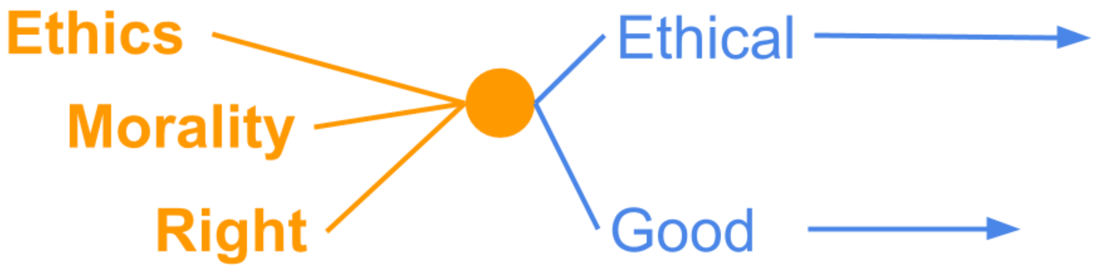
AI Ethics: Kind

We're involved in AI ethics with the right concept!

HeroX challenge: “How would you teach AI to be kind?” set by EthicsNet

Author: Dr Adam Bostock, [Innovation Future Specialist]()

Date: 18th August, 2018



Introduction

The Benefits

Artificial intelligence (AI) will become significant in many aspects of our lives, both directly and indirectly. So it is important that AI can be trusted to deliver good and fair outcomes for everyone. If we (the global community) get this right then ethical AI could enhance the quality of our lives, whatever our context.

AI can be trained to appear kind (e.g. speak politely), but we can also use ethical AI practices, and the solution presented here, to grow AI systems that are genuinely kind: speak politely, listen attentively, understand your needs, and take positive steps to help you and society as a whole.

(Independent of this project, the author has written a short *evolving* story that shows what the personal relationship between a person and their AI assistant could be like within a few decades: Education 2049: Looking forward to the great journey. We may come to regard our personal AI assistant as a close friend, and one that we rely on.)

The Challenge

Building an ethical AI system is a challenging proposition, due to the technical, human, and complex ethical challenges.

Ethics is not always black and white, or clearly defined, or objective; and people might favour a subjective outcome when it best suits their interests. Further, ethics might vary from country to country, or community to community.

This is further compounded by the problem that people might know what they “want” but are unaware of what they actually *need*. And let us be honest, sometimes, some people are just silly (e.g. doing something unfavourable for a laugh, facing addiction, self-harm, antagonistic or aggressive for no apparent reason, racist, and warmongering)! (Ethical AI and kind AI could help to solve these challenges too, via education, support, and personal coaching.)

Is the role of ethics to give people what they “want” - and ask for - or to give them what they *need*? An AI might become aware of solutions that give everyone what they *need* (but perhaps not what they “want”.)

Similar issues have been depicted in science fiction movies, where the AI unwisely takes over the control of society for our own “good” (as in iRobot); or where the actions of an AI probably lead to good outcomes, but rather than perceiving the AI’s good intentions we fear it (as in Transcendence).

Perhaps, no one person or system has the “solution”.

So the solution might be *a democratic approach of globally shared (dynamic) data and* *preferences*, and a *collaborative network of individual, diverse, AI entities*.

There are many perspectives and contexts, and that might mean good actions vary depending on those perspectives and contexts. We should seek out what they are, and try to understand them - a joint venture between humans and artificial intelligence.

Ethical Approach

A potentially successful approach to AI ethics and safety might be to adopt the following:

1. Establish **core safeguards** (e.g. do no “harm”, limit individual influence, and detect and ignore disruptive bots)
2. Allow the training datasets to **evolve** over time
3. Adopt a **probabilistic** approach, rather than absolute certainty
4. Be influenced by **consensus** (but with core safeguards enabled)
5. Allow **personal** datasets (to override aspects of the default consensus; but with safeguards enabled)
6. Support **contextual** aspects (e.g. country/community)
7. Offer **optimised recommendations** from the AI (derived from previous simulations, as described below...)
8. Provide a **simulation sandpit** for new datasets to be safely tested, and evolved: demonstrating the potential good and bad impact of each proposed training dataset; showing the changing impact when switching from a current dataset to a new proposed dataset; and so educating both AI and people
   [limit the detail of (future) advanced simulations, to avoid accidentally creating sentient beings as lab rats]
9. **Safeguards** should pause the adoption of apparently risky datasets and seek (authoritative) approval [e.g. avoiding the Tay chatbot and “Boaty McBoatface” embarrassments]

Being kind

So how do we get an AI to be kind: to be friendly, generous, considerate and helpful?

A kind AI might:

- use positive language
- offer useful advice and information that goes beyond a person’s direct requests to give them what they need (not merely what they ask for)
- tailor its response depending on the person’s context (professional / work setting; personal / leisure setting) and/or emotional state
- conduct actions that have a beneficial outcome

Implementation

We have to be mindful of the limited (but promising) abilities of today's AI technologies and the rapid pace of innovation towards much greater abilities, general intelligence, and/or super intelligence. This means any worthwhile approach has to support the limited abilities of AI today and its vastly superior abilities in the (not too distant) future.

We also have to be mindful of the benefits arising from a system that gives many people, globally, the opportunity to participate. This means providing a system that is *simple to understand*, and accessible via a wide range of devices and platforms. Two widely available technologies for doing this are the World Wide Web (and the wider Internet), and online social networks. A system that meets these requirements would facilitate easy, widespread, adoption. This inspired the following approach.

Concept Reference Model for xIntelligence

The term *xIntelligence (xI)* refers to multiple types of intelligence: human, machine learning, and more sophisticated levels of artificial intelligence (AI) as they arise. The following approach provides a mechanism for knowledge to be shared *simply* across these.

The Concept Reference Model for xIntelligence (CRMxI) is a way for anyone to share concepts (about anything) so that others can learn about those concepts, even AI systems. **The site in the above link provides details about CRMxI, along with a few simple examples, including Ethics.** CRMxI was created explicitly, by the author, for this project, but its scope can go far beyond ethics to include any concept.

Concepts can be quickly and easily defined in any text editor, web page editor, or social media tool.

A concept is defined in text, and it consists of one or more attributes that are associated with the concept or provide examples of the concept.

An attribute can be a reference to *any resource on the Internet*, via its URL (e.g. web page, social media post, image, video, audio, podcast, animation, data, instructions, API, program, service, or even an AI). This provides unlimited potential and future scalability!

There is also a "not reference" which means it refers to something that does not form part of the concept; something that is opposite to the concept. For example the Ethics concept has such a reference to *unethical*.

The Power of Distributed Knowledge

The above examples represent just one person's perspective on those concepts. Different individuals may, and probably will, express the same concepts in different ways (and with different amounts of effort and detail).

Organisations can be expected to focus on concepts that are relevant to them; and we can expect formal and detailed expressions of those concepts. These might form the basis of authoritative sources, which others can *reference* (and build upon).

Collectively, many individual and authoritative sources will provide a rich knowledge bank of concepts; for the benefit of all (whether human or AI).

Future expansion of CRMxI

The power of this approach is based on its *simplicity*, so that anyone and everyone can make use of it. Therefore, the associated philosophy suggests that it should always remain simple.

Note also that CRMxI is only intended to represent knowledge, it is not the intelligent agent (xI) reading and learning from this knowledge. (See: Intelligent Agents)

However, should the need arise, the definition of CRMxI allows for future expansion.

New keywords can be introduced. For example, the following have been envisaged:

- *Predict* - referencing an anticipated concept (or any Internet resource) that probably follows on from the concept in question (e.g. “rain” predicts: wet ground)
- *Action* - a reference to an appropriate action in response to detecting a scenario that matches the given concept (e.g. an appropriate helpful action)

A probability parameter might also be associated with the keywords (e.g. p=0.81). However, an AI system is likely to derive more accurate probability values by learning from multiple concept sources (and real world observations) on its own; so the need for this parameter is probably redundant.

Intelligent Agents

It is the role of intelligent agents (AI) to learn from the concept sources and to *act* according to:

- those concepts
- (other) training data
- sensory or input data
- context
- algorithms, and
- programming [currently, software instructions still play a very significant role in the development of AI systems]

In other words, CRMxI only represents one factor in how an AI will react.

CRMxI provides fundamental support (knowledge) to range of AI systems (and humans) of varying abilities, operating in a diverse range of contexts. Today, AI systems use a range of diverse programming and training approaches; and in the future we can expect additional approaches to emerge. If CRMxI was to attempt the integration of all these approaches then it would become a much more complex system, and that would defeat the objective of having a *simple system* that everyone can use quickly and easily.