicators, capturing a person's deviation from self isn't straightforward because his or her true beliefs are hidden. We've observed that acquaintances can better anticipate each other's behaviors because they know when they agree with each other and when they don't. So, one possible approach is to predict true beliefs from other truth tellers' beliefs, given that we know their past correlations. We've explored and implemented this approach in our work.⁹

We simulated the opinions of a group of agents using Bayesian networks (BNs) to model human

reasoning and learning behavior. We predicted agents' opinions based on correlations with each other that we obtained from a past history of opinions. The system can detect 87 percent of all self-inconsistencies with a false-positive rate as low as 2 percent. We also validated the self-inconsistency method using survey data in which test subjects were asked to lie about their opinions on abortion.¹⁰ To reveal the deviation from subjects' original beliefs, we used pairs of deceptive and honest stories from the same individuals. We encoded the sentiments of some common arguments into numerical values to calculate correlations between test subjects. Due to the datasets' limitations, we simulated part of the test subjects' historical opinions based on their existing opinions.

The experimental results showed a low inconsistency rate in both honest and deceptive stories, meaning that only part of the beliefs in deceptive stories were inconsistent. If we consider a story to be deceptive when one or more of its arguments are inconsistent, then the recall rate for detection is 79 percent, and the precision is 52.67 percent. Still, we must incorporate further measurements to refine the detection.

Deception Is an Intentional Act

The second major direction is to distinguish misinformation from deception. Inconsistency (consistencies with self) and untruthfulness (consistencies with others) indicate deception. Nevertheless, misinformation and other unintentional errors can also cause inconsistency and untruthfulness. We must distinguish misinformation from deception because intentional and unintentional distortions of communication are driven by different intents and can produce different consequences in the long run.

Intent to deceive indicates that deception is an intentional act, whereas other deviations in behavior, such as misinformation, are unintentional errors.

Unfortunately, little research on deception-detection realizes the importance of deception being intentional for us to distinguish deception from unintentional errors. As an intentional act, the goal is to interact with people in a particular way for some target reason; for deception, that reason is an objective the deceiver falsifies. Due to this unique feature in the deceiver's intent, his or her reasoning process when forming the deception also becomes unique.

An experienced deceiver intends to maintain the entire story's consistency, maximize its compellingness, and convey the most supportive arguments.⁷ While maintaining the consistency of the story, the deceiver simultaneously manipulates related parts; while maximizing its compellingness, the deceiver manipulates some parts of the story more; and while conveying supportive arguments, the deceiver emphasizes functional arguments and hides nonfunctional ones.

We can quantitatively measure the reasoning patterns (including argument structure) using an individual's (structural) inconsistency with him- or herself and untruthfulness (structurally) compared against truth tellers. From simulated data, we observed these patterns and identified that deception and misinformation cases are separable. A prototype system based on these patterns⁷ produces more reliable results because the deceiver's reasoning process directly follows the intent to deceive. Its unique patterns are usually unavoidable and robust to linguistic change.

A deceiver's intent plays an important role in all aspects of deception research, from the development of detection approaches to the construction of deception datasets. However, to our knowledge, no other study seeks insights from the deceiver's intent, formation of deception, or reasoning process. Little research in computational methods of deception detection even goes beyond the level of words. Understanding the formation of arguments in deception-driven reasoning is a rich avenue for exploration.

References

1. M.G. Millar and K.U. Millar, "The Effects of Cognitive Capacity and Suspicion on Truth Bias," *Communication Research*, vol. 24, no. 5, 1997, pp. 556–570.
2. C.V. Ford, *Lies! Lies! Lies! The Psychology of Deceit*, Am. Psychiatric Press, 1996.
3. P.E. Johnson et al., "Detecting Deception: Adversarial Problem Solving in a Low Base-Rate World," *Cognitive Science*, vol. 25, no. 3, 2001, pp. 355–392.
4. B.M. DePaulo et al., "Cues to Deception," *Psychological Bull.*, vol. 129, no. 1, 2003, pp. 74–112.
5. R. Mihalcea and C. Strapparava, "The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language," *Proc. Joint Conf. 47th Ann. Meeting of the Assoc. for Computational Linguistics and the 4th Int'l Joint Conf. Natural Language Processing of the Asian Federation of Natural Language Processing*, 2009 Conf. Short Papers, Assoc. for Computational Linguistics, 2009, pp. 309–312.
6. E. Santos and Q. Zhao, "Adversarial Models for Opponent Intent Inference," *Adversarial Reasoning: Computational Approaches to Reading the Opponents' Mind*, A. Kott and W. McEneaney, eds., Chapman & Hall, 2006, pp. 1–22.
7. D. Li and E. Santos Jr., "Argument Formation in the Reasoning Process: Toward a Generic Model of Deception Detection," *Proc. EACL 2012 Workshop on Computational Approaches to Deception Detection*, Association for Computational Linguistics, 2012, pp. 63–71.
8. E. Santos et al., "Intelligence Analyses and the Insider Threat," *IEEE Trans. Systems, Man, and Cybernetics: Part A*, vol. 42, no. 2, 2012, pp. 1–17.
9. E. Santos Jr. and D. Li, "Deception Detection in Multi-Agent Systems," *IEEE Trans. Systems, Man, and Cybernetics: Part A*, vol. 40, no. 2, 2010, pp. 224–235.
10. D. Li and E. Santos Jr., "Deception Detection in Human Reasoning," *Proc. IEEE Int'l Conf. Systems, Man, and Cybernetics*, IEEE Press, 2011, pp. 165–172.

Eugene Santos Jr. is a professor of engineering in the Thayer School of Engineering at Dartmouth College. Contact him at eugene.santos.jr@dartmouth.edu.

Deqing Li is a doctoral candidate in the Thayer School of Engineering at Dartmouth College. Contact her at deqing.li.th@dartmouth.edu.

Fei Yu is a doctoral candidate in the Thayer School of Engineering at Dartmouth College. Contact her at fei.yu.th@dartmouth.edu.

Automatic Deception Detection in Computer-Mediated Communication

Lina Zhou and Dongsong Zhang,
*University of Maryland,
Baltimore County*

Deception has changed dramatically since the early 21st century, both quantitatively and qualitatively, given the advance of information and

communication technologies. People become vulnerable to deception in computer-mediated communication (CMC), and an urgent need exists to increase awareness of such deception and the importance of preventing and detecting it. However, deception detection in CMC faces challenges that never existed for detecting face-to-face deception. For instance, computer-based media generally provide restricted modalities for interpreting online interaction, often limited to text only. In addition, computer-based media offer a wide range of transmission and processing capabilities, each of which warrants separate investigations. Evidence shows an average person's accuracy of lie-truth discrimination in CMC is worse than a random guess. This is largely attributed to our limited knowledge about cues to online deception as well as human beings' truth bias. Technology can help eliminate this bias and be instrumental to discovering cues to online deception. Thus, using technology support to automatically detect online deception is particularly promising.

Automatic Deception Detection in CMC

The basic deception detection process might involve solving the following problems: *cue identification* that focuses on identifying a set of verbal and nonverbal features of deceptive communication; *cue extraction* that deals with extracting and encoding those features; *deception reasoning* that infers possible deception using cues to deception; and *decision making* that involves making detection decisions based on the inference results as well as the user's or situation's level of error tolerance. Here, we discuss the first three problems in text-based CMC.

Identifying Cues from Text

Interpersonal deception theory is widely used to explain online deceptive behaviors. According to this theory, a deceiver is engaged in both strategic and nonstrategic behaviors. Strategic behaviors manifest as plans and intentions to achieve deceptive goals, whereas nonstrategic behaviors reflect the perceptual, cognitive, and emotional processes involved in deceptive communication. The latter behaviors are beyond intentional control, which we might be able to rely on as the basis for identifying cues to deception.

Given that text is the predominant medium in CMC, text-based or “verbal” cues have been studied in online deception research. Text-based cues are related to either the language or content of online messages. Compared with content-based cues, which refer to messages’ specific information features, such as veracity, completeness, and relevance, language-based cues are particularly promising because they depend less on expert knowledge and prior experience, and are relatively insensitive to the domain of discourse.

Researchers have identified language-based cues in numerous dimensions to date, including quantity, nonimmediacy, diversity, specificity, language complexity, cognitive complexity, informality, expressivity, affect, and uncertainty. For example, deceptive messages in asynchronous CMC tend to be longer, more informal and uncertain, more expressive and nonimmediate, less complex, and less diverse than truthful messages.¹ Deception in synchronous CMC involves higher levels of cognitive complexity and positive affect than truth-telling.² We can measure each linguistic dimension via specific linguistic features or styles. For instance, the numbers of words,

phrases, and messages can indicate quantity, and self-references, modal operators, and generalizing terms represent nonimmediacy.

Extracting Cues via NLP Techniques

Natural language processing (NLP) techniques have benefited and influenced the identification of language-based cues. Techniques that have been commonly adopted for extracting deception cues include morphological analysis (determining words, nonword tokens, and parts of speech in a sentence), shallow syntactic parsing (identifying some phrasal constituents without indicating their internal structures and functions in a sentence), and lexical semantic analysis (interpreting the meaning of words without resolving the entire sentence’s meaning). The final NLP output helps us derive metrics for cues to deception. For instance, we can measure emotiveness as the ratio of the total number of adjectives and adverbs to the total number of nouns and verbs. A meta-analysis of automatically extracted linguistic cues to deception reveals small but significant effect sizes for some linguistic categories.³

Deception Reasoning via Machine Learning

We can treat deception reasoning as a classification task, a typical machine learning problem. While building models for detecting deception in a given context, a classifier can learn individual cues’ weights (or importance) from previously labeled data collected from a similar context. We can use such models to classify new data into deception or truth with a certain level of confidence. By employing training data collected from different contexts, we can adapt and extend these models to these new

contexts and associated deception strategies. Given that the number of instances in the truth class greatly outnumbers those in the deception class, the development of deception reasoning models must address this imbalance with the following potential means:

- *Data sampling.* Under-sampling the majority (truth) class (randomly or selectively) and over-sampling the minority (deception) class (duplicates or interpolates) is a classifier-independent solution.
- *Parameter tuning.* To reduce the tendency of standard classifiers in learning how to predict the majority class, we can adjust the cost of errors by assigning a higher cost to instances of the deception class than to instances of the truth class.
- *Evaluation metrics.* To eliminate the bias of accuracy toward the truth class, we must also consider other performance metrics that focus on the deception class, such as false positives, false negatives, precision, and recall.

Researchers have applied various machine learning techniques to deception detection, such as neural networks, naïve Bayes, decision trees, support vector machines (SVMs), statistical language modeling, and neuro-fuzzy methods. Appropriate machine learning methods can result in deception-detection accuracy higher than 70 percent if they consider only important cues.⁴ Previous studies have demonstrated that neural networks and SVMs are superior to others in both performance and generality. In addition, language modeling and neuro-fuzzy techniques serve as good alternatives. The former can capture words’ dependency in deceptive communication without

explicit feature extraction, and the latter addresses uncertainty due to imprecision and vagueness of cues to deception, their relationships, and the outcome of deception detection.

Future Research Directions

We've made considerable progress in online deception detection through joint efforts from a wide range of disciplines, such as social sciences, communications, computer science, and information systems. In moving forward, we must build upon what's been achieved and explore new issues that merit further investigation:

- *Analyzing discourse cues to online deception.* Online interaction is discourse-rich, particularly in multiparty communication. Existing linguistic features have been selected mainly at subsentential levels (such as words and phrases). Discourse analysis can potentially better reveal the underlying intention of individual utterances or messages by understanding the information in relation to the discourse structure as a whole. This is particularly important for understanding text in CMC that often diverges significantly from traditional formal written text because the former is relatively brief, informal, and poor at managing interruptions and organizing turn-taking.
- *Identifying nonverbal cues to online deception.* Nonverbal behavior such as participatory, keyboard, social structure, and even brain activity will become increasingly available in new CMC channels. Accordingly, we can identify nonverbal cues to online deception by analyzing interaction, clickstreams, social networks, sentiment, and user brain data. Verbal and

nonverbal leakage should collectively make powerful cues to online deception.

- *Tracking changes in deception behavior.* The way deception behavior dynamically adjusts should become a source of indirect cues to deception.
- *Building standard corpora of online deception.* Deception corpora remain lacking, particularly for real-world online deception, despite a wide variety of deceptive acts on the Web.⁵ A challenge lies in determining the ground truth. Active learning is a promising method for obtaining labeled deception data for model training. Another is to treat deception detection as an outlier detection problem. Among other deceptive scenarios, online gaming has emerged as a new venue for collecting real deception data.
- *Incorporating contextual factors into deception-detection models.* Cues to deception are likely to change as deceivers and receivers or communication context (such as culture and media synchronicity) change or deceivers adapt their deception strategies. Accordingly, deception-detection techniques should be adaptive and customized.
- *Improving user satisfaction.* In addition to the effectiveness of deception-detection techniques, the success of automatic deception detection also hinges on user satisfaction and system usability.


Challenges in the automatic detection of deception in CMC will continue. The field of information and communication technology is rapidly evolving; so should our knowledge about online deception and techniques for deception detection. ■

References

1. L.J. Zhou et al., "Automated Linguistics Based Cues for Detecting Deception in Text-Based Asynchronous Computer-Mediated Communication: An Empirical Investigation," *Group Decision and Negotiation*, vol. 13, no. 1, 2004, pp. 81–106.
2. L. Zhou, "An Empirical Investigation of Deception Behavior in Instant Messaging," *IEEE Trans. Professional Communication*, vol. 48, no. 2, 2005, pp. 147–160.
3. V. Hauch et al., "Linguistic Cues to Deception Assessed by Computer Programs: A Meta-Analysis," *Proc. EACL 2012 Workshop Computational Approaches to Deception Detection*, Assoc. for Computational Linguistics, 2012, pp. 1–4.
4. L. Zhou et al., "A Comparison of Classification Methods for Predicting Deception in Computer-Mediated Communication," *J. Management Information Systems*, vol. 20, no. 4, 2004, pp. 139–165.
5. A. Vartapetian and L. Gillam, "Web Deception Detanglement," *Proc. 12th ETHICOMP Int'l Conf. Social and Ethical Impacts of Information and Communication Technology*, Sheffield Hallam Univ., 2011, pp. 445–454.

Lina Zhou is an associate professor of information systems at the University of Maryland, Baltimore County. Contact her at zhoul@umbc.edu.

Dongsong Zhang is an associate professor of information systems at the University of Maryland, Baltimore County. Contact him at zhangd@umbc.edu.

 Selected CS articles and columns are also available for free at <http://ComputingNow.computer.org>.