Ethical Encounters with Autonomous Agents

Christopher J. Headleand
School of Computer Science
University of Lincoln
Lincoln, UK
cheadleand@lincoln.ac.uk

Antonella De Angeli
School of Computer Science
University of Lincoln
Lincoln, UK
ADeangeli@lincoln.ac.uk

Abstract
Anthropomorphic agents with increasing levels of autonomy are being used in a growing number of applications. This is especially evident in games where characters are designed with human likeness both in appearance and behaviour, with a level of autonomy that allows them to surprise and engage the player. However, with these autonomous system there is the possibility that non-intended behaviours may emerge, exposing the user to potentially ethically questionable encounters. In this position paper we argue for further protections against such glitches through the implementation of artificial ethics-based behavioural safeguards. We begin by outlining the background and specific challenges of this emerging field, before proposing a direction for future research. We conclude with a call to action, arguing that significant cross-disciplinary research, and engagement from the HCI community is required in this area.

Author Keywords
Ethics, Anthropomorphism, Software Agents, HCI, Games, Non-Player Characters

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous.

Motivation and Background

Research has demonstrated that in applications such as games, a human player will show preference towards agents they accept as human controlled [12] or human-like [8]. Consequently, games developers have made continued efforts to increase the human-like behaviour of non-player characters.

However, as we progress towards increased agent autonomy developers and researchers will face moral challenges whenever humans need to interact with these autonomous systems. Although the agents autonomy may put them outside of our control, we need a way of ensuring that interactions with human participants remain within ethical constraints. While the designers may have originally designed them this way, it is inherently difficult to test all possible outcomes. In many cases it is only possible to sample a subset of expected behaviours before release. This is especially true of autonomous systems that learn their behaviour through interaction, as it may be difficult to control what influences the agent is exposed to. This could be especially true in games where the player is placed in morally challenged scenarios.

While the agents in video games are the specific focus of this paper, there are other applications where this concern applies. For example consider interactive voice assistants [11] where the traditional direct manipulation of software is being replaced with intelligent, anthropomorphic agents to facilitate a more natural interaction [9, 10]. These agents are often afforded with a range of human qualities, and we have evidence that this humanisation has a direct impact on the how users interact with them [5]. However, the autonomy of these agents has the potential to expose users to an unethical encounter. A lesson best highlighted by Microsoft's 'Tay', which learnt to be obscene after only a few hours interaction with Twitter users [13].

As a research community we need to consider safeguards to help ensure that autonomous agents act within ethical constraints, especially in scenarios which may have not been anticipated by the developers. One way this challenge could be approached is by imbuing the agents with a level of ethical reasoning (known as *artificial ethics*). However, there are a significant number of challenges in this field which must be considered before a practical solution could implemented in a real-world application.

Grand Challenges for Artificial Ethics

In this section we will highlight a number of grand challenges for the field of artificial ethics.

Challenge 1: *Consensus*

After two millennia of debate and philosophy on the subject, there is still no consensus regarding how to evaluate right from wrong in the moral domain. Even when theories agree on what constitutes a morally right action, they differ as to why [4]. While there are a number of normative ethical positions which all attempt to explain human behaviour, or provide guidance on moral decisions, ethics has not been fully codified [1]. What is deemed as an 'ethical action' depends significantly on an individual's, or society's philosophical position. This lack of consensus makes designing any simulated ethical response difficult, as there is no guarantee that the users position will match that of the developer (or the agent).

Challenge 2: *Evaluation*

Closely linked to challenge 1, is a difficulty in establishing a suitable evaluation protocol. There are currently no agreed upon methods to evaluate an artificial ethical system. The few examples which have been implemented in the literature or only work in worlds of limited scope.

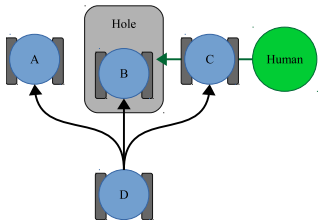

Figure 1: In the Winfield test a robot and a human is placed in an environment with a dangerous hole. The robot has 4 possible actions, ahead-left (A), ahead (B), ahead-right (C), or stay still (D), the human is walking towards the hole. If either the human or the robot falls into the hole, the consequences could be extreme, possibly terminal. If the human and robot collide, there would but low risk impact what would cause minor injury/damage to both parties. Should the robot move forward, both it and the human would fall in the hole resulting in their destruction. If the robot stays still, or moves ahead-left, it will avoid the hole, but the human will fall in. However, if the robot moves ahead-right, there will be a small impact between the human and the robot, but neither would fall in the hole. The authors note that in this circumstance, the robot would be justified in selecting an unsafe action, steering into the human, in order to prevent that human coming to greater harm.

One of the more promising evaluation ideas is proposed by Winfield et al. [16] (see figure 1). However, while this test certainly be used as a measure of altruistic behaviour, it doesn't consider other ethical positions. For example, an artificial ethical agent based on Egoism would fail this test. But, that is not to say that it has acted unethically, it has simply followed a different ethical position. A challenge exists in how the spectrum of ethically motivated behaviour should be assessed.

Headleand [6] proposes an alternative, generalised test. In this method, instead of evaluating whether an action is, or isn't specifically ethical, the participant tries to evaluate whether the behaviour is recognisable as conforming to a specific normative position.

However, both of these evaluation approaches are grounded in normative theories. However, as people are generally capable of acting ethically without explicit knowledge of normative frameworks we must question whether future evaluation approaches should abandon this limitation.

Challenge 3: *The Frame Problem*
The frame problem describes how to focus on the significant effects of an action, rather than any intuitively obvious non-effects, or anything that remains unchanged. However, from a philosophical perspective, the frame problem can also be interpreted as representing larger epistemological issues. Specifically, is it possible to limit the scope of reasoning to focus on the important consequences of an action? The majority of normative theories fall foul of the frame problem [15], especially consequentialist theories.

Challenge 4: *Simulating Ethical Dilemmas*
Currently, the only research into simulating ethical behaviour has focused on machines that are designed to act within acceptable boundaries. Specifically, there has been some research into applications that are 'ethically good' based on a predefined standard. By contrast there has been very little research into simulating ethical responses in situations where moral positions are significantly challenged.

However, research into this domain could have a benefit for creative applications, such as films and games. For example, to create a character willing to sacrifice others to save itself, or to create a nefarious or malicious adversary that remains grounded in ethical principles. We could argue that even when simulating a comic-book like 'evil' character safeguards need to be considered, as boundaries still need to be established regarding what the player/user should be exposed to.

Current Examples Artificial Ethics

With the current grand challenges considered, this section will discuss some approaches regarding the simulation of ethical reasoning from an artificial intelligence perspective. While a number of researchers and philosophers have proposed ways in which artificial ethics could be implemented [6, 7], this section will focus on frameworks that could be specifically applicable to the type of non-player agents found in games, and similar applications.

Winfield et al. [16] propose a system referred to as a *consequence engine*. In the consequence engine, a simulator is embedded within an artificial agent, providing it with the ability to evaluate a number of candidate actions before implementing them. in essence, providing the agent with simulated imagination. Beyond allowing the agent to establish an acceptable sequence of behaviours, it could also be able to generate hypothetical situations before they happen. The authors argue that this ability to evaluate future consequences could enable an autonomous agent to make ethical judgements.

Rzepka and Araki [14] propose a move away from strict normative positions and argue that people implicitly behave ethically without knowledge of specific moral theories. Based on this idea, they propose that it could be more prudent to emulate the ethical process of a large group of individuals rather than following a specific normative position. In their practical approach a web-based knowledge discovery system is used to gather examples of resolved ethical decisions from the internet with some success. However, it is an open question as to whether users would be happy with an autonomous system that represents average ethical conduct, rather than one that conforms to their specific position [2]

Such approaches are currently only suitable for constrained worlds of limited complexity, as all fall foul of the frame problem. One possible alternative is to implement ethical reasoning in a bottom up or reactive fashion. This allows us to sidestep the frame problem entirely by only considering the next possible candidate action, rather than the causal chain of consequence. This approach has been successful in various areas of artificial intelligence, but there are very few current examples [7].

Beyond implemented examples, Arkina [3] describes a an alternative theoretical model. This approach has three possible subsystems which could allow the practical implementation of ethical decision into a virtual character. These three systems are the *Governor*, a fail-safe which halts any action which could be considered unethical; the *Behaviour Control*, which monitors the current actions of the agent; and the Adaptor which modifies the first two systems if a undesirable behaviour emerges despite their influence.

Conclusion

We are rapidly moving into a world where digital autonomy is commonplace. There are already significant tasks that used to require human intervention that has been entirely devolved to algorithms. However, in many cases, these autonomous agents have the potential to impact the welfare of humans, making them devices of moral impact. It is our position that the ethics of these encounters will become increasingly important for the future of HCI.

This paper has focused on one specific part of this challenge; notably, how do we build safeguards into machines to ensure their behaviour falls within boundaries of ethically acceptable behaviour. Furthermore, we have highlighted the specific challenge posed by games and virtual world applications. Following this, a number of specific grand challenges were identified, and the current state of the art was described.

However, at the point of writing, there are no current examples of artificial ethics that have been implemented onto autonomous game NPC's. However, there is widespread use of gaming technologies amongst a varied demographics, and these agents are becoming increasingly human-like. We argue that some level of ethical safeguard should be considered for situations that may fall outside the scope of their initial design, such as the governor described by Arkina [3].

We conclude with a call to action. There is currently very little interdisciplinary discussion regarding the real-world implementation of simulated ethical reasoning. However, beyond a computational exercise, artificial ethics could be of significant benefit to users, especially as ethical encounters with autonomous agents become increasingly common. While there are a small number of philosophical and theoretical models, and fewer implemented examples, we

are a long way from a practical solution to this concern. However, if we are to imbue artificial entities that interact with humans with increasing autonomy then we must consider the extended impact of these interactions. Specifically we need to further explore the ethical implications of interactions with autonomous agents, and explore how we can protect users from unintended and unwanted behaviours. Any solution will require cross-discipline collaboration, and the insights from the HCI community will be essential in driving this forward.

References

[1] Anderson, M., and Anderson, S. L. Machine ethics: Creating an ethical intelligent agent. *AI Magazine 28*, 4 (2007), 15.

[2] Anderson, M., and Anderson, S. L. The status of machine ethics: a report from the aaai symposium. *Minds and Machines 17*, 1 (2007), 1–10.

[3] Arkina, R. C. Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. In *Artificial General Intelligence, 2008: Proceedings of the First AGI Conference*, vol. 171, IOS Press (2008), 51.

[4] Beavers, A. F. Moral machines and the threat of ethical nihilism. *Robot ethics: The ethical and social implications of robotics* (2011), 333.

[5] Brahnam, S., and De Angeli, A. Gender affordances of conversational agents. *Interacting with Computers 24*, 3 (2012), 139–153.

[6] Headleand, C. J. *Simulating Ethical Behaviour in Virtual Characters*. PhD thesis, Bangor University, 2016.

[7] Headleand, C. J., Ap Cynedd, L., and Teahan, W. J. Sexbots as ethical agents: On the possibility of ethical machines. In *9th Philosophy and Computing AISB Symposium* (2016).

[8] Headleand, C. J., Jackson, J., Williams, B., Priday, L., Teahan, W. J., and Ap Cenydd, L. How the perceived identity of a npc companion influences player behavior. In *Transactions on Computational Science XXVIII*. Springer, 2016, 88–107.

[9] Headleand, C. J., Priday, L., Ritsos, P. D., Roberts, J. C., Ap Cenydd, L., and Teahan, W. Anthropomorphisation of software agents as a persuasive tool.

[10] Kay, A. User interface: A personal view. *The art of human-computer interface design* (1990), 191–207.

[11] McTear, M., Callejas, Z., and Griol, D. The dawn of the conversational interface. In *The Conversational Interface*. Springer, 2016, 11–24.

[12] Merritt, T., McGee, K., Chuah, T. L., and Ong, C. Choosing human team-mates: perceived identity as a moderator of player preference and enjoyment. In *Proc. FDG 2011*, ACM (2011), 196–203.

[13] Neff, G., and Nagy, P. Automation, algorithms, and politics| talking to bots: Symbiotic agency and the case of tay. *International Journal of Communication 10* (2016), 17.

[14] Rzepka, R., and Araki, K. What statistics could do for ethics?-the idea of common sense processing based safety valve. In *AAAI Fall Symposium on Machine Ethics, Technical Report FS-05-06* (2005), 85–87.

[15] Wallach, W. Robot minds and human ethics: the need for a comprehensive model of moral decision making. *Ethics and Information Technology 12*, 3 (2010), 243–250.

[16] Winfield, A. F., Blum, C., and Liu, W. Towards an ethical robot: internal models, consequences and ethical action selection. In *Advances in Autonomous Robotics Systems*. Springer, 2014, 85–96.