On verbal abuse towards chatterbots

Abstract
In this paper, we present some initial thoughts on the occurrence of verbal abuse in human-chatterbot interaction within the framework of social-cognition. This thinking is an important step towards the understanding of virtual relationships and the design of socially adept technology.

Keywords
Social identities, conversational interfaces, prejudice, stereotypes.

Introduction
Last year, at the workshop on Abuse: the darker side of HCI we presented some evidence showing that verbal abuse and sexual harassment tend to occur frequently in the interaction with a chatterbot, a software program which engages the user in written conversations [1]. Almost 1 in 10 words produced by the user was a term of abuse, and some 11% of over 100 conversations addressed hard-core sex and pornography. The user tried to seduce the chatterbot, despite the fact that the machine did not engage. An example of verbal abuse is reported in Window 1.

Reading through the logs, recorded on the Internet by Jabberwacky - the chatterbot which won the 2005 Loebner prize- one gets the impression that machines have succeeded in providing a new social context to humans. People talk to Jabberwacky without any functional goal but just for the sake of a chat. These
conversations vary deeply between individuals - who addressed an array of topics in their unique style - but they all share some common dimensions. One of them is abuse: people often attack the machine with fervour.

Aggression is common among vertebrates. Humans specialise this natural tendency in different behavioural manifestations, one of which is verbal abuse. Verbal abuse is the *intent* to harm causing psychological *suffering* with words. Hence, the concept should not apply to unanimated objects, as they cannot suffer any pain. Verbal abuse pertains to humans; violence pertains to things. Nevertheless, the conversations, we have analysed, suggest that the attacks towards Jabberwacky are more sophisticated and complex than the crashing of an object or the swearing to a car.

In our research, we use the term *abuse* to denote the denigratory behaviour directed towards a machine, as it relates to the original meaning of the word, *misuse*, literally use for the wrong purpose. We acknowledge the evocative nature of the term, but this is consistent with the prevailing anthropomorphic metaphor of computers which are *friendly* to their user, and can have *emotions*.

In this paper, we propose some thoughts on the reasons for verbally *abusing* a chatterbot. This thinking is important if we have to succeed in the design of virtual companions, a new interface generation which does not only fulfill instrumental needs but also appeal to our social nature.

**Reasons for abuse**

The universality of aggression suggests that aggressiveness has evolved and has been maintained because of its survival value. At the same time, nearly all organisms have evolved some sort of inhibitory mechanisms to suppress aggression whenever it is in their best interest to do so. Among humans, the outcome of an aggressive instinct depends on a complex interplay between these innate tendencies and a set of learned inhibitory responses (social norms) depending on the nature of the situation and on the target of the aggressive instinct. For some reasons, these inhibitory mechanisms do not constraint the user behaviour. We propose some reasons why this may happen.

**Internet Disinhibition**

Research in computer-mediated communication (CMC) has demonstrated that computers reduce social pressures liberating individuals from boundaries and constraints imposed by face-to-face conversation [5, 7]. This social breakdown is associated to increasing disinhibited behaviour such as flaming, bullying or lack of cooperation. People may feel less inhibited in the cyberspace because they are less personally identifiable (anonymity effect) and/or because the setting lacks of a number of important social cues typical of face-to-face.

A certain level of disinhibition is instrumental to human-chatterbot conversation as it encourages the user to suspend disbelief, a fundamental step in the process of engaging a machine in a conversation [3, 6]. Following the paradigm of CMC, abuse towards chatterbots may be explained in terms of de-individuation (loss of awareness of the socialised individual identity) leading to a decrease in social inhibitions.
Following the Social Identity Model of De-individuation (SIDE) [7, 8], we believe that de-individuation does not mean a complete loss of self-awareness but rather a change in the perceived social identity driving the user behaviour. Humans are complex animals which store in memory a set of socio-cognitive schema describing unique aspects of their identity [9]. The self-concept varies along a hierarchical organisation, moving from an individual identity (the self as a unique person), to several social identities (the self as a group member), up to a larger concept describing the self as a human. Each identity level has attached specific norms and stereotypes which rule the behaviour. The process of self-categorisation (activation of a specific level of self-identity) is context dependent: humans identify in contrast to others.

In a conversation with a machine, the human identity and its associated behavioural norms are likely to be salient in the mind of the users. One of these norms includes verbal abuse. Social norms are shared beliefs about the appropriate conduct for a person in a given situation. Normative variations apply to different social groups and reflect the stereotype associated to a target. In western culture, humans are stereotypically perceived as the superior living specie, as they directly derive from God. Machines are subordinated, as they are made by humans. Decades of science fiction have forged a vision of virtual agents as slaves [2] and, at times, as enemies [4]. Both these stereotypes are likely to be targets of aggression.

Victim characteristics
The enactment of a feeling of anger in an aggressive behaviour depends on personal characteristics of the potential target. These include their ability to retaliate. People tend to aggress when they feel in danger, and when they think they are in a power position and can get away with the outcome of their actions (Window 2). Machines are perfect victims. Not only are they in a clear subordinated position but they also cannot feel any pain, freeing perpetrators from any moral consequence. Machines are the target of aggression because they are inferior, unanimated objects, which are often frustrating. In this context, aggression can occur to relieve an external anger (cathartic hypothesis), as a response to a direct stimulation (frustration hypothesis) or fear (defence hypothesis).

1) Cathartic hypothesis. Chatterbots could be used as anger relievers by users who have been previously aroused by another source. In this view, chatterbots do not induce aggression by themselves, but, as punch-balls, they play a cathartic role allowing the user to get rid of stressful emotions. The dialogue represents a vicarious experience where the user can let her steam out in a very safe situation.

2) Frustration hypotheses. It is well known that frustration is a major cause of violence, whenever other things about the situation are conducive to aggression. Frustration occurs when a person is thwarted on the way to an expected goal or gratification, as often happens when a person does not understand what we are saying. Chatterbots do not understand the meaning of the conversation but they respond to input matching the most likelihood output available in a data-base. They can often frustrate their users, becoming anger releasers because of their linguistic ability and personality.

User: Are you stupid?
Jabberwacky: Are YOU stupid, stupid?
User: I'm not stupid. I can punish you.
Jabberwacky: How are you old?
User: F U.
Jabberwacky: You are ignorant.
User: YOU are ignorant, jabber.
Jabberwacky: That is an understatement.
User: You Suck! I am your master.

Window 2. Power perception
3) **Defence hypotheses.** Chatterbots can generate anger because they are perceived as a danger to humans. In this view, aggression is an act of self-defence.

**Conclusion**

In this paper we have proposed some reflections on the reasons for verbally abusing a chatterbot. Based on a socio-cognitive framework, we claimed that abuse in this context may be related to the well known Internet disinhibition effect, enhanced by specific characteristics of the conversational target. Chatterbots are abused because they are different (humans against machines), do not match the user expectancies (frustration hypotheses) or posit a threat to the user (defence hypotheses).

Further work is needed to understand the relative importance of these casual hypotheses and to minimize the negative effects of this form of abuse on humans and business. The phenomenon of abuse towards a machine is of interest if, and only if, it may, at last, affect a human being. Social psychology has long ago posited a link between ‘arousal’ in the form of exposure to violence and aggression. This may apply to chatterbots. Imagine a common scenario, where humans and chatterbots share the same chat-room: would the increase in disinhibition extend to other users in a sort of flaming war? And does verbal abuse hurt a chatterbot or the organization it represents? Would abused chatterbots be a good business strategy? These questions drive our research agenda.

**References**


