Deceptive and Counter-Deceptive Machines
Papers from the AAAI 2015 Fall Symposium

Impression Management, Mindshaping and the Social Function of Fibbing

Paul Bello and **Will Bridewell**
Naval Research Laboratory
4555 Overlook Ave. SW
Washington, DC 20375

Abstract

In a symposium focused on deception and counter-deception in machines, one might be immediately drawn to a narrow conception of those phenomena which highlight the pernicious ways in which they might be used. On the broader notion of fibbing that we describe in our talk, the social function of being *fast and loose with the truth* takes center stage as a tool for accomplishing a wide variety of socially centered goals. We briefly review the FIDE framework, described in (Isaac & Bridewell 2014; Bridewell & Bello 2014), including the conceptual resources it requires and the variety of fib-related concepts it supports. FIDE delineates between the aforementioned concepts as ends, and the strategic means by which the fibber might achieve these ends. In doing so, we show that certain types of difficult to conceptualize behavior, most notably *bullshitting* (Frankfurt 2006) and responses to bullshitting, are instances of a kind of strategy for impression management that serves higher-order social goals.

The Social Function of Fibbing

Deception is often conflated with a host of other activities that share some of its salient characteristics, but are different in kind. Harry Frankfurt draws a distinction in his now-classic analysis of the concept of bullshit between the latter, and deception *simpliciter* (Frankfurt 2006). For Frankfurt, the mark of bullshit is a lack of concern for the truth or falsity of what is said. To be clear, the content of bullshit can certainly *be* true or *be* false, but these are the accidents rather than the essence of bullshitting. On the other hand, deceitful statements are partially constituted by their false content. Moreover, falsity is an inseparable part of the deceiver's plan, whereas the bullshitter merely aims to present himself in a particular way. The difference between these two categories rests in the *ulterior motives* that underwrite each one. When the bullshitter sounds off on a topic, the motive is not necessarily to have the audience believe that he endorses the various propositions embedded in his bullshit, but rather to have the audience draw other sorts of inferences about his character, likes, dislikes, and so on.

For various reasons, discussions of deception are often set against the backdrop of a deceiver who has less than savory motives. But a few moments of reflection may bring to mind cases of deceit that serve the greater good: undercover police work, intelligence operations in the military, and even white lies and omissions. Similarly, *pace* Frankfurt, bullshitting can also serve the greater good—or at least be pro-social. Consider the endless number of idle conversations people have about the weather, their professions, sales at local stores, *ad infinitum*. How many times a day do people say "Oh, that's interesting," or utter a short, "mm-hm," to keep up appearances in the midst of an otherwise dreadfully boring or off-putting conversation? These exchanges are especially common when one's interlocutor is situated somewhere in the social hierarchy such that norms prescribe deference and respect. The motive of the bullshitter in these ubiquitous cases is to manage the other's impressions such that certain relevant social norms remain inviolate.

The question remains: how can we pull these complicated cases apart? Our answer to date depends crucially on the idea of using a speaker's beliefs about content and his ulterior motives to distinguish among lying and its close cousins, such as pandering and paltering (Isaac & Bridewell 2014). However, bullshitting is difficult to locate within this classificatory scheme. One can bullshit as a means to lie, pander, or palter. On this characterization, we claim that bullshitting is better considered to be a strategy for impression management than a concept or specific class of deception. In general agreement with Frankfurt, we see bullshitting as impression management, but we have also identified cases of impression management that fail to neatly fit the mold of Frankfurt's definition. The approach that we discuss in our talk involves decomposing lies, bullshit, and their cousins. What we find in all cases is that the fibber values a secondary activity (or goal or norm) above upholding what we call the *norm of truthfulness*. Distinguishing between lying, pandering, paltering, bullshitting and other forms of impression management is a matter of identifying the secondary goals and norms that supersede truthfulness (Bridewell & Bello 2014).

To this end, we describe the FIDE framework and its representational commitments in an attempt to rigorously pin down these closely related concepts. In line with the expressed goals in the call for submissions, we hope that a formal treatment will make the full range of concepts that we have discussed amenable to computational treatment and eventual implementation in intelligent systems.

Acknowledgments

The authors are grateful to Alistair Isaac for ongoing discussions and contributions to the FIDE framework. This research was funded by ONR under award numbers N0001414WX20179 and N0001415WX01339. The opinions expressed in this paper are solely the authors and should not be taken to reflect the policy or position of the United States Government or the Department of Defense.

References

Bridewell, W., and Bello, P. F. 2014. Reasoning about belief revision to change minds: a challenge for cognitive systems. *Advances in Cognitive Systems*, 3:107–122.

Frankfurt, H. G. 2005. *On bullshit*. Princeton, NJ: Princeton University Press.

Isaac, A. M. C., and Bridewell, W. 2014. Mindreading deception in dialog. *Cognitive Systems Research*, 28:12–19.