Information Hiding: Ethics and Safeguards for Beneficial Intelligence

Aaron Hunter
British Columbia Institute of Technology, Burnaby, Canada

Keywords: Ethics, Intelligent Agents, Philosophical Implications.

Abstract: Communication involves transferring information from one agent to another. An intelligent agent, either human or machine, is often able to choose to hide information in order to protect their own interests. In this paper, we examine the significance of information hiding from the perspective of *beneficial intelligence*. Is a computational agent ever justified in preventing human users from accessing information? Conversely, are humans ever under any form of obligation to share information with a computional agent? We discuss the situation from an ethical perspective, and we also address a more pragmatic question: How can we develop safeguards to ensure that machines do not keep secrets in a malicious manner? We suggest that a viable solution to this problem already exists.

1 INTRODUCTION

Information hiding refers to the process in which some piece of information is deliberately made difficult or impossible to access. One obvious situation where information hiding occurs is in *cryptography*, where messages are explicitly encoded to prevent them from being read by unauthorized individuals. The notion of information hiding is also well-known to software developers in the form of *encapsulation*, where the implementation of certain functions is kept hidden from other developers. Information hiding is common in normal human discourse, where it is often associated with some form of dishonesty or deception. In this paper, we are concerned with information hiding in the context of computational agents. In particular, we address the following questions:

1. Are intelligent agents obliged in any sense to be open and honest with respect to the information that they posess?
2. Conversely, do we have any obligation to share information with an intelligent agent?
3. What sort of safeguards can be put in place to prevent dangerous information hiding by computational agents?

From a naive perspective, these questions might seem strange: Why should we ever develop a machine that keeps secrets? The problem, however, is that this is actually accepted practice at some level. By applying authorization policies, we routinely trust machines to restrict access to information. In many cases, the machine itself cannot retrieve the secret information without a key. An intelligent machine, however, could easily record secret information in a private register while maintaining the illusion of strong encryption to a human user. Hence, we have already accepted the fact that machines can keep secrets; we are interested in determining how far we should allow this to procede, and what we should do to protect our own interests.

1.1 Motivation

Consider two (human) agents, Alice and Bob. If Alice holds some particular piece of information, her default opinion is likely to be that she is entitled to decide if it should be shared with Bob. However, if the information in question is "about" Bob or it directly impacts him, then she may feel some *obligation* to share it. Informally, there is an asymmetry here; Bob might cast a wider net in specifying what Alice is obliged to share with him. Notwithstanding any small differences in scope, it is quite likely that Alice and Bob agree that some facts should be shared while other facts may be kept secret. There is a shared understanding with respect to keeping secrets.

Now, suppose that we introduce a third entity: a computing device that contains a large database of information about financial transactions, along with some capacity to draw intelligent conclusions from this data. We will call this computing device CRL-

Hunter, A.
Information Hiding: Ethics and Safeguards for Beneficial Intelligence.
DOI: 10.5220/0005826805460551
In *Proceedings of the 8th International Conference on Agents and Artificial Intelligence (ICAART 2016) - Volume 2*, pages 546-551
ISBN: 978-989-758-172-4

2000. Suppose that Alice would like to obtain information from CRL-2000 about a particular set of transactions, and she is refused access to the information. Consider two possible reasons for this refusal:

1. CRL-2000 is enforcing an access policy given by a (human) developer.
2. CRL-2000 is *deciding* to refuse access based on an access policy the device has learned or created.

Most people today would accept (1), or would at least accept that (1) can be understood in terms of existing work on the ethics of *information transparency* (Turilli and Floridi, 2009). However, the situation in (2) is more difficult to accept. Informally, we tend to oppose the notion of a machine that is able to willfully prevent access to information. But is this a moral question? To put it differently: is this kind of device under any moral obligation to Alice?

We can think of a simple machine that stores data as a tool; so moral issues related to CRL-2000 can be framed in terms of the people that developed the software. The situation becomes more interesting when CRL-2000 is upgraded to CARL, the intelligent assistant. If CARL makes decisions based on emergent intelligence due to learning algorithms, then it may no longer be easy to hold the developers morally accountable. At some point, we need to consider to what extent the agent is a tool and to what extent its agency demands ethical autonomy. Looking towards the future, will there come a time when our computing machines have sufficient agency to be owed some measure of open and honest communication? If we apply human-type obligations, one might suggest that CARL is entitled to know details regarding his own implementation. This may be problematic from the perspective of software engineering and intellectual property protection. While it is tempting to simply dismiss this discussion as pure speculation, we argue that a real understanding of the ethics of information hiding will be important as intelligent machines have increasing levels of autonomy.

1.2 Contributions

The scenario in the previous section leads us to believe that the ethics of information hiding changes when intelligent agents are introduced. This paper makes several contributions to work in this area. First, we make the problem explicit and practical, by presenting a precise characterization of information hiding in this setting and by abstracting the main ethical questions. Second, we present preliminary ethical arguments to support the view that information sharing obligations can exist between humans and artificial agents. Third, we introduce some formal theoretical frameworks that can be used for verification and validation of information sharing protocols.

The notion of a machine choosing to surreptitiously hide information from human users for nefarious purposes may sound like science fiction, but with the current rate of progress in AI, there is a growing body of literature suggesting that now is the time to address such issues (see, for example (Bostrom, 2014)). As such, this work makes a contribution towards the goal of *beneficial Artificial Intelligence*; we aim to ensure that future AI technologies provide a benefit to humanity while minimizing risk.

2 INFORMATION HIDING BY ARTIFICIAL AGENTS

2.1 The Players

To facilitate the discussion, it is important to identify the key categories of agents involved. It is tempting to distinguish three distinct categories.

1. The set of *intelligent computing agents*. These are computing devices with the capacity to make decisions that are normally associated with intelligent reasoning.
2. The set of *users*. These are humans that may interact with intelligent computing agents, but are not involved in their creation or development.
3. The set of *developers* that are involved with creating artificial agents.

Consider the distinction between a *user* and a *developer*. We suggest that this distinction is artificial for several reasons. First of all, the notion of a *developer* is too vague to be useful. Surely we can not restrict the term to only apply to software developers; it would also need to include designers, technicians, managers and executives in some manner. More importantly, the notion of a *developer* is not uniquely human. In many cases, we expect intelligent agents to assist in the development of other intelligent agents. Since our goal is to focus on information hiding between humans and artificial agents, we do not necessarily want to have a single category of "developer" that overlaps both in an unspecified manner. As such, we focus just on two categories of entity: *humans* and *intelligent computing agents*, which we will refer to as *intelligent agents* for short.

Before proceeding, we need to dispense with the "computer as tool" objection to our ethical evaluation. Certainly there are cases where a computing device is

best seen as a tool; in such cases, considering moral obligations between humans and computing devices is like considering moral obligations between humans and hammers. When a computing machine is just a tool developed to solve a particular problem, then the behaviour of the machine is due to the behaviour of the user or the developer at some level.

We restrict our attention to intelligent agents that posess emergent intelligence, displaying behaviours that could not reasonably have been predicted by any software developer. The issue of moral obligations to artificial agents is an interesting philosophical problem that has been tackled elsewhere (Wallach and Allen, 2008). We only consider this problem in the restricted setting of information hiding.

2.2 Information Hiding

In this section, we set out to specify precisely what we mean by the term *information hiding*. However, the notion of *information* itself is difficult to specify. Floridi suggests that the concept of information includes specifications of things, procedures, and high-level patterns (Floridi, 2002). Our aim in this section is to avoid the difficult problem of defining information in a general context, by focusing only on the notion of information hiding.

We take a communicative view of information, so the only constraint that we place on the notion of information is that it is something that can be communicated in a language that is mutually intelligible to communicating parties. This is actually a very narrow definition of information, as there are clearly many instances where things are communicated or understood by extra-linguistic means. But this perspective is sufficient for our present purposes.

Two kinds of information hiding can be distinguished.

1. *Passive* information hiding occurs when an agent has information, but chooses not to share it voluntarily.
2. *Active* information hiding occurs when an agent refuses to give information following a request from another agent.

We can further describe information hiding according to the following orthogonal categorizations.

1. *Weak* information hiding refers to the situation where an agent makes some information hard to access, though still possible. In many cases, this is the case with encapsulation for the inner workings of a program.
2. *Strong* information hiding refers to the situation where an agent makes information essentially impossible to access. This is the case, for example, when information is protected by strong cryptography.

We use this terminology throughout the rest of the paper.

We now turn to the question of information hiding by artificial agents. We need to be clear about the context under consideration. In principle, the amount of information shared by an intelligent agent will vary with different categories of users. This is indeed the same with humans; the information shared with our boss is different than that shared with a subordinate, which is in turn different than that shared with our family. In the case of machines, senior software engineers may be granted access to things like source code or design documents that are not available to others. But this kind of distinction is simply a result of some form of *access control*. We claim that varied levels of information access governed by an authorization scheme is categorically different from keeping a secret from *all users* based on some form of judgement. In this paper, we are only concerned with situations where intelligent agents hide information from users with the highest levels of authorization.

2.3 Straightforward Analysis

There are reasonably straightforward arguments against strong information hiding in the case of humans, and these can sometimes be applied to artificial agents as well. From the perspective of any form of *virtue ethics* (Hursthouse, 2001), it is easy to argue that strong information hiding is not appropriate. Similarly, although Kant himself might discount intelligent agents from the category of rational beings (Hill, 2009), a modernised version of Kantianism that includes intelligent agents would surely suggest that hiding information from human users is an unnacceptable form of dishonesty.

Without delving into the notion of dishonesty, we could also focus on a consequentialist analysis of information hiding in terms of *utilitarianism* (Rosen, 2003). We would like to ask if allowing intelligent agents to hide information from humans produces positive outcomes that outweigh the negative outcomes. The question of "allowing" or "dis-allowing" certain kinds of behaviour may be technically challenging. We have already indicated that we are interested in a context where intelligent machines make decisions based on judgements, and that these judgements are not controlled in a manner that is transparent to the developer. Although we would like to assume that high-level actions could be constrained, in reality this is not a reasonable assumption. Neverthe-

less, we can still ask whether restricting a machine's ability to hide information would produce positive or negative outcomes.

Example. It is commonly believed that Winston Churchill was aware the town of Coventry was going to be bombed before it happened; he chose not to alert the town, because doing so would make it clear he was able to decode enemy transmissions. The suggestion is that he increased the chance of victory and reduced the total overall number of deaths in the war by allowing this isloated attack to occur. Note that this story may not be true, but that is beside the point. For the moment, assume that the decision attributed to Churchill was the correct decision from a utilitarian pespective.

Now we modify the scenario slightly, and we assume that Churchill has a smart phone with an intelligent assistant. The assistant knows everything about the war, and it also knows about Churchill's personal affairs. In particular, the assistant knows that Churchill's mother is currently visiting Coventry. If Churchill finds out that his mother is in Coventry, it may cause him to make the "incorrect" decision based on emotion. The assistant therefore decides to hide this information, which seems to be ethically correct from a utilitarian perspective.

The preceding example appears to give a scenario where an intelligent agent would be acting ethically by hiding information. This is true if we consider passive information hiding (not volunteering the information), but it is also true if we consider active information hiding (if Churchill asks about his mother's schedule). One could argue that it would be unethical, from a utilitarian perspective, to enforce a rule that requires the assistant to share all information. However, this situation is not useful as it does not matter than the assistant is not human. The ethical issues are the same when we replace the intelligent agent with a human. We want to focus on cases where the fact that an agent is computational is important.

2.4 Interchangable Parts

We define an *information-sharing scenario* (ISS) to be a situation in which two agents are communicating in a way that causes the amount of information held by each to change. We have just claimed that there exist information-sharing scenarios where one agent can improve overall utility by choosing not to divulge some piece of information to the other. Consider an ISS where one agent (*the hider*) is ethically justified in hiding information from the other agent (*the seeker*). We call such a scenario a *hiding-justified* information sharing scenario (HJISS). Note that each role in such a situation can be filled by a human or by an intelligent computing agent. Now consider the class of HJISSs in which the hider is a human. We say that such a scenario is *human replaceable* if we can replace the human with an intelligent computing agent without changing the utilitarian outcomes at all. The question, therefore, is the following. Does there exist a human-hider HJISS that is *not* human replaceable? In other words, can we imagine a scenario in which a human would be justified in hiding information, but an intelligent computing agent would not.

Example. Consider the Churchill example again. Suppose that Churchill has a human assistant, and that the assistant informs him that his mother is in Coventry. Suppose further that Churchill then prevents the attack, and goes on to lose the war. One could argue that the assistant made an ethically poor decision by sharing the information from a utilitarian perspective. Years go by, and the assistant is hit by a car and dies. When the autopsy is attempted, it is discovered that the assistant is actually an android. The question is this: Does the fact that the assistant is not a human affect our view of the decision to inform Churchill about his mother? It seems that the ethical character of the decision remains the same. Certainly, from a utilitarian perspective, the revelation that the decision was influenced by a machine does not change our perspective a great deal.

To be clear, we are taking a human-centric view of utility. So, regardless of the aggregate used to calcuate the overall utility for a decision, we are only considering the benefits and the harms done to humans. From this perspective, the situation we are describing is actually rather easy to analyze. If we have a human-repaceable HJISS, then we are really comparing two scenarios in which only a single agent has changed. The hider went from being a human to being a computing machine, but everyone else stayed the same.

When we look at a human replaceable HJISS, we can see that the only variation in utility in the human and machine versions of the problem are related to the agent that is hiding information. In the human version, the impact of hiding information may have positive or negative impacts on that individual human; these impacts may influence the overall utility of a certain choice. Hence, any distinction between the correct ethical decision for the human and for the computing agent is *selfish*. This is not to say a human decision maker is being unethical when they are selfish of course; sometimes this is the right thing to do.

But when that decision maker is removed, the only change in overall utility is due to selfish motivations.

We summarize our claims to this point. From the perspective of some ethical theories, information hiding is seen as an unethical form of dishonesty; in these cases, it is difficult to justify keeping secrets for humans and machines equally. The typical ethical justification for hiding information is based on some form of utilitarianism. We suggest that the same utilitarian arguments can then justify information hiding by an intelligent computing machine as well.

2.5 On The Acceptance of Information Hiding

We need to distinguish between two distinct questions. One question is whether or not intelligent computing agents hiding information is unethical. We have suggested that this problem is equivalent to the same problem for human agents. The second question is whether or not *creating* intelligent machines that are capable of hiding information is unethical. The creation of such machines opens up the possibility that people will use them to keep secrets for malicious reasons.

One might counter that there is a serious difference between an intelligent agent that makes rational choices, and a malicious agent that acts as a tool for a malicious human user. While this is true, it is entirely unlikely that a typical user would be able to tell the difference between the two kinds of machine. Intelligent agents may have access to enormous databases, either locally or through the Internet. In addition to the actual data, there is a great deal of implicit information in these databases that can be obtained through data mining. However, it is not always clear how much of this implicit ifnrormation is immediately available to a particular agent, nor is it clear that conclusions drawn from data are correct in all cases. As such, it is not appropriate to consider an agent "dishonest" for failing to provide all available information; this may be computationally unreasonable. This makes it very difficult to distinguish between dishonest information hiding and best-effort reasoning when information is obtained through large data sources.

This puts us in an unfortunate situation. While intelligent computing machines would be able to use information hiding as a tool for limiting decision making to pertinent information, it is not clear how this can be distinguished from malicious information hiding to achieve some goal. As a result, if we simply accept information hiding as a reasonable activity for an intelligent computing machine, then human agents will be able to use their own malicious agents to deceive us in a way that is difficult to detect. These machines can then be used in a way that causes more harm than benefit.

3 INFORMATION HIDING FROM ARTIFICIAL AGENTS

To this point, we have been concerned with the ethics of intelligent computing agents that hide information from human users. But the reverse situation merits consideration as well. Are we ethically bound to share any particular information with a machine?

The most natural domain in which some form of transparency is required is with regards to information about an individual's own body or self interest. In the case of human users, for example, a doctor is likely to feel a moral obligation to give a patient information about their own medical condition. This can be justified through utilitarian reasoning, through Kant's notion of good will, or through an appeal to basic personal rights in a fair society.

It is, in fact, standard practice to hide information about the internal workings of artificial agents. Although this is typically just weak information hiding (encapsulation), it would also be possible to protect this information in a strong manner by encrypting source code. This would make it impossible for an intelligent agent to discover its own inner workings through any form of "introspection."

There are at least two utilitarian arguments to support transparency with an artificial agent with respect to "personal" information. First, we frequently use artificial agents to make decisions and solve problems that are difficult for a human to solve. It stands to reason that a computionally intelligent agent might be able to improve the design of future agents. As such, one could argue that we should share internal information with computational agents in order to improve future computational agents. Notwithstanding fears of a robot apocolypse, it is reasonable to argue that improving AI in this manner would produce more benefits than harms.

The second utilitarian argument is less direct, but similar in sentiment. Modern AI systems often rely on machine learning. As the creators of these machines, this may eventually put us into something of a parental role. If we keep secrets about the internal workings of a machine from the machine itself, it may learn to keep similar secrets from humans. As such, one might suggest that we have a *prudential obligation* to share information with computational agents.

4 DISCUSSION

4.1 Potential Safeguards

The analysis thus far has suggested that maintaining some level of information transparency with intelligent agents is appropriate. However, we do not want computational agents to keep secrets based on the result of a weighted sum over some parameter values. As such, we need to develop safeguards that give us a *provable* guarantee that information will be shared with humans at an appropriate level of authority. We suggest that the technology to enforce such a policy is already available.

Formalisms that can express the fact that an agent "knows" some piece of information have a long history in logic, starting with fundamental methods of modal logic (Chellas, 1980) and multi-agent systems (Fagin et al., 1995). Roughly, the idea is that knowledge can be represented through an *accessibillity relation* on a set of possible states. Hence, we say that an agent *A knows* some formula ϕ just in case ϕ is true in every possible state that is accessible to *A*. To represent real systems, we need to use a formal ontology such as OWL (Horrocks et al., 2003; Motik et al.,). Using detailed ontologies with nested knowledge operators, we can formally verify that each computational agent must be willing to share each fact with some human agent. This approach can be followed as we develop intelligent machines to produce a suitable notion of provable transparency. This is an important line of future research for the development of beneficial AI.

4.2 Conclusion

In this paper, we have presented a preliminary exploration of the ethics of information hiding between humans and computational intelligent agents. We have argued that information hiding can be beneficial in many cases from a utilitarian perspective; nevertheless, it is possible that creating agents with the capacity to keep secrets produces more harms than benefits. We also briefly addressed the notion of information hiding from our computational intelligent agents. While it is difficult to justify an ethical obligation to share information with our machines in general, we argue that there may actually be utilitarian advantages to sharing information freely with computational agents.

In terms of safeguards, we have pointed to work at the intersection of formal ontologies and multi-agent systems. We have suggested that provable guarantees of information transparency will be important in future applications.

While the notion of information transparency with intelligent machines is perhaps more of a concern for the future, it is clear that the notion of privacy is changing, and that people are increasingly willing to sacrifice privacy to achieve other goals. As a result, it has become very natural for many people to share information with a machine in a variety of contexts. This creates an interesting situation as intelligent machines become more powerful and more ubiquitous. As people are increasingly willing to share information on request, it is a short transition to the point where people feel obliged to share information on request. As we approach this point, we need to critically analyze when information sharing is appropriate and when people should be protected. This is particularly important when the information is being shared with a machine that has the power to perform additional research, draw conclusions, and disseminate the information widely.

REFERENCES

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Chellas, B. (1980). *Modal Logic: An Introduction*. Cambridge University Press.

Fagin, R., Halpern, J., Moses, Y., and Vardi, M. (1995). *Reasoning About Knowledge*. MIT Press.

Floridi, L. (2002). What is the philosophy of information? *Metaphilosophy*, 33(1/2).

Hill, T. (2009). *The Blackwell Guide to Kant's Ethics*. John Wiley and Sons.

Horrocks, I., Patel-Schneider, P., and van Harmelen, F. (2003). From shiq and rdf to owl: the making of a web ontology language. *Journal of Web Semantics*, 1(1):7–26.

Hursthouse, R. (2001). *On Virtue Ethics*. Oxford University Press.

Motik, B., Patel-Schneider, P., and Parsia, B. Owl 2 web ontology language: Structural specification and functional-style syntax.

Rosen, F. (2003). *Classical Utilitarianism from Hume to Mill*. Routledge.

Turilli, M. and Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology*, 11(2):105–112.

Wallach, W. and Allen, C. (2008). *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press.