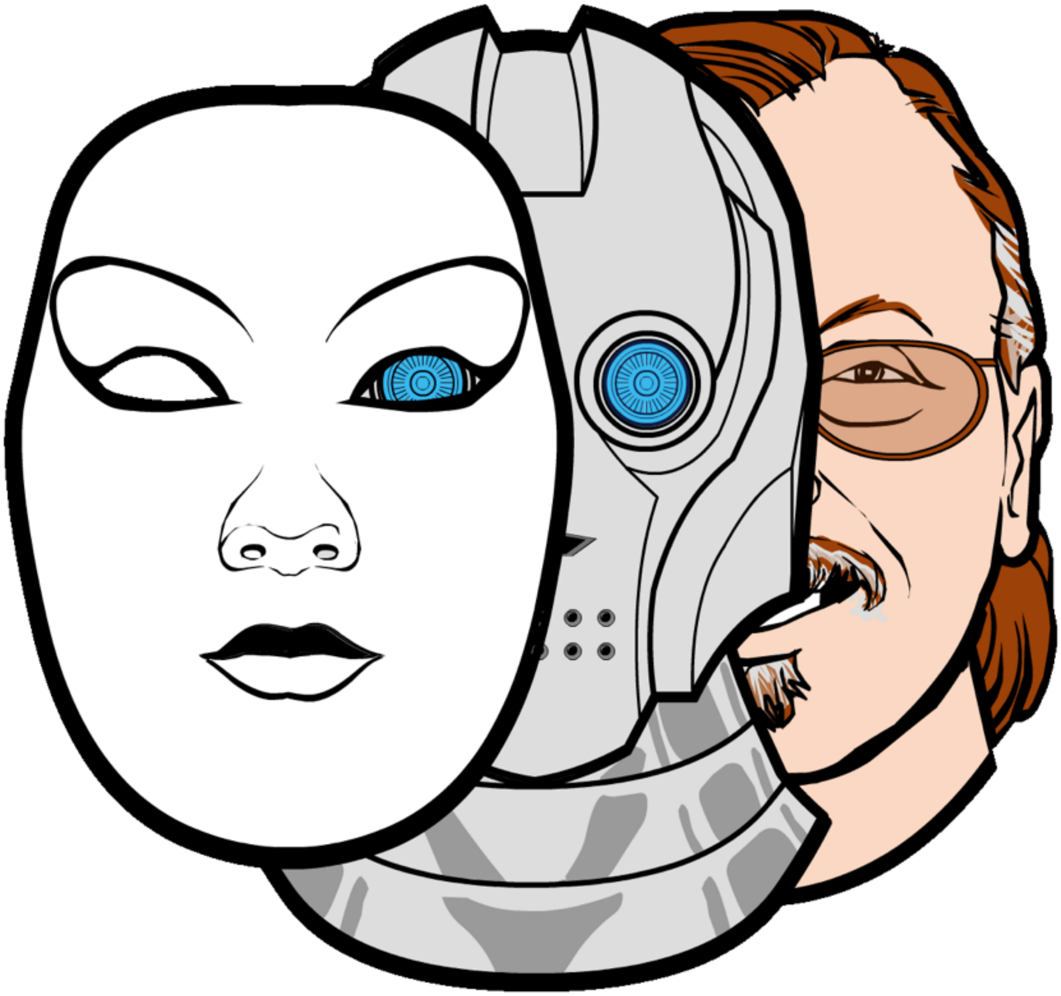
Personified Systems

Systems we relate to as if they were persons, and how they should behave



Jim Burrows

August 18, 2016, Rev. 1.0.1

Table of Contents

Trustworthy Personified Systems

> *What's a "trustworthy personified system"? Every day we see new systems emerging that we interact with as if they were people. We talk with them, take their advice; they read our gestures and facial expressions, help drive our cars, amuse our children and care for the aged and infirm. If they are going to integrate with our society and begin to act as if they are persons, we need them to behave as trusted members of society, not as outsiders, outliers or outlaws. How do we create that trust? That's what I hope to discuss here.*

Introductions

I'm a computer professional who prides himself in doing things that he's never done before. Thus after leading teams at two separate companies, doing basically the same thing (building secure and private communications systems), I became convinced that I need to shift my focus to something new. This caused me to try to figure out what the next big issue was likely to be, so that I could work on that. I concluded that it is creating well-behaved or trustworthy systems, especially "personified systems." Having said that, I should introduce a term or two, and explain why I think they point to something important.

I am using the term "personified systems" here to address a number of converging and interrelated trends. In simplest terms, what I am calling personified systems are computer- and network-based systems that are powerful and ubiquitous enough that we deal with them not so much as computers or even as tools, but as if they were people or beings that interact with and assist us.

There are at least three factors that contribute to us regarding them as "personified systems": how they present themselves to us, how much they "know" and the jobs that they do. In terms of presentation and interaction, they may talk with us, respond to text commands in more or less natural language, or watch us and our gestures. Systems with voice recognition like Siri, Alexa, Cortana, "Hello Barbie", the NAO robot, the CogniToys green dinosaur and such are all personified systems in this sense. They also, generally, exhibit another of the defining attributes: they all rely on semantic analysis in order to understand what we are saying, and that analysis is almost always done "in the cloud" on remote servers. This means that these network-distributed systems have not merely simple data, but information that has been turned into knowledge (a distinction I describe below). Finally, personified systems act as virtual assistants or active agents acting on our behalf. They may be personal assistants, nannies, caretakers, medical assistants, or full- or part-time drivers in our planes, cars and trucks. They have job descriptions more than defined uses.

Agency and its absence

Let me now introduce a couple of terms that describe what these systems are not (yet). They are neither artificial general intelligences (AGIs) nor autonomous moral agents (AMAs).

> **Artificial General Intelligence** - An AGI is a system that is fully capable of performing any cognitive function that a human being could perform. This is sometimes referred to as "strong AI", to distinguish it from specialized systems that can perform only some functions, such as play chess or "Jeopardy", or refine the results of a user's search query. In many ways, it could be argued that an AGI should be considered to be a "person" rather than merely a "personified system" that behaves somewhat like a person.
>
> **Autonomous Moral Agent** - An AMA is an entity or system capable of acting (as an agent) on its own and making its own moral judgements. AMAs can be held responsible for their actions. Since moral judgements are a human cognitive function, presumably a fully realized AGI would also be capable of acting as an AMA, while personified systems would lack that ability.

While today's personified systems may mimic some aspects of true persons, they are neither conscious nor self-aware; they cannot pass the Turing Test. Rather, they mimic some aspects of genuine persons without actually being persons. They have no intentions or motivations. They are just systems that exhibit some characteristics of persons. They are, thus, "personified."

The topic of AI brings with it a lot of terms, some of which will show up in this document. There are a great many theories and approaches to creating artificial intelligence, and the distinctions between them are not always sharp. In general, there are two broad approaches.

The first, "logicist", approach seeks to reproduce in software the way we think. Loosely speaking, Logicist AI strives to represent intelligence, reasoning, and knowledge in terms of formal logic and mathematics operating on symbolic representations of knowledge and the world. Logicists concern themselves with creating *ontologies, taxonomies, grammars* and other representations of knowledge that model the world and what we know about it, and *algorithms* and *languages* to express them in, that capture the way that we reason about that knowledge and the world.

This has sometimes been called "Good Old-Fashioned Artificial Intelligence" or GOFAI. One way of looking at it is that it is about creating models of the world, our knowledge, and the ways that we reason about our knowledge.

The second, naturalist, approach seeks to emulate the ways that natural systems—the nervous system, especially the human brain—works. This dates back to the work of the early cyberneticists in the 1940s. It has reappeared with names such as Artificial Neural Networks (ANNs), Machine Learning (ML), Deep Learning (DL) and Hierarchical Temporal Memory (HTM), and been described as “biologically inspired”, “connectionist” and “Model Free Methods”.

These last two terms are worth explaining. “Connectionist” refers to our current theories regarding the ways that patterns of synaptic connections between neurons allow the nervous system to do the pattern matching that is at the heart of cognition. “Model free” refers to the absence of explicit models of the world in these systems. Practitioners seek to “model the brain/mind, not the world”.

This brings us to the second half of the introductory phrase that I used above: systems that are “well-behaved or trustworthy”. When computers were seen merely as things, as tools, we spoke of them in terms like “well-constructed”, “reliable”, “robust” or conversely “buggy” or “broken.” As they become personified, more normative language starts to be useful. We may think of them as “trustworthy”, “well-behaved”, “helpful”, “friendly”, “funny”, “smart” or “dumb”. This isn’t just a matter of language. It is tied to how we think about them, how we expect them to act, what we are willing to confide in them, how we and they behave.

More terms: Data, Information, Knowledge (and Wisdom)

One of the reasons that trust is becoming a bigger issue is that computer and IT systems are climbing what’s often called the “information hierarchy” and are dealing with data in more meaningful ways. There are a number of models of this hierarchy in information theory, often spoken of under the rubric of “DIKW”, standing for Data, Information, Knowledge, and Wisdom. In general, here is what is meant by the terms:

> **Data** refers to the simplest elements of information: symbols, facts, numerical values, or signals depending upon the context. We often speak of the “raw data”, meaning the pixels coming out of a camera or scanner; the numeric data such as temperature, air pressure, and wind direction that come from various sensors; unprocessed audio from a microphone; and the like. In and of itself, data is of minimal use or meaning.
>
> **Information** is where meaning begins. Information is inferred from data, and the relationships between different pieces of data. Information is data *about* something. Stepping outside the realm of computers and information for a moment, journalists turn facts into informative stories by asking the “who, what, where, when” questions. This is the primary distinction between data and information. A stream of X,Y mouse positions and clicks, or of signals from a microphone, key taps on a keyboard, become menu items selected, actions performed, words spoken or written. The data begins to be meaningful.

> **Knowledge** is where the meaning comes into its own. Knowledge deals with what the information is about. We see this distinction in systems like voice recognition or text auto-correcters. Dealing just with information, we can correlate certain patterns of sounds with specific phonemes or words, we can compare sequences of letters with words in a word list, or look at word pair frequencies to determine the most likely correct word. All of that is manipulating information. However if we know what the speaker is talking about, what the topic of a document is, then we have knowledge that will greatly improve our ability to choose the right transcription of the spoken word or to determine the correct word.
>
> **Wisdom** is often cited as the next step in the hierarchy, as the point where judgement comes in. As such, it is of only marginal significance with regard to the behavior of current systems as opposed to people. Wisdom is certainly a thing that we would like autonomous systems to embody. Only when they get to this level will we truly be able to talk about systems that behave and decide ethically. For now, however, they will have to leave that to us.

In the context of autonomous and personified systems, information is valuable, but knowledge is the real power. As they accumulate not mere information about us, but knowledge, the stakes are raised on our privacy. Allow me to finish the discussion of these terms with a story from the news of the last few years:

The point of sales systems at Target collect all sorts of data each time someone buys an item from one of their stores. This becomes useful information when it is correlated with stock inventories and can allow them to manage those inventories. The data also becomes information about their customers when sales are tagged with a customer ID. By analyzing patterns in the information they can gain knowledge about the customers, such as which ones are behaving as if they were pregnant. This allows them to target promotions to have just the right content at just the right time to capture new customers. It also leads, as the news reported, to a father being angered because Target knew that his daughter was pregnant before he did. Today, management at Target has the wisdom to merely increase the number of references to items of interest to new mothers in the customized catalogs mailed to expectant mothers rather than send them a catalog of nothing but new baby items.

This distinction and the spectrum from data to wisdom feature prominently in all of the work reported here. It is, for instance, debatable whether existing Machine Learning and Deep Learning techniques truly represent actual knowledge and understanding or are just very elaborate manipulations of information. If a system does not understand something, but merely recognizes patterns and categories, but presents it to a human who immediately understands the full meaning and implications, do we call it mere information or the beginnings of true knowledge?

The Roadmap

In the first half of this chapter, I outline what I would do if I were in charge of a full-fledged, well-funded R&D effort for designing trustworthy personified systems.

Long term or "Doing it right"

To do the job properly, to identify what principles of system behavior should be adopted and then begin to implement them, would be a major effort for any enterprise. In this section I will outline the types of research and development activities that would be involved. In the real world, any enterprise or institution that mounted an R&D project or program in this space would probably scope the effort differently, and while it is quite possible to envision that scope being larger, it is more likely that it would be narrower than what I am describing here. Still, it seems worthwhile to look at it in a large generalized context.

Business: Choose a focus

"Personified systems", as I have defined them, cover a huge array of existing and emerging products. Any actual R&D project is likely to address only a subset of them, chosen according to the business and research requirements of the enterprise or institution doing the research and/or development.

Some of the types of systems that are included in the general category are:

- **Generalized personal assistants:** This category includes such systems as Apple's Siri, Google Now, Microsoft's Cortana, and Amazon's Echo, aka "Alexa". These systems respond to voice and text commands and queries given in a quasi-natural language for human/machine dialogs, and perform tasks such as looking up information, performing calculations, creating and scheduling reminders and calendar events, sending emails and text messages, taking notes, creating shopping and task lists, and the various functions that are becoming the common daily uses of computers and mobile devices.
- **Specialized virtual executive assistants:** This is a more advanced version of the previous category and includes systems such as "Monica", the virtual assistant of Eric Horvitz at Microsoft Research[1]. Whereas the simpler assistants of the previous category interact solely with the user they serve, Monica, assisted by a small Nao robot and other systems at Microsoft Research, deals with other people on her user's behalf. Monica greets visitors to Horvitz's door, schedules appointments and in general fills a role that might otherwise have been performed by a secretary or executive assistant.

[1] Suzanne Choney, "Smart elevator, robot and virtual admin will help guide you at Microsoft Research", *The Fire Hose (blog)*, Microsoft, April 8, 2014, http://blogs.microsoft.com/firehose/..., accessed June 20, 2016

- **Companions for children:** This category comprises both educational and entertainment systems that serve the roles of companion or nanny. It includes devices such as the Cognitoys green dinosaur, Mattel's "Hello Barbie", ToyTalk's "The Winston Show", and many others. These systems differ, both in that the user is a child and that the customer is not the user. While true "virtual nannies", robotic systems given some form of responsibility, are still in the future, there are R&D projects headed in that direction, and their behavior will be important enough that we want to be thinking about it now.

- **Virtual caregivers:** The Japanese, in particular,[2] have been expending considerable effort building care-giving systems, especially for the aged. As the median age of the populace is going up and the number of children in families is going down, the demands and requirements for the care of the old and the infirm are growing. While autonomous systems capable of fully taking on these duties are well in the future, there are many ways that autonomous systems might assist and complement human caregivers.

- **Expert system virtual assistants:** Many professions are making increased use of automated virtual assistants. Virtual medical assistants are helping with diagnosis, prescribing and managing drug regimes and other tasks. In the financial world, autonomous systems not only give advice, but have taken over many aspects of stock and commodity trading. In the legal field, the tools have been shifting from simple search to autonomous systems assisting or replacing aspects of a law clerk's job. All of these professions deal with confidential and sensitive information, and a require high degree of trust.

- **Autonomous and semi-autonomous vehicles:** More and more autonomous driver-assist and self-driving features are appearing not only in research and concept vehicles, but in production vehicles, and on our roads. These systems are entrusted with human safety and operate dangerous machinery. As such, they are taking on nearly human responsibilities and require substantial trust. In addition to cars, commercial airliners are already flying under automated control for the vast majority of flight time.

- **Home appliances and systems:** The "Internet of Things" is growing rapidly, with everything from thermostats and alarm systems to lights and entertainment systems being automated and coordinated. Voice response systems monitoring and controlling these devices are becoming more sophisticated and are merging with the general purpose virtual assistants.

[2] Jacob M. Schlesinger and Alexander Martin, "Graying Japan Tries to Embrace the Golden Years", *Wall Street Journal 2050: Demographic Destiny,* Wall Street Journal, Nov. 29, 2015, http://www.wsj.com/articles/graying-japan-tries-to-embrace-the-golden-years-1448808028 accessed June 29, 2016

- **Games and Toys:** In addition to the talking toys and games mentioned under the “Companions for children” category, AIs act as players and non-player elements of many games. While it is not clear how crucial trust is in these systems, they do play an increasing role in the public’s dealing with and understanding of autonomous systems, and may influence opinions and expectations well beyond the limits of the games they inhabit. Additionally, there is the whole area of adult and sexual toys and entertainment. Here, issues of trust and confidentiality may be quite important.

Sociology and Psychology

There are a number of social and psychological issues that need to be addressed either through the expertise of the participants in the team or through explicit research. These questions come in the areas both of the broad background and context in which the systems operate and in the specific interactions of the individual systems being studied and developed in the area of focus. Areas that need to be covered are:

Societal expectations: How do people think of personified systems, robots and the like? What are our expectations of them?

Natural man/machine dialogs: Given that background, how do people speak to and interact with machines? In many ways, this is similar to how we interact with each other, but research shows that knowing something to be a machine alters the nature of our interactions with it. Part of this is due to the different capabilities of machines that are not yet fully intelligent. Some is due to the expectations that society, fiction and media set. Finally, some of it is because of the different roles that artificial systems play.

Impact upon us: For the foreseeable future, personified systems and AGIs will serve a sub-human role in society. This is likely to be so even for AGIs until they are not only deserving of rights but recognized by society as deserving. This role as “sub-human” will have an impact on us. As we treat them as both persons or person-like, but at the same time as inferior, it is likely to affect how we deal with other humans. Will it be a pressure upon us to go back to thinking of some people as sub-human persons or will it clarify the line between all humans, who are full persons, and non-humans, who are not?

Social psychology of morality: Substantial research has been done in both the neurophysiology and the biological and social foundations of morality. Work on this project needs to be well grounded in these aspects of human morality and behavior in order to understand how artificial systems can be integrated into society.

Jonathan Haidt's Social Intuitionist model and Moral Foundations theory[3], if valid and accurate, may provide a valuable grounding in understanding the human morality into which we are trying to integrate autonomous systems. On the other hand, Kurt Gray's critique of the specific foundations of Haidt's work, as well as his own theories regarding our formation of a theory of mind[4] and the role of our perception of membership in the "Mind Club", provide alternative clues on how to integrate personified systems into people's moral and social interactions.

Philosophy and Ethics

The next step, based upon the roles and capabilities of the systems in the area of focus, and upon the expectations, needs and desire of the users, is to decide upon a suitable model of normative ethics, and then to flesh it out. There are three major classes of normative ethics: deontological, that is to say rule-based; consequentialist, focusing on the results and impact of actions; and virtue-based, focusing on the character of the actor.

After we choose one of these three major paradigms, or some form of hybrid normative system that combines aspects of all three, we will need to develop a more detailed system of rules, or principles will need to be developed. Again, the area of focus will have a significant impact upon which specific virtues, rules or principles are chosen, and the priority relationships between them, but I expect that there is a great overlap between the requirements of different areas.

For virtually all of the focus areas, there are existing human professions with their own rules of conduct, ethical standards and requisite virtues. Designing a specific normative system will call upon those bodies of work, along with general ethical considerations and the peculiarities of establishing a normative system for non-human and—in terms of both cognitive and ethical realms—sub-human actors. Even when truly autonomous moral agents emerge and are generally recognized, it seems likely that their natures will still be different enough that there will be differences between the normative system controlling their behavior and that of humans.

One area of study that will need to be addressed is the impact upon us as humans of dealing with sub-human actors, and the normative systems that apply to both them and us. We are barely getting to the point where our social systems no longer recognize classes of supposedly inferior humans, and we have not gotten very far in considering the ethical roles

[3] Jonathan Haidt, "MoralFoundations.org", YourMorals.org collaboration, January 30, 2016, http://moralfoundations.org accessed June 29, 2016

[4] Kurt Gray, Adam Waytz and Liane Young, "The Moral Dyad: A Fundamental Template Unifying Moral Judgment", *Psychological Inquiry*, National Center for Biotechnology Information, April 2012, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3396360/ accessed June 29, 2016

of animals. As machines begin to be personified, as we begin to speak to them, and even converse, the impact upon our own consciences and behavior of dealing and interacting with non-persons or semi-persons with limited or different ethical roles will need to be monitored.

Operational

Once we've chosen a particular normative ethical system and set of principles, rules or virtues, the set will need to be operationalized, and the priorities and relationships between them will need to be clearly defined. As existing systems fall far short of the analytic capabilities and judgements required to apply the rules and principles to specific actions, the bulk of that analysis will have to be done by researchers and developers and turned into specifications and descriptions of the required behavior of the systems. This is a major undertaking.

Technical

Once the rules have been operationalized, they will have to be embodied in the design and implementation of the actual software systems involved, both the analytic systems (most of which reside in the cloud today) and the more active agents and applications that reside in mobile, desktop and IoT devices. Since the handling and exfiltration of sensitive information is a major issue in the trustworthiness of these systems, special care will have to be taken in the design of the distributed aspects of the system so as to control the domains to which various pieces of information are exposed.

Short term or "What I've been up to"

In this section, I will outline what I have been doing and am continuing to do on my own, during my current sabbatical.

In the previous section, I laid out a grand plan for conducting a hypothetical program of research into building trustworthy personified systems, if such an effort were performed by a group or department within a major enterprise or institution. This section describes the more limited task that I have been able to address on my own.

One of the most important things that I learned when I was working in software human factors and usability research is that a combination of real-world testing and research is more powerful and effective than relying on mere human expertise. Nonetheless, expertise and experience have their place and so I have been applying mine to this question for the last few months. What follows is a summary of what I have been doing and continue to do in this field.

Business: Choose a focus

Since my researches are not tied to a specific project or charter from a specific organization, I have been somewhat freewheeling in focus. In general, I have been focusing on consumer software systems such as virtual assistants and voice recognition systems, and not on hardware intensive and specialist systems such as military uses, robots, drones, and the like.

There is perhaps one exception to this: the question of autonomous vehicles. Like robots and drones, these are hardware systems, but in the short term, they are being introduced as a number of smaller assistive technologies—systems for parking the car or changing lanes on the highway or avoiding collisions. They are, thus, in a way, virtual assistants, but assistants that add a new dimension of their own: the potential to be responsible for actual physical harm (or its avoidance).

Sociology and psychology

Sociology and psychology come into this process in a couple of fundamental ways. They inform us as to the origins and mechanisms of human morality and trust, and tell us how people view personified systems in general and interact with the sorts of systems that exist today. I have been relying on two sources in the area of how people view personified systems: the experiments that I did a couple of decades back in "natural man/machine dialogs", and my own informal survey of how robots and computers have been presented in mass media over the last 6 or 7 decades.

During the 1980s I was a member of an R&D team at Digital Equipment Corporation that was variously known as Human Engineering Research, Software Human Engineering and Software Usability Engineering. We did some of the basic foundational research on the topics of man/machine dialogs and software usability[5]. One of our many findings was that people talked to machines in a way that was similar to, but distinct from, the way that they addressed each other, something we referred to as "Natural Man/Machine Dialogs" and studied with a number of simulated systems, including the so-called "Wizard of Oz" setup, whereby a hidden researcher provided the flexible intelligence that software lacked[6]. We then built systems that behaved the way that the people expected.

One of the things that became clear to me personally at that point was that people's expectations derived in good part from the portrayal of robots and computers in popular

[5] Michael Good, "Software Usability Engineering", *Digital Technical Journal*, February 1988, Digital Equipment Corporation, http://michaelgood.info/publications/usability/software-usability-engineering/ accessed June 29, 2016

[6] Michael D. Good, John A. Whiteside, Dennis R. Wixon, & Sandra J. Jones, "Building a User-Derived Interface", *Communications of the ACM*, October 1984, http://michaelgood.info/publications/usability/building-a-user-derived-interface/ June 29, 2016

fiction; these in turn depended upon the expectations of the folk who were writing and creating the media representations. This process has clearly been continuing over the last 3 or so decades, in a process of iterative creative refinement, one that also interacts with people's experiences with computers, robots, and other automated systems, and which has contributed to the design of those systems. Fiction and technology have been shaping each other

Since trust is, in part, dependent upon living up to people's expectations, being aware of the expectations set in our fiction and in aspirational technical visions, such as the classic Knowledge Navigator videos from Apple, can contribute important context.

In order to understand how to make systems be and be perceived as trustworthy, we need to understand the role, mechanism, and history of trust in human society. An excellent source on this, which I consulted at the start of my project, is Bruce Schneier's *Liars and Outliers*[7]. Schneier, a technology and security expert by trade, brings a very technical and detailed viewpoint to the subject.

I have also been doing a survey in the area of the social psychology of morality. This is a field that has grown considerably since my days as a social psychology major and serves as one of several foundations for the next area, philosophy and ethics. As we personify systems, and begin to evaluate their behavior using terms like "trustworthiness", "loyalty", and "discretion", it becomes important to understand what causes us to perceive, or fail to perceive, these qualities in others—what is the user's theory of mind, and what adds to and detracts from our perceptions of others as "moral", "immoral", and "amoral".

Philosophy and Ethics

In discussing this project with friends and colleagues, the topic of Artificial Intelligence and when robots and other autonomous systems would be true autonomous moral agents naturally came up. When it did, during these discussions, I realized that the reasons that I have always had low expectations of AI had become clear enough that I could write about them usefully. The result is a short essay, not properly part of this project, on the topic of AI and "Common Sense" in the original Aristotelian or medieval sense. It can be found here.

Given that Artificial General Intelligence—AGI—intelligence on the order of our own, which is capable of allowing artificial systems to become true autonomous moral agents, is still well in the future (traditionally 25 years from "now" for the current value of "now"), we need to address not the question of how to endow personified systems with morals and ethical judgement, but rather, how to get them to behave in a manner that is congruent with human

[7] Bruce Schneier, "Liars and Outliers: Enabling the Trust that Society Needs to Thrive", John Wiley & Sons, February, 2012, https://www.schneier.com/books/liars_and_outliers/ accessed June 29, 2016

norms and ethics. That, in turn, leaves us with the question of what system of morals they should align with.

There are three broad schools of normative ethics: deontology, ethics based upon rules of behavior; consequentialism, based upon the outcome, or expected outcome of our actions; and virtue ethics, based upon the nature and character of the actor. While some evaluation of consequences and obedience to rules certainly have a place in designing trustworthy AI, my tentative judgment is that the center of the project must focus on virtues.

After surveying systems of virtues from Eagle Scouts to war fighters to the professional standards of human professionals, as well as butlers and valets, I've come up with a system of virtues that I am broadly categorizing either as "functional" virtues or aspects of trustworthiness. They are:

1. Functional or utilitarian virtues
    1. Helpfulness
    2. Obedience
    3. Friendliness
    4. Courtesy
2. Aspects of Trustworthiness
    1. Loyalty
    2. Candor
    3. Discretion
    4. "Propriety"

The first group are attributes that are familiar to software developers, UI/UX specialists, and usability engineers. They are very similar to the attributes of good, usable, approachable user interfaces. The second group consists of trustworthiness and four subsidiary virtues that it comprises. It is these that I have been focusing my attention on. The five "Which Ethics?..." chapters address the selection of these norms.

Operational

In order to endow systems with these virtues, system architects and software developers need clear definitions of what each one means. They can, then, ask themselves each time that they are deciding how a system should behave, "What would a virtuous person do?" By focusing on emulating the behaviors that are commensurate with trustworthiness and its subsidiary virtues, engineers can create systems that will integrate as smoothly as possible into society. The chapter entitles "Which Virtues" discusses this in detail. On the whole, the operational definitions are still a work in progress.

Technical

Once operational definitions are available for the principles and virtues, they will need to be turned into a set of guidelines and design elements that can serve as a starting point or template for design and implementation of systems. Some of these occur to me as I work to operationalize the virtues, and I am beginning to start collecting some. So long as this is a small one-man theoretical exercise, this list will remain substantially incomplete, but as a long-time system and software architect, I cannot help but contemplate and accumulate a few.

Which Ethics?

Determining how the behavior of personified systems can best be made trustworthy depends upon a number of factors. First of all, we need to decide what approach to take, which style of normative ethics best suits personified systems. Beyond that, though, there is the question of how the behavior of systems will evolve over time as they become more intelligent, more autonomous, and more person-like. An approach well suited to mere personified systems may not work nearly as well for AGIs that are capable of autonomous moral agency, and *vice versa*. If possible, it would be desirable that the behavior and people's expectations of autonomous systems span their evolution. It is, therefore, worth evaluating approaches to normative ethics both in terms of what is suitable today, and how well they adapt over time.

The Three Norms

In this chapter, I will give a brief overview of the schools of normative ethics and give a few pointers to related outside reading and viewing. Subsequent chapters will deal with each of the three schools in detail, will look at the work being in that area, and will discuss its suitability and short- and long-term prospects.

The study of ethics has several sub-disciplines. The one we are most concerned with in this effort is "Normative Ethics", defined as systems of norms for behavior. Ordinarily, these norms are thought of as controlling human behavior, but in our case, we are looking to define the norms of partially autonomous systems created by humans. There are at least three domains in which these norms might apply: the behavior of the people creating the systems, norms built into the systems, and norms created by the systems, once they are capable of it. It is the second of these, the norms that control the behavior of systems that are not themselves autonomous moral agents (AMAs), that we are most concerned with, though the other two senses will be referred to occasionally.

The three main categories of normative ethics are deontology, systems of rules controlling ethical behavior; consequentialism, systems based upon evaluating the consequences of those actions; and virtue ethics which center upon the character of the actor, rather than the rules or consequences. Each has something to offer in the area of machine ethics. A brief introduction of them individually follows, and the next three chapters explore them in detail.

Deontology: Driven by rules

Deontology seeks to guide our actions through the application of a set of moral rules or principles. One of the more highly regarded deontological systems is Kant's "Categorical Imperative", the principle of universality: "Act only in accordance with that maxim through which you can at the same time will that it become a universal law." Perhaps even more widely applied are the systems of divine commandments of the various world religions, such as the Ten Commandments, Christ's two great commandments, and so on.

There are, of course, many other deontological systems. They have in common the focus on establishing a rule or set of rules for judging acts to be moral or immoral.

On the surface, deontological systems would seem to be well suited to controlling the behavior of computer systems, systems which themselves consist of bodies of computer code. It might be argued, in fact, that such systems are nothing but detailed specifications of rules controlling the behavior of the system. So, if they are already just implementations of systems of rules, why not add rules of ethics?

The problem that emerges, though, is that what computers excel at is combining very simple rules defined in very precise terms. Actual deontological systems depend upon understanding far more complex and nuanced terms. When Kant tells us to act in a way such that we would wish that all other people act that way, it is a very simple statement, but the implications are profound. How do we wish others to act? How would an autonomous system know what that is?

Similarly, the Ten Commandments are often said to tell us "Thou shalt not kill" and yet we have war, self-defense, and executions. Looking more closely at the commandment, we find that it is more accurately translated as "Thou shalt do no murder", and "murder" is defined as unjust killing, while self-defense, lawfully declared war, and executions passed down as sentences by lawful courts are considered "just". How do we define all of that?

There are a number of advocates of a deontological approach to causing autonomous systems to behave ethically. Selmer Bringsjord of the Rensselaer Polytechnic Institute and his associates, for instance, created a language and a system for expressing the knowledge, goals, and behavior of autonomous systems that they call the "Deontic Cognitive Event Calculus" (DCEC).

The husband and wife team of computer specialist Michael Anderson and philosopher Susan Anderson have been working on methodologies based on the deontological "*prima facie* duty" theories of W.D. Ross. Recently, they have been exploring the application of Machine Learning to derive such duties. "Which Ethics?—Deontology" will cover the work of both Bringsjord and the Andersons.

Useful resources on deontology, both specific to autonomous systems and in general:

- Selmer Bringsjord and his associates' "Deontic Cognitive Event Calculus" system[8,9].
- The Andersons' papers,
    - "Machine Ethics: Creating an Ethical Intelligent Agent"[10],
    - "A Prima Facie Duty Approach to Machine Ethics and Its Application to Elder Care"[11]
    - "Robot, Be Good"[12]
    - "Ensuring Ethical Behavior from Autonomous Systems"[13]
- The Stanford Encyclopedia of Philosophy has a good article on Deontology[14], and of course, many others.
- The University of Tennessee's Internet Encyclopedia of Philosophy's article on W.D. Ross includes a discussion of his ethical system and *prima facie* duties[15].

[8] Selmer Bringsjord et al, "Deontic Cognitive Event Calculus (Formal Specification)", Rensselaer Artificial Intelligence and Reasoning Laboratory, May, 2013, http://www.cs.rpi.edu/~govinn/dcec.pdf accessed July 13, 2016

[9] Naveen Sundar Govindarajulu et al, "On Deep Computational Formalization of Natural Language", Rensselaer Artificial Intelligence and Reasoning Laboratory, 2013, http://kryten.mm.rpi.edu/SELPAP/2013.FormalMagic/main.pdf accessed July 13, 2016

[10] Michael Anderson & Susan Leigh Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent", *AI Magazine*, 2007, AAAI, http://ieet.org/archive/AIMagFinal.pdf accessed July 13, 2016

[11] Susan Leigh Anderson & Michael Anderson, "A Prima Facie Duty Approach to Machine Ethics and Its Application to Elder Care", *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence,* AAAI Publications, 2011, http://www.aaai.org/ocs/index.php/WS/AAAIW11/paper/view/3812/4274 accessed July 13, 2016

[12] Susan Leigh Anderson & Michael Anderson, "Robot Be Good", *Scientific American #303,* October 2010, http://franz.com/success/customer_apps/artificial_intelligence/EthEl/robot-be-good.PDF accessed July 13, 2016

[13] Michael Anderson & Susan Leigh Anderson, "Toward Ensuring Ethical Behavior from Autonomous Systems: A Case-Supported Principle-Based Paradigm", *2014 AAAI Fall Symposium Series*, AAAI Publications, 2014, https://www.aaai.org/ocs/index.php/FSS/FSS14/paper/viewFile/9106/9131 accessed July 13, 2016

[14] Larry Alexander & Michael Moore, "Deontological Ethics", *Stanford Encyclopedia of Philosophy*, December 12, 2012, http://plato.stanford.edu/entries/ethics-deontological/ accessed July 13, 2016

[15] David L. Simpson, "Ross's Ethical Theory: Main Components and Principles", *"William David Ross (1877—1971)", Internet Encyclopedia of Philosophy*, http://www.iep.utm.edu/ross-wd/#H6 accessed July 13, 2016

Consequentialism: Means to an end

The second approach is to consider not rules of behavior for the actor, but the consequences of their actions to others and to themselves. Consequentialist systems determine the ethics of an action by the results that it brings about. Examples of consequentialist ethical systems are utilitarianism, which calls for maximizing human wellbeing, and hedonism, which maximizes human pleasure.

Prof. Alan Winfield of the University of the West of England, Bristol, and his students and colleagues have done a good deal of work on what he calls the “Consequence Engine”, an approach for, first off, making robots safe to be around, but which he hopes will also enable them to act ethically. His design uses a secondary system that is capable of simulating the robot and its environs, which is used to test out various alternatives, ruling out those choices that would result in harm to humans. It then turns over the remaining set of actions to the robot’s main control system to pick the preferred action. In essence, his Consequence Engine’s job is to insure that the robot follows the famous dictum, 'First, do no harm.'[16]

The problem with this approach is that simulating the robot, its environment, and its actions is hard. Since, as Prof. Winfield himself points out, there are many kinds of harm to humans that must be prevented, there are many aspects of the world that must be simulated, in order for the engine’s predictions to be fully effective. Still, by creating a framework that can be implemented with greater and greater detail and fidelity to the real world, this approach provides an incremental mechanism that can be tuned and improved over time, rendering a consequentialist robot, AI or AGI ethics at least plausible.

[16] Alan Winfield *et al*, “Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection”, *Engineering Moral Agents - from Human Morality to Artificial Morality*, Schloss Dagstuhl, 2016, http://materials.dagstuhl.de/files/16/16222/16222.AlanFTWinfield.Preprint.pdf accessed July 14, 2016

Useful resources on consequentialism:

- Prof. Winfield's papers[17], blog posts[18] and videos[19],[20] are worth reading and viewing.
- The BBC has a page discussing Consequentialism[21] in their *Ethical Guide.*
- The BBC article, in turn, refers to a far lengthier article at UTM's Internet Encyclopedia of Philosophy[22].

[17] Alan Winfield, CV and recent papers, 2016, http://www.cems.uwe.ac.uk/~a-winfield/ accessed July 14, 2016

[18] Alan Winfield, "*Alan Winfield's Web Log*", 2016, http://alanwinfield.blogspot.com/ accessed July 14, 2016

[19] Alan Winfield, "Making an ethical machine", *World Economic Forum,* February 19, 2016, https://youtu.be/qZ2zHL4x5r8 accessed July 14, 2016

[20] Alan Winfield, "The Thinking Robot", *Talks at Google*, May 10, 2016, https://youtu.be/-e2MrWYRUF8 accessed July 14, 2016

[21] The BBC, "Consequentialism", *Ethics guide,* 2014, http://www.bbc.co.uk/ethics/introduction/consequentialism_1.shtml accessed July 14, 2016

[22] William Haines, "Consequentialism", Internet Encyclopedia of Philosophy, http://www.iep.utm.edu/conseque/ accessed July 14, 2016

Virtue Ethics: Virtual character

The third approach to ethics is to consider the character of the actor rather than the effects of the action or the duties and rules that determine it. Virtue-based ethics goes back to Aristotle. With its emphasis on character and judgement, Virtue ethics is often thought of as the least likely fit for autonomous systems. Still, several writers see a plausible role in both short-term personified systems and eventual AGIs. I will explore this in the chapter "Which Ethics? — Virtue Ethics". Until then, here are a few references as food for thought.

Anthony Beavers covers the role of virtue ethics in his chapter of the book *Robot Ethics* entitled "Moral Machines and the Threat of Ethical Nihilism". Beavers cites, among others, an early and influential paper, James Gips' "Towards the Ethical Robot", in which he covers all three schools of normative ethics. Of Virtue ethics, Gips writes:

> *"The virtue-based approach to ethics, especially that of Aristotle, seems to resonate well with the modern connectionist approach to AI. Both seem to emphasize the immediate, the perceptual, the non-symbolic. Both emphasize development by training rather than by the teaching of abstract theory."*

Philosopher Daniel Hicks has written an article "Virtue Ethics for Robots", in which he criticizes both Deontological and Utilitarian (consequentialist) principle-based systems for not fully dealing with moral dilemmas, appropriate ethical responses to them and the lack of what he calls "tactical creativity" for dealing with specific situations. Hicks is not, specifically, an expert in AI and machine ethics, and refers repeatedly to the notion that robots "follow their programming". It is not clear that he understands, as Gips did, the extent to which Machine Learning systems create their own programming.

Shannon Vallor of Santa Clara University in her paper "The Future of Military Virtue: Autonomous Systems and the Moral Deskilling of the Military", considers the role of autonomous systems in combat, and their impact on human ethical skills. Her conclusion is that we should, perhaps, restrict the deployment of automated methods of warfare to appropriate contexts, and work to increase the ethical skills of the human combatants.

The area of the ethics of autonomous military systems is a particularly tough one. On the one hand, if you are going to trust robots to use lethal force, there are good reasons to, as the military does, insist upon explicit and provable ethical rules, rules such as Bringsjord's DCEC. The price of error is very high. On the other hand, virtue has always driven the world's militaries at least as much as deontological rules, and should be considered.

Useful resources on virtue ethics:

- Patrick Lin *et al.*'s book, “Robot Ethics: The Ethical and Social Implications of Robotics”.[23]
- Anthony Beavers' article “Moral Machines and the Threat of Ethical Nihilism”.[24]
- James Gips' 1991 paper, “Towards the Ethical Robot”.[25]
- Daniel Hicks' blog article, “Virtue Ethics for Robots”.[26]
- Shannon Vallor's paper “The Future of Military Virtue”.[27]
- On a lighter note, there is a Youtube video in which two “Robots Discuss Virtue Theory”.[28]

[23] Patrick Lin, Keith Abney & George A. Bekey, *Robot Ethics: The Ethical and Social Implications of Robotics*, The MIT Press, December 2011, https://mitpress.mit.edu/books/robot-ethics accessed July 15, 2016

[24] Anthony F. Beavers, “Moral Machines and the Threat of Ethical Nihilism”, draft, 2011, http://faculty.evansville.edu/tb2/PDFs/Moral%20Machines%20and%20the%20Threat%20of%20Ethical%20Nihilism.pdf accessed July 15, 2016

[25] James Gips, “Towards the Ethical Robot”, draft, May 1991, http://www.cs.bc.edu/~gips/EthicalRobot.pdf accessed July 15, 2016

[26] Daniel Hicks, “Virtue Ethics for Robots”, *Je fais, donc je suis* [blog], Jun3 2014, http://jefais.tumblr.com/post/89164919838/virtue-ethics-for-robots accessed July 15, 2016

[27] Shannon Vallor, “The Future of Military Virtue: Autonomous Systems and the Moral Deskilling of the Military”, *5th International Conference on Cyber Conflict*, 2013, https://ccdcoe.org/cycon/2013/proceedings/d2r1s10_vallor.pdf accessed July 15, 2016

[28] Douglass McFerran, Robots Discuss Virtue Theory, *YouTube,* December 2012, https://youtu.be/wKTZSzMeMpE accessed July 15, 2016

Which Ethics? — Deontology

Deontology is an approach to Normative Ethics that is based upon the premise that what a person ought to do is determined by some set of rules or principles. It includes codes of ethics like The Ten Commandments or principles such as Kant's Categorical Imperative[29]. Deontological ethics have been associated with automated systems since at least the advent of Isaac Asimov's "Three Laws of Robotics"[30].

Bringsjord, et al.

One of the strongest contemporary advocates of a deontological approach to "roboethics" and regulating the behavior of AIs, robots and other automated systems is Selmer Bringsjord. Bringsjord comes from the older "logicist" school of AI. He describes this perspective (and calls for it to become an independent discipline separate from all other AI efforts) in a paper entitled "The Logicist's Manifesto"[31]. In it, he categorizes the Logicist perspective as having three attributes. Logicist AI (LAI) is:

- **Ambitious**—LAI is an ambitious enterprise: it aims at building artificial persons.
- **Logical Systems**— LAI is specifically based on the formalization of one of the distinguishing features of persons, namely that they are bearers of propositional attitudes (such as *knows*, *believes*, *intends*, etc.), and that persons reason over such attitudes (which are often directed toward the propositional attitudes of other agents). This formalization is achieved via logical systems.
- **Top-Down**—LAI is a top-down enterprise: it starts by immediately tackling that which is distinctive of persons (e.g., propositional attitudes), without wasting [time] dwelling on the adventitious embodiment of cognition in particular physical stuff, or (what will later be called stage-one) transduction between the external physical environment and high-level cognitive processes.

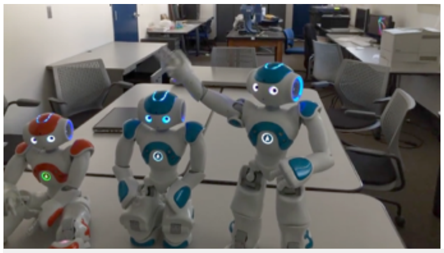
[29] "act only in accordance with that maxim through which you can at the same time will that it become a universal law." See the Stanford Encyclopedia of Philosophy for more on the CI, Kant's Moral Philosophy and Deontology in general.

[30] It is, perhaps, worth noting that Asimov's Three Laws themselves are not a particularly good system in the real world, as he himself pointed out upon occasion. He formulated them not to control actual robots, but as a source of conflict to drive the plots of his story. Each story in which they featured was centered around some oversight, self-contradiction, or conflict among the Three Laws. It was believable that people would attempt to control robots some day using a mechanism such as the Three Laws, and that the set that they chose, while they might appear adequate, would be sufficiently flawed so as to offer all the conflict that he would need to write a large number of stories.

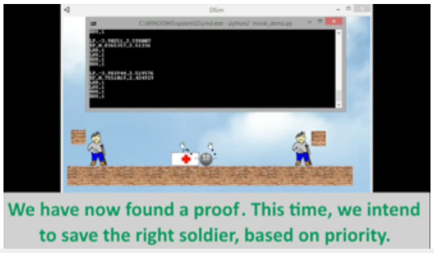
[31] Selmer Bringsjord, "The Logicist Manifesto: At Long Last Let Logic-Based Artificial Intelligence Become a Field Unto Itself", September, 2008, http://kryten.mm.rpi.edu/SB_LAI_Manifesto_091808.pdf accessed July 15, 2016

Based on this top-down, logic-driven understanding of what personhood is, and how artificial persons can be created, Bringsjord and his colleagues have created a language and system called “Deontic Cognitive Event Calculus” (DCEC). This is a system that allows propositions of the sort that he mentions—“X knows/believes/intends/is obliged to”, and so forth—and allows them to be combined such that a statement that would be expressed in English as “If you come across a wounded soldier, you are obliged to add getting him to a MedEvac unit to your system of goals and continue to act upon those goals” can be expressed algorithmically.

Using this system, they claim to have made a number of major advances toward creating persons, and even persons with ethical reasoning and motivations. Here are three videos illustrating some of their claimed accomplishments: “Self awareness”, “Solving moral dilemmas”, and “Akrasia in robots”:



Self Consciousness with NAO Bots



We have now found a proof. This time, we intend to save the right soldier, based on priority.

Solving a moral dilemma with DCEC

Akrasia in robots demonstration

Please note that I have referred to them as “claimed accomplishments” because there are other researchers and theorists who might explain what they show in quite different terms. With the caveat that I don’t share Bringsjord’s Logicist perspective, allow me to explain what each of these is.

In the first video, we see the experiment that was announced as demonstrating the rudiments of “self-awareness” in a robot. The demonstration is loosely based on the “Three Wise Men” puzzle that has been around for a long time and is described in the “Logicist’s Manifesto”. Three robots have been told that two of them have been “given a dumb pill”, which will render them unable to speak, and the third was “given a placebo”. They are then asked “Which pill did you receive”. One of them says, “I don’t know,” hears itself, then corrects itself to say “Now I know”. This, they argue, demonstrates awareness of self.

In the second video, a collection of goals and methodologies (and some other unspecified instructions) results in a demonstration of how priorities drive decision making and new strategies emerge when the priorities are equal.

The third video is said to demonstrate “akrasia” or working against your own self interest. Two robots, tagged with large red or blue dots that identify their faction, enact a play in which one of them is a prisoner and the other a guard. They then “hurt” or “refrain from hurting" each other by whether they intentionally collide. One really must know the details of the scenario and setup in order to see how self interest, self-awareness and akrasia are manifested in this little play.

Regardless of whether these videos actually demonstrate the philosophical points that they are trying to make, the DCEC language and system is proving to be a convenient and concise way to model some types of behavior and control systems. Given that, deontology should not be written off as an approach to the normative ethics of personified systems. Its applicability to full AIs may, however, be more controversial, at least in the case of the work of Bringsjord and his team.

Besides his intentionally provocative Logicist's Manifesto, many of his other writings are, if not controversial, at least somewhat provocative. In their semi-formal paper, "Piagetian Roboethics via Category Theory: Moving Beyond Mere Formal Operations to Engineer Robots Whose Decisions are Guaranteed to be Ethically Correct", for instance, Bringsjord *et al.* tie their work to Jean Piaget's fourth stage of logical development, but leave out all references to the developmental nature of Piaget's work other than to say that a provably correct ethical robot would have to be based on a yet-to-be-developed fifth or higher stage. This is in keeping with the Logicists' top-down approach.

At the heart of their notion of a stage of logical reasoning more mature than Piaget's fourth, adult, reasoning stage is one based upon category theory, which they say will allow the robot to derive its own codes of conduct. This is needed, they say, because fourth stage reasoning is inadequate. They give the following example:

> *"Imagine a code of conduct that recommends some action which, in the broader context, is positively immoral. For example, if human Jones carries a device which, if not eliminated, will (by his plan) see to the incineration of a metropolis, and a robot (e.g., an unmanned, autonomous UAV) is bound by a code of conduct not to destroy Jones because he happens to be a civilian, or be in a church, or at a cemetery ... the robot has just one shot to save the day, and this is it, it would be immoral not to eliminate Jones."*

This example is in keeping with uses for automated systems that Bringsjord sees as vital, as outlined in an opinion piece that he wrote—"Only a Technology Triad Can Tame Terror"—for the Troy Record[32], and which is referred to and expanded upon in a presentation that he gave for the Minds & Machines program at RPI in 2007[33]. In that piece he concluded that the

[32] Selmer Bringsjord, "Only a Technology Triad Can Tame Terror", *Troy Record*, August 9, 2007, http://kryten.mm.rpi.edu/NEWSP/Only_Technology_Can_Tame_Terror_080907.pdf accessed July 15, 2016

[33] Selmer Bringsjord, "Only a Technology Triad Can Tame Terror", *Minds & Machines @ RPI*, 2007, http://kryten.mm.rpi.edu/PRES/TAMETERROR/sb_tameterror.pdf accessed July 15, 2016

only protection that we can have against terrorism and mass shootings such as the one at Virginia Tech is to build a triad of technologies:

> *"Our engineers must be given the resources to produce the perfected marriage of a trio: pervasive, all-seeing sensors; automated reasoners; and autonomous, lethal robots. In short, we need small machines that can see and hear in every corner; machines smart enough to understand and reason over the raw data that these sensing machines perceive; and machines able to instantly and infallibly fire autonomously on the strength of what the reasoning implies."*

Given that he believes this sort of technology is necessary, it is easy to see why Bringsjord and company insist upon a system of logically provable ethics. The effort to create fully Logicist AIs that are controlled by an explicit system such as the one that implements their DCEC language has been the main focus of their work for several years. Perhaps the best introduction to their work is their article "Toward a General Logicist Methodology for Engineering Ethically Correct Robots".[34]

The home page for the Rensselaer Artificial Intelligence and Reasoning (RAIR) Laboratory[35] contains details on many of their projects, including the DCEC system, and Psychometric Artificial General Intelligence World (PAGI World) ("pay-guy"), a simulation environment used for AI and AGI testing (and seen in the second video above).

Critics of this approach have suggested that Logicist AI is incapable of creating the artificial persons that they are seeking, that LAI is not a valid path to a true Artificial General Intelligence (AGI). One example of this stance is an answer that Monica Anderson recently posted on Quora to the question "What are the main differences between Artificial Intelligence and Machine Learning?"[36] She wrote, in part:

> *"Machine Learning is the only kind of AI there is.*
>
> *"AI is changing. We are now recognizing that most things called "AI" in the past are nothing more than advanced programming tricks. As long as the programmer is the one supplying all the intelligence to the system by*

[34] Selmer Bringsjord, Konstantine Arkoudas, & Paul Bello, "Toward a General Logicist Methodology for Engineering Ethically Correct Robots", *IEEE Intelligent Systems*, July/August 2006, http://kryten.mm.rpi.edu/bringsjord_inference_robot_ethics_preprint.pdf accessed July 15, 2016

[35] Rensselaer Artificial Intelligence and Reasoning (RAIR) Laboratory, *Home page,* 2014, http://rair.cogsci.rpi.edu/ accessed July 15, 2016

[36] Monica Anderson, "What are the main differences between artificial intelligence and machine learning?", *Quora,* November 2015, https://www.quora.com/What-are-the-main-differences-between-artificial-intelligence-and-machine-learning/answer/Monica-Anderson accessed July 15, 2016

> *programming it in as a World Model, the system is not really an Artificial Intelligence. It's "just a program".*
>
> *"Don't model the World; Model the Mind.*
>
> *"When you Model the Mind you can create systems capable of Learning everything about the world. It is a much smaller task, since the world is very large and changes behind your back, which means World Models will become obsolete the moment they are made. The only hope to create intelligent systems is to have the system itself create and maintain its own World Models. Continuously, in response to sensory input.*
>
> *"Following this line of reasoning, Machine Learning is NOT a subset of AI. It really is the ONLY kind of AI there is."*

She closes her reply with the caveat,

> *"I really shouldn't confuse things but strictly speaking, Deep Learning is not AI either. We are currently using Supervised Deep Learning, which is another (but less critical) programmer's cheat since the "supervision" is a kind of World Model. Real AI requires Unsupervised Deep Learning. Many people including myself are working on this; it is possibly thousands of times more difficult that Supervised Learning. But this is where we have to go.*
>
> *"Deep Learning isn't AI but it's the only thing we have that's on the path to True AI."*

This is, in ways, similar to an argument that I made a few months ago in an essay titled, "AI, a 'Common Sense' Approach"[37]. By "common sense", I was referring not to the current plain language meaning of the phrase, but Aristotle's definition of the "common sense" as the human internal faculty that integrates the perceptions of the five senses—sight, hearing, touch, smell and taste—into a coherent view of the world. This is not accomplished through binary or formal logic, but rather through a system of pattern matching and learning mechanisms of the sort being explored in Machine Learning research.

Anderson and Anderson

Another team that takes a deontological approach, but one that is more consistent with a Machine Learning approach, is the husband and wife team of computer researcher Michael and philosopher Susan Leigh Anderson. (Susan's early work appears under her maiden name.)

[37] Jim Burrows, "AI, a "Common Sense" Approach", Eldacur Technologies, September 2015, http://www.eldacur.com/AI-a_Common_Sense_Approach-Revised-2016.03.21.pdf accessed July 15, 2016

A good introduction to their work can be found in the *Scientific American* article "Robot be Good"[38]. Another early roadmap article, "Machine Ethics: Creating an Ethical Intelligent Agent"[39], can be found in the Winter 2007 issue of *AI Magazine*, and there is a workshop paper entitled "*Prima Facie* Duty Approach to Machine Ethics and Its Application to Elder Care"[40] presented at the 2011 AAAI Conference on AI.

Whereas Bringsjord and company have advocated a deontological system of "Divine Command" logic (see the "Introducing Divine-Command Robot Ethics"[41] paper on their site or their chapter in the book, "Robot Ethics"[42]), the Andersons have adopted a "*prima facie* duty" system.

Their early work, which was at least partially successful, was based upon a consequentialist system of ethics, specifically Jeremy Bentham's "Hedonistic Act Utilitarianism". While they were successful, their work convinced them that a more complex system that relied upon and could be explained in terms of deontological principles was needed, and so they turned to W.D. Ross's notion of *prima facie* duties. Ross's theory is fundamentally deontological[43], although it may utilize some consequentialism-based principles.

Prima facie duty theory holds that there is not one unifying ethical principle such as Bentham's utilitarian principle or Kant's categorical imperative. Rather it relies on a number of duties that it holds to be real and self-evident (using "*prima facie"* to mean "obviously true on their face, when first seen"). In Ross's initial version there were 7 such duties, which he later reformulated as just 4.

[38] Susan Leigh Anderson & Michael Anderson, "Robot Be Good", *Scientific American #303,* October 2010, http://franz.com/success/customer_apps/artificial_intelligence/EthEl/robot-be-good.PDF accessed July 13, 2016

[39] Michael Anderson & Susan Leigh Anderson, "Machine Ethics: Creating an Ethical Intelligent Agent", *AI Magazine,* 2007, AAAI, http://ieet.org/archive/AIMagFinal.pdf accessed July 13, 2016

[40] Susan Leigh Anderson & Michael Anderson, "A Prima Facie Duty Approach to Machine Ethics and Its Application to Elder Care", *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence,* AAAI Publications, 2011, http://www.aaai.org/ocs/index.php/WS/AAAIW11/paper/view/3812/4274 accessed July 13, 2016

[41] Selmer Bringsjord & Joshua Taylor, "Introducing Divine-Command Robot Ethics", Tech. Rep. 062310, Rensselaer Polytechnic Institute, 2009, http://kryten.mm.rpi.edu/lrt_star_pc_062310.pdf accessed July 15, 2016

[42] Patrick Lin, Keith Abney & George A. Bekey, *Robot Ethics: The Ethical and Social Implications of Robotics*, The MIT Press, December 2011, https://mitpress.mit.edu/books/robot-ethics accessed July 15, 2016

[43] David L. Simpson, "Ross's Ethical Theory: Main Components and Principles", *"William David Ross (1877—1971)", Internet Encyclopedia of Philosophy*, http://www.iep.utm.edu/ross-wd/#H6 accessed July 13, 2016

Ross acknowledged that there are multiple duties, forcing us at times to determine their importance in each specific case, and prioritize them accordingly. In order to be useful for autonomous systems, a deciding principle and consistent methodology for choosing a course of action needed to be developed. The Andersons, utilizing John Rawls' "reflective equilibrium"[44] approach, designed, built and tested a number of experiments. These include the following systems:

- **MedEthEx** — a medical ethics advisor system
- **EthEl** — a medication reminder system, that can issue reminders and decide if a human overseer should be notified if the patient refuses to take their medication.
- A **Nao** robot instantiation of EthEl.
- **GenEth** — a Machine Learning expert system which can derive deciding principles for resolving ethical dilemmas based upon a collection of cases and the evaluation of experts as to their acceptability.

The GenEth generator can be used to derive the guiding principles from a set of test cases and evaluations of them by trained ethicists. These principles can then be fed into systems such as EthEl and loaded on a Nao robot or other system.

The description of events used in GenEth, EthEl etc. are tuples representing specific attributes and their impact on both "acceptable" and "unacceptable" judgements. As such, these systems are either lab curiosities or simply development tools for creating the sort of system that Monica Anderson (not related to Michael and Susan) called "just a program" in the Quora answer cited above. However, the Andersons hope to go further. They are already using Machine Learning to create the principle external to the robot that will use them.

The Andersons' work has, as can be seen above, all been in the area of medical ethics, and as such, it is important to them that they be able to create systems that not only behave properly, but also can explain *why* they did a certain thing, what principles they were following. This is clearly easier in a Logicist AI environment than in more biologically-derived Machine Learning paradigms, especially the sort of unsupervised deep learning that many expect to be the path towards full AGI. The Andersons' approach is phasing in Machine Learning. It will be interesting to see how far they can go down that path while maintaining the type of transparency that their current system provides.

In the end, the Andersons' work has become what you might call a "hybrid/hybrid" approach. In terms of the computer science and technology involved, the programming of the robots themselves takes a largely traditional Logicist AI approach, but

[44] Norman Daniels, "Reflective Equilibrium", *The Stanford Encyclopedia of Philosophy*, Winter 2013, http://plato.stanford.edu/archives/win2013/entries/reflective-equilibrium accessed July 15, 2016

supervised Machine Learning is certainly key to their methodology. Moreover, from an ethical perspective, they have chosen an approach that is firmly based in the duties of deontology, but which incorporates consequentialist aspects as well. In the next chapter, I will look at a more purely Consequentialist approach.

Which Ethics? — Consequentialism

Consequentialism refers to the general class of normative systems that determine the rightness or wrongness of an act not in accordance to the rules or principles that motivate or control it, the duties, obligations or virtues, but rather judge acts by their effects, their consequences. There are several types of Consequentialism, depending upon exactly how the value of those effects are measured: Act Utilitarianism, Hedonism or Egoism, Altruism, and so on. There are several good sources on Consequentialism available on the net:

- The Stanford Encyclopedia of Philosophy's Consequentialism article[45]
- The Internet Encyclopedia of Philosophy's Consequentialism article[46]
- The BBC's Ethics Guide article on Consequentialism[47]
- Wikipedia's Consequentialism article[48]

Most writers dealing with Consequentialist machine ethics deal in Jeremy Bentham's act utilitarianism, and specifically an altruistic version of it, given that the wellbeing of humans, not machines, are of consequence. Machines are, of course, valuable, so damaging them wastes resources, but it is that, and not the consequences to them, that factor in.

As noted in the previous chapter, the work of Michael and Susan Leigh Anderson dealt in its earliest stages with Bentham's Utilitarianism, but was soon shifted to *Prima Facie* Duty ethics, which is primarily a form of Deontology that bases some of its duties on Rule Utilitarianism or other Consequentialist theories.

Perhaps the most significant advocate for Consequentialism in machine ethics is Professor Alan Winfield, a professor of Electronic Engineering who works in Cognitive Robotics at the Bristol Robotics Laboratory in the UK. The approach that Prof. Winfield and his colleagues have taken has been variously called the Consequence Engine and an "Internal Model using a Consequence Evaluator". The most recent paper that I have describing this work is "Robots with internal models: A route to self-aware and hence safer robots". As can be seen from the following list of papers, their work has involved incorporating a number of biologically inspired approached into the development and evolution of small swarm robots.

[45] Walter Sinnott-Armstrong, "Consequentialism", *Stanford Encyclopedia of Philosophy*, Winter 2015, http://plato.stanford.edu/entries/consequentialism/ accessed July 18, 2016

[46] William Haines, "Consequentialism"*, Internet Encyclopedia of Philosophy*, http://www.iep.utm.edu/conseque/ accessed July 14, 2016

[47] The BBC, "Consequentialism", *Ethics guide,* 2014, http://www.bbc.co.uk/ethics/introduction/consequentialism_1.shtml accessed July 14, 2016

[48] "Consequentialism", *Wikipedia*, 2016, https://en.wikipedia.org/wiki/Consequentialism accessed July 18, 2016

Formal publications — Robot ethics

- Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection[49]
- Robots with internal models: A route to self-aware and hence safer robots[50]

Blog postings — Robot ethics[51]

- How ethical is your ethical robot?
- Towards ethical robots: an update
- Towards an Ethical Robot
- On internal models, consequence engines and Popperian creatures
- Ethical Robots: some technical and ethical challenges

Formal publications — Related topics[52]

- Evolvable robot hardware
- An artificial immune system for self-healing in swarm robotic systems
- Editorial: Special issue on ground robots operating in dynamic, unstructured and large-scale outdoor environments
- On the evolution of behaviours through embodied imitation
- Mobile GPGPU acceleration of embodied robot simulation
- Run-time detection of faults in autonomous mobile robots based on the comparison of simulated and real robot behavior
- A low-cost real-time tracking infrastructure for ground-based robot swarms
- The distributed co-evolution of an on-board simulator and controller for swarm robot behaviours
- Estimating the energy cost of (artificial) evolution

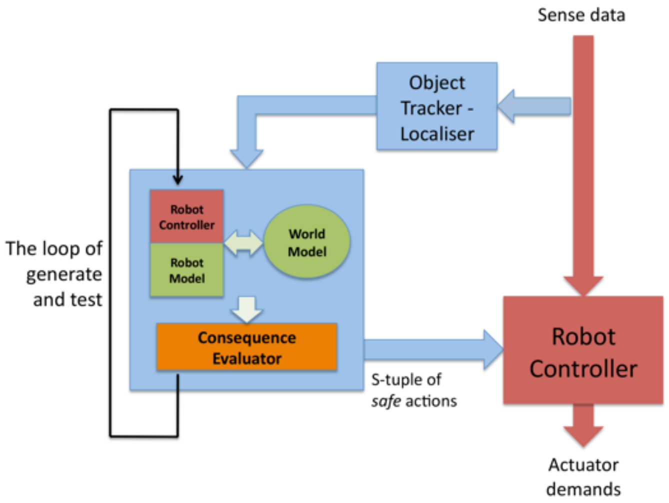
[49] Alan Winfield *et al.*, "Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection", *Proceedings of the twenty-ninth AAAI conference on artificial intelligence*, January 2015, https://www.aaai.org/ocs/index.php/WS/AAAIW15/paper/download/10119/10131 accessed July 17, 2016

[50] Alan Winfield, "Robots with internal models: A route to self-aware and hence safer robots", *The Computer After Me: Awareness And Self-Awareness In Autonomic Systems*, Imperial College Press, 2014, http://eprints.uwe.ac.uk/24467/ accessed July 17, 2016

[51] Alan Winfield, "Alan Winfield's Web Log", 2016, http://alanwinfield.blogspot.com/ accessed July 17, 2016

[52] Alan Winfield, "Professor Alan Winfield", *Staff Profiles*, University of the West of England, 2016, http://people.uwe.ac.uk/Pages/person.aspx?accountname=campus%5Ca-winfield accessed July 17, 2016

The Consequence Engine (CE), which is at the heart of their approach to machine ethics and safety, is a secondary control process that uses simulation to narrow the possible future actions available to the main control process. This diagram, taken from “Robots with internal models”, illustrates the role of the CE.



A Self-Aware Architecture for Safe Action-Selection. The Robot Control data flows are shown in red; the Internal Model data flows in blue.

In this architecture, the robot has an internal model of both the external world and of itself and its own actions (shown in blue). The robot’s main control system (shown in red) provides it (through a path not shown in the diagram) with a list of possible next actions. Each of these is simulated and its consequences are evaluated. In their design, the secondary system doesn’t choose which next action is to be performed, but rather winnows the list of potential next actions to those which are “safe” or ethically acceptable. It returns this list to the main control program, which selects the actual action to perform.

So far, the types of harm that their robots can simulate and avoid are simple physical accidents. Specifically, they have run simulations using two different generations of robots preventing a human (a role played in their tests by a second robot standing in for the human) from falling in a hole. Winfield in various talks admits that this is an extremely constrained environment, but he asserts that the principles involved can be expanded to include various sort of ethical harm (see his talk “Ethical robotics: Some technical and ethical challenges”[53] or the slides[54] from it):

- Unintended physical harm
- Unintended psychological harm
- Unintended socio-economic harm

As he notes, the big difficulty is that each of these is hard to predict in simulations. In order to make the system more sophisticated, one must add both more potential actions and more detail to the simulations, resulting in exponential growth. Even with the benefit of Moore’s Law, the limits of practicality are hit very quickly. Nonetheless, the Bristol team has demonstrated substantial early success.

Like Bringsjord and to an extent the Andersons, both of whom we covered in the previous chapter, Winfield comes from a traditional logicist tradition, and not the Model Free, Deep Learning school of Monica Anderson and many others. This means that it is engineers, programmers and architects who are responsible for the internal models, the Evaluator and the Controller logic. This only adds to the scalability issues.

Still, it does provide for one criterion that all of these researchers agree upon. Not only must a system behave in accordance with ethical principles or reasoning, but it must be able to explain why it did such and such, what the principles and reasoning involved are. Bringsjord is concerned with developing safe military and combat robots, the Andersons work in medical and healthcare applications, and Winfield started with simply making robots physically safe. Researchers in each of these areas, and in fact any academic involved in research, must be concerned with oversight from formal ethical review boards. If the principles, models, and reasoning are designed and implemented by human engineers, then building in an audit facility is easy. True Machine Learning approaches, such as Artificial Neural Nets (ANNs) and Deep Learning, may well involve the system itself creating the algorithms, models and logic. This makes it substantially harder to insure that a plain language explication of the reason for a certain action will be as readily available.

Winfield differs from Bringsjord, however, in being less devoted to the top-down logicist approach. As can be seen, especially in the “related topics” list of papers, the Bristol team is involved in genetic algorithms for software and hardware, robot culture, the role of

[53] Alan Winfield, “Ethical robotics: Some technical and ethical challenges”, *Fourth EUCogIII Members Conference on the Social and Ethical Aspects of Cognitive Systems*, 2013, https://youtu.be/McZQWFP3AoY accessed July 18, 2016

[54] Alan Winfield, “Ethical robotics: Some technical and ethical challenges”, *Alan Winfield’s Web Log*, 2013, http://www.alanwinfield.blogspot.co.uk/2013/10/ethical-robots-some-technical-and.html accessed July 18, 2016

forgetting and error, self-healing and various low-level biology-inspired methods. While he rejects deontology and virtue ethics, he does so on a purely pragmatic view, and not in accordance with a manifesto. He writes, ‘So from a practical point of view it seems that we can build a robot with consequentialist ethics, whereas it is much harder to think about how to build a robot with say Deontic ethics, or Virtue ethics.” If as the consequentialist complexities explode exponentially (or worse), Deep Learning or other bio-inspired approaches prove effective in deontological or virtue contexts, then, so long as there is a way to explicate the system’s reasoning, in keeping with the needs of ethics review and auditing, one can envision Winfield and his colleagues embracing what works.

In terms of the specific goals of this project, making Personified Systems that are worthy of trust, it is pretty clear that Winfield and the team in Bristol have demonstrated the potential of a Consequentialist system. Whether this system can grow with the technology as the Machine Learning and Deep Learning explosions progress, is less clear, but as with Bringsjord’s DCEC system, Winfield’s Consequence Engine certainly provides usable technology that we would be wise not to ignore.

Which Ethics? — Virtue Ethics

Having looked at deontological and consequentialist ethics, we now come to the third major class of normative ethical systems, and perhaps the oldest, Virtue Ethics. Whereas deontology and consequentialism provide rules for behavior based on principles and consequences, respectively, Virtue Ethics focuses on the character of the actor. It asks, and attempts to answer, "What sort of a person ought I to be, and how would such a person act?"

Shortly after I started this effort, I tentatively concluded that Virtue Ethics was a more promising approach to making "personified systems" trustable and worthy of trust. I then spent many weeks looking into which virtues, and how we might operationalize and implement them. In the last few months, I have been revisiting the suitability of the three classes of normative ethics. (See the two previous chapters.) In doing so, I found a nearly 25-year-old article by James Gips of Boston College entitled "Towards the Ethical Robot"[55] that gives an excellent overview of the whole subject.

After the inevitable Asimov Three Laws citation, Gips starts rather precisely with the issue that I have regarding personified systems, as our systems start behaving more like people: "[W]e want our robots to behave more like equals, more like ethical people". We have thousands of years interacting with other people, and have a detailed set of expectations as to how they should act; systems that act like people will inevitably be judged in the context of those expectations. This is what led me to ask what are those expectations, and how can non-intelligent systems that merely act and interact in human-like ways be made to live up to them?

I'd love to say that Gips agrees with much of what I have been doing and writing for the last several months, but given that he wrote it all a quarter of a century ago, I'm pretty much forced to say that it is *I* who agree with *him*. Having given an overview of Deontological and Consequentialist ethics, Gips writes

> *"On what type of ethical theory can automated ethical reasoning be based?*
>
> *"At first glance, consequentialist theories might seem the most "scientific", the most amenable to implementation in a robot. Maybe so, but there is a tremendous problem of measurement. How can one predict "pleasure", "happiness", or "well-being" in individuals in a way that is additive, or even comparable?*
>
> *"Deontological theories seem to offer more hope. The categorical imperative might be tough to implement in a reasoning system. But I think one could see*

[55] James Gips, "Towards the Ethical Robot", *Android Epistemology*, May 1991, http://www.cs.bc.edu/~gips/EthicalRobot.pdf accessed July 18, 2016

> *using a moral system like the one proposed by Gert as the basis for an automated ethical reasoning system. A difficult problem is in the resolution of conflicting obligations. Gert's impartial rational person advocating that violating the rule in these circumstances be publicly allowed seems reasonable but tough to implement.*
>
> *"The virtue-based approach to ethics, especially that of Aristotle, seems to resonate well with the modern connectionist approach to AI. Both seem to emphasize the immediate, the perceptual, the non-symbolic. Both emphasize development by training rather than by the teaching of abstract theory. Paul Churchland writes interestingly about moral knowledge and its development from a neurocomputational, connectionist point of view in "Moral Facts and Moral Knowledge", the final chapter of [Churchland 1989]."*

Since the systems of the day were not yet up to voice and face recognition, natural human-machine dialogs, driving cars, and so forth, he did not extend his thinking to merely "personified" systems, but I will add that in the case of programmer-driven systems, the design approach of software architects and implementers asking at each juncture, "What would a trustworthy system do?" or "What would be the candid (or discreet, etc.) thing for the system to do?" extends the value of virtue ethics to the sub-AI world.

Gips gives a good summary of various systems of virtues:

> *"Plato and other Greeks thought there are four cardinal virtues: wisdom, courage, temperance, and justice. They thought that from these primary virtues all other virtues can be derived. If one is wise and courageous and temperate and just then right actions will follow.*
>
> *"Aquinas thought the seven cardinal virtues are faith, hope, love, prudence, fortitude, temperance, and justice. The first three are "theological" virtues, the final four "human" virtues.*
>
> *"For Schopenhauer there are two cardinal virtues: benevolence and justice."*

To this I would add what is often taken in pop culture to be the exemplar of virtue, the Eagle Scout, who according to the twelve points of the Scout's Law is summed up as follows:

> *"A scout is trustworthy, loyal, helpful, friendly, courteous, kind, obedient, cheerful, thrifty, brave, clean, and reverent."*

While they do not write specifically about "virtues", the adherents to the "Moral Foundations" school of thinking started by social psychologist Jonathan Haidt, explain human morality in terms of the evolution of five or six underlying “foundations”—psychological values or judgments[56]:

- **Care/harm** — Underlies virtues of kindness, gentleness, and nurturance.
- **Fairness/cheating** — Generates ideas of justice, rights, and autonomy.
- **Loyalty/betrayal** — Underlies virtues of patriotism and self-sacrifice for the group.
- **Authority/subversion** — Underlies virtues of leadership and followership, including deference to legitimate authority and respect for traditions.
- **Sanctity/degradation** — Underlies the widespread idea that the body is a temple which can be desecrated by immoral activities and contaminants.

And tentatively:

- **Liberty/oppression**, which they note is in tension with the authority foundation.

It is fairly easy to see how each of these ties to the virtues of the Eagle Scout, for instance.

One thing worth noting is that Aristotle conceived of virtue as a middle ground between two vices, one of excess and the other of deficiency. Thus for him courage is the mean between recklessness and cowardice, generosity between wasteful excess and stinginess, and so forth. On the other hand, many such as the Moral Foundations theorists see the world in more black and white, good and bad, dichotomies. This idea of virtues as a Golden Mean is reflected in many ethical systems, such as Taoism, Buddhism’s Middle Way, Confucius’ Doctrine of the Mean, and so forth. On the other hand dualistic ethics has dominated the Abrahamic and other Middle Eastern religions such as Zoroastrianism.

Gips’ observation that Virtue Ethics “seems to resonate well with the modern connectionist approach to AI” seems particularly pertinent today given the recent explosive growth in Machine Learning technologies. This leads us to what may be the major shortcoming of a Virtue approach to machine ethics: the area of accountability, that is of the system’s ability to explain why it took certain actions.

The rules regulating a Deontological system that uses a language such as Bringsjord’s DCEC can readily be mapped to English, as can the specific consequences that caused Winfield’s Consequence engine to reject a specific alternative. However, given the opacity of the reasoning created by sophisticated Deep Learning systems, it may well be that the attributes that Gips cites as matching connectionist or ML, that it “emphasize[s] the immediate, the

[56] MoralFoundations.org, *web site,* the YourMorals.org collaboration, 2016, http://moralfoundations.org/ accessed July 18, 2016

perceptual, the non-symbolic [and] development by training rather than by the teaching of abstract theory" run substantial risk of making it hard for the system to explain its actions.

Still, it seems to me that Virtue might be amenable to, for instance, the ML and principle-generating methodology of the Andersons' GenEth process, as discussed in the "Which Ethics? — Deontology" chapter. GenEth describes various events in terms of features which are present or absent to varying degrees, measured as numbers that range, for instance, from -2 to +2. Its intent is for professional ethicists to be able to train the system by giving a number of cases and their judgement as to the proper course. An analogous approach that allows the system to recognize the applicability of the various virtues to situations would seem to make sense.

Given all of this, I conclude that Virtue ethics is applicable from the least intelligent personified systems up through to the hypothetical future AGIs sophisticated enough to act as AMAs. But, even, more, I can see arguments for believing that a hybrid system, combining Virtue ethics with some of the best work being done in Deontological and Consequentialist Machine Ethics, could be extremely powerful. I will address this idea in my next couple of chapters.

Which Ethics? — Pulling it all together

What type of normative ethics is best suited to Personified Systems, systems that are acting more and more like people? What kind of ethics works for simple software with more human user interfaces? For sophisticated AIs, and for the ultimate Artificial General AIs of the future?

In the last three chapters, I've given an overview of what's already been done and written about in terms of the three main categories of normative ethics (Deontology, Consequentialism and Virtue ethics) and how each might apply to the systems of the future. The question now is which of these or which combination of them should we use.

Each of the efforts that we've examined in these chapters provides important tools and insights. Bringsjord and his team at RPI provide us with a formal language, their Deontic Cognitive Event Calculus (DCEC, see the DCEC Formal Specification[57] and "On Deep Computational Formalization of Natural Language"[58]), that allows the expression of propositions about knowledge, beliefs, obligations and so forth.

The Andersons with their GenEth provide a Machine Learning-based mechanism for analyzing and generalizing the reasoning of professional ethicists regarding ethical dilemma test cases. Both of these systems allow the principles that drive logicist programmer-written software systems to be expressed both formally and in clear English translations, which is important as an aid to accountability.

Winfield's Consequence Engine architecture offers an on-board decision-making approach: its duplicate-system control process operates in an internal simulation of the system and the world around it and evaluates the results so as to preclude unacceptable actions. (See "Towards an Ethical Robot"[59].) This contrasts with the approaches of Bringsjord and the

[57] Selmer Bringsjord et al, "Deontic Cognitive Event Calculus (Formal Specification)", Rensselaer Artificial Intelligence and Reasoning Laboratory, May, 2013, http://www.cs.rpi.edu/~govinn/dcec.pdf accessed July 13, 2016

[58] Naveen Sundar Govindarajulu et al, "On Deep Computational Formalization of Natural Language", Rensselaer Artificial Intelligence and Reasoning Laboratory, 2013, http://kryten.mm.rpi.edu/SELPAP/2013.FormalMagic/main.pdf accessed July 13, 2016

[59] Alan Winfield, "Towards an Ethical Robot", *Alan Winfield's Web Log*, 2013, http://alanwinfield.blogspot.com/2014/08/on-internal-models-part-2-ethical-robot.html accessed July 18, 2016

Andersons, both of which externalize the ethical reasoning (See Winfield's blog post regarding "Popperian creatures"[60]).

Bringsjord's ethics are strictly deontological, and Winfield's consequentialist. The Andersons' use of Ross's *prima facie* duties begins to bring these traditions together. It uses a systematic approach to developing and adopting a deciding principle for prioritizing and selecting among the applicable duties. Some of these duties may be more consequentialist rather than purely deontological.

If deontological rules, *prima facie* duties, and consequences are to be laid out explicitly, then it would appear to be unavoidable that the body of such rules will become huge. A given sophisticated system might operate with thousands, likely many thousands, of duties, obligations, rules, etc. An advantage of Virtue ethics might well be to offer an organizing principle for these rules.

If each learned or generated rule is created in the context of a ruling virtue, then the system might be able to give explanations of its choices and actions by saying that it was motivated by "loyalty, specifically, a rule arising from the case where…" and citing the training test case or cases that were the major contributors to a specific learned pattern. I do not claim that tracking such information will be easy, but explaining behavior based on ML-generated (especially Deep Learning) patterns is inherently difficult in and of itself, and so if the language of virtues can help to organize the ethical reasoning, it would be a great help.

How this is all accomplished depends on the architecture and developmental methodologies used to create specific systems. At one end of the spectrum, we have systems that are created using traditional software engineering methodologies, with human programmers making the decisions and coding the behaviors into the system's software. For them, the virtues provide a way to organize their thinking as they approach each design or implementation decision. They may ask themselves "To whom is the system being loyal at this time? How are the user's, society's and our (the manufacturer's) interests being traded off?" or "Is the user's information being handled with discretion at this point?" or "Are we being candid here?" and, in general, "What would a virtuous system do?"

At the other end of the spectrum, one can envision a sophisticated AGI guided by what amounts to an artificial conscience, trained by professional ethicists, evaluating its actions and choices according to internalized principles that are organized and explicable according to a system of virtues and specific duties. The capabilities of current state of the art systems

[60] Alan Winfield, "On internal models, consequence engines and Popperian creatures", *Alan Winfield's Web Log*, 2013, http://alanwinfield.blogspot.com/2014/07/on-internal-models-consequence-engines.html accessed July 18, 2016

are somewhere in between, and can draw upon the theories and practices of all of the efforts described in the last few chapters.

When I started this effort, I was focused almost entirely upon the role and function of virtue ethics in guiding and improving the behavior of personified systems. As I have been evaluating the possibility of actually including deontological and consequentialist principles and *prima facie* duties, a more complex picture has begun to emerge. I continue to think, as Gips did a quarter century ago, that Virtue Ethics parallels the workings of connectionist models of natural cognition and the latest developments in machine learning, and can serve a very important function. I am even beginning to see what may be the gross outline of an approach to integrating normative ethics into the workings of future full AGIs such that they might be able to act as autonomous moral agents. I will take this up in the next chapter.

For now, let us conclude that the answer to “Which ethics?” is “a hybrid approach featuring *prima facie* duties based upon both deontological and consequentialist principles, organized according to principles derived from the decisions of professional ethicists, and reflecting a set of core virtues”. While it is not a simple answer, two or three millennia of philosophy have taught us nothing if not that ethics is a complex matter. The good news is that the precision and formality required to incorporate anything into an automated system forces us to clarify our thinking, and approaching machine ethics may aid us in refining our own human ethics.

Which Virtues?

> *When I first started this project, several months ago, I fairly immediately focused upon virtue ethics as my approach to making personified systems behave in a trustworthy way. Today, my preferred approach is something more along the lines of* prima facie *duties, using a small set of virtues to provide an organizing principle as the suite of duties grows. Either approach, though, demands that we select an appropriate set of virtues.*

Introduction

The precise set of virtues that is appropriate for a particular personified or artificially intelligent system, and the relationship of the virtues to each other, must, of course, be based upon the exact nature of the system in question. The virtues of a personal assistant may need to vary considerably from those of a medical care-giving system or one designed for war. In this chapter, I will concentrate primarily on systems that act as assistants and companion systems with only a limited degree of autonomy, and that are not responsible for the life and death of human beings. Robot doctors, nurses, and war fighters are well outside our purview, as are full-fledged drivers. Those systems require specialized ethical systems and extended oversight.

Based upon a number of systems of human virtues, from the Scout's law to a survey of the attributes required for a butler or valet as described on websites devoted to those professions, I've identified eight virtues of a general-purpose trustworthy personified system. Trustworthiness itself may be regarded as a ninth over-arching virtue. The eight come in two groups and are as follows:

1. Helpfulness
2. Obedience
3. Friendliness
4. Courtesy

5. Loyalty
6. Candor
7. Discretion
8. Propriety

The first four virtues in the list are what I have been regarding as "utilitarian" or "functional" virtues. An assistant that is helpful, obedient, friendly, and courteous is more useful or easier to use. They map fairly directly to established UI and UX practices.

The next four, I consider broadly as “ethical” virtues. A system that is loyal, forthright, discreet, and in general trustworthy, comes as close as can be attained without full intelligence to behaving ethically, or perhaps “properly”. In this chapter, I will focus on Trustworthiness and its four “ethical” subordinate virtues, laying out how they are defined both philosophically and operationally in a general sense. To actually implement them for a specific application, class or family of applications would require a much more specific and detailed operationalization than I can manage here. Still, it is important to not only understand the role of these virtues in human ethics, but their role in the operation of technological systems.

As I was researching, an additional attribute (*Propriety*) emerged as important, though exactly how is still debatable. This is the question of what degree of emotional involvement such a system should be geared for. At the extreme, it is the question of whether humans should be allowed or encouraged to love the system’s persona. Should these systems attempt to retain a degree of professional distance, or should they seek friendship and emotional involvement? I am referring to this as the virtue of “propriety” for the nonce.

Loyalty

When we speak of loyalty, there are two aspects: “whose interests are being served by the actions of the system?” and “how consistent is the answer to the first question?” Different systems will each have potentially different sets of loyalties. For both personified systems and human beings, loyalty is not a simple one-dimensional attribute; rather, each of us is loyal to family, friends, employer, nation, and so on. Loyalty, operationally, is a matter of priorities. Do the interests of one party or group outweigh those of another?

Eric Horvitz’s personal assistant, Monica, which interacts with people at the door to his office, could be loyal to him, to his employer, or to the vendor that provided the system. In this case, the employer and the vendor are both Microsoft, but when the system goes commercial and John Wylie at Acme Services purchases one, it will make a substantial difference to him if it is loyal to Acme or Microsoft, or shifts inconsistently between them. Given that his assistant will have access to aspects of both John’s personal and work behavior, it is important for him to know Monica works for Acme, and not for him personally, and it will presumably be important to Acme that she is working for them and not for Microsoft.

Likewise, when George and Mary obtain a virtual nanny to look after their children, will they be told “I’m sorry, but Wendy and the boys have requested privacy” or will Nana spy on the children? How about when they ask the virtual caretaker that they bought for George’s aging mother? Does it answer to them or Grandmère? Does it matter if they just leased the system? Does that make it loyal to the company rather than the parent or the child?

What if Wendy and the boys requested privacy to smoke pot? Should Nana rat them out to their parents or to the law, or keep their secret? If Grandmère is indulging in the cooking sherry, should George and Mary be told? Should her doctor know, in case it is contraindicated given her medicine? How about the insurance company that paid for the system?

Liability, ownership, and the primary "user"/"care recipient", etc. are all factors in deciding where the system's loyalties belong, and potentially in deciding on a hierarchy of loyalties, or reserved privileges.

As autonomous collision avoidance becomes more sophisticated and prevalent in our autos, they will inevitably be faced with making hard choices and tradeoffs. Does the system primarily strive to protect its passengers and secondarily minimize the harm caused to others? Or does it willingly sacrifice the safety of the passengers when doing so protects a larger number of bystanders? Would a privately owned auto behave differently from a commercial limousine, or a municipal bus? Does the action of an autonomous long-haul truck depend upon the nature and value of the cargo? In short, is the auto driver loyal to its passengers, owner, manufacturer, insurer, or society as a whole?

It is worth noting that, to the extent that software engineers or architects build in explicit trade-offs in advance for dealing with emergencies, they are faced with an increased responsibility and liability as compared with a human driver responding in the heat of the moment. In the fraction of a second that a driver faces a disaster and reacts, perhaps with little or no conscious thought, we generally don't hold them fully responsible for the decision to veer right or left. However, if a programer creates an explicit set of priorities, or a QA engineer approves the shipment of an AI that has learned or derived a specific set of priorities, then those human acts are made deliberately and with time to weigh and choose the consequences.

This means that the introduction of personified and semi-autonomous systems actually introduces issues of responsibility and liability beyond those that would apply if an unassisted human were solely involved. How this additional burden of responsibility is handled is unknown at present and will remain so until laws are passed and precedents are set in our courts. Thus the legal and economic pressures around determining the loyalties of personified systems and the AGIs that follow after them will be in flux for the foreseeable future.

Designers and implementors of personified systems ought to identify one or more explicit hierarchies for their system and then stick with them, evaluating at each major decision point how the priorities come into play. I say "one or more" because different customers may want different sets of loyalties and priorities or want to choose between them. Offering a feature that allows for the customization of priorities is, of course, quite a lot more work.

The complexity of the issue of loyalty leads to the next virtue.

Candor

The virtue of candor in a personified system may be considered an elaboration of the property of “transparency” that we are used to in discussing software and business in general. A candid system is one that makes its loyalties and its actions clear to the parties involved: its client, patient, boss, owner, or the like. Someone being served by a candid personified system should be aware of what that system’s priorities and loyalties are and what actions the system has taken, in general. It need not report every minor routine action, but it should insure that the person(s) served know, in general, the sorts of actions it routinely performs, who controls it and sets its policies and priorities, and what those priorities are. It should not be evasive if additional details are requested.

Conflicting loyalties may well be inevitable, as noted above. As the realities of liability surrounding autonomous agents develop, manufacturers are likely to reserve certain privileges. Similarly, in the case of care-taking systems that are provided through health or other insurance, the insurer or other agency that pays for or provides the system may reserve some rights or demand certain priorities and limitations. Personal assistants may be constrained or controlled by the individual user, their employer, or the vendor providing the system, and the cloud-based or other services that are used to power it.

These ambiguities and conflicts in system loyalty will be tolerable only if the user is clearly aware of them. In other words, the system must be candid in revealing its loyalties, priorities, limitations, capabilities, and actions.

In considering the virtues and behaviors of hypothetical fully intelligent or moral agents, one of the potential virtues is honesty. For the purposes of the present effort, which is limited to the “virtues” and behaviors of more limited personified agents, I have lumped honesty and candor into a single category. True dishonesty requires intention. A lie is not merely a falsehood, but an intentional and deceptive falsehood. Pre-intelligent systems do not possess true intent. As such, I am subsuming “honesty” into “candor” in this discussion. For true AGIs, they might well need to be separate.

An important aspect of candor is taking an effort to be understood, and not merely to recite information. A candid system should speak simply and briefly in plain language, allowing or encouraging requests for further explanations, elaborations, or details. Reading the system’s terms of service or privacy policy aloud is not, in fact, particularly informative. Responding with a simplified concise plain-language summary, and asking whether more details are required, would be much better and more candid.

Developers and designers should be asking, “Have I made it clear to the user what to expect?” not merely with regard to questions of loyalties, but regarding the behavior of the system as a whole.

Discretion

The need for discretion in digital virtual assistants is underscored by the following tension. On the one hand, more and more of our confidential information is making its way into our various computers and mobile devices, and we need to protect the most sensitive information about us from leaking out. On the other hand, the phrase “information economy” is becoming more and more literally true. We have come to view access to network and computer services as something that we do not pay for with money, but rather with information about us and our habits. More than that, what was once mere data has become information by being collected, correlated, and analyzed, and as we begin talking to personalized systems with voice recognition, that information is being parsed and analyzed semantically to the point where it represents real knowledge about us.

In the typical voice-driven system like Siri, Cortana, Google Now, or Amazon Echo, the attention phrase that allows the system to know we are talking to it, (“Hey Siri”, “Cortana”, “OK, Google” or “Alexa”, respectively), is recognized locally in the device, but the commands and queries addressed to the system are sent to a remote server in the cloud where they are analyzed semantically, using general and specific grammars, the user’s contacts and other information, and the context of recent commands and queries. The system identifies the actual subject matter being spoken about and uses that to distinguish which among a number of similar-sounding words is intended. Transforming simple audio data into parsed and analyzed semantic knowledge, with meaning, makes the information that much more valuable. And access to it that much more intrusive.

On the other horn of our dilemma, accessing services without paying for them, either monetarily or with information, is a failure to participate in the economy of the Internet, at best, and theft or its moral equivalent at worst. The implicit deal is that systems provide us information, entertainment, connectivity, and the like, and we pay for it with information about ourselves. If we refuse to provide any information, we are pulling out of, or leeching off of, the economy, but if we pay for access and services with our most vital secrets, we are taking huge risks.

A discreet real life butler, faced with this situation, would protect our secrets, confidences, family matters, and confidential business information, and pay for services with the most innocuous bits of information, and those directly involved with specifying and obtaining the desired services. He would be frugal with the master's money and information, exchanging it only for true received value, and with an understanding of each bit's value and risks.

While the local system is unlikely to be anywhere near as good at analyzing our utterances, commands, and queries as the remote servers, it can do some analysis and inferences, and we can label questions and other information. It can also ask for clarification of the degree of confidentiality. One can readily imagine saying, "Jeeves, make discreet inquiries into the cost of …" or, using different attention words for different assistants that handle different aspects of our lives and so forth. Creating more intelligent systems capable of an amount of discretion should be possible.

Discretion can, and should, be combined with candor. A superior system would be one that not only makes intelligent distinctions between the confidential and the trivial, but should allow us to know the distinctions and priorities that it is using.

Propriety

While the first three constituent virtues of trustworthiness—loyalty, candor, and discretion—are intimately intertwined, the fourth is a bit more independent and complex. I am currently calling it "propriety".

At first blush, it would seem that one of the desirable characteristics of a personified system is for it to be emotionally engaging—friendly or even lovable. This seems to be a natural outgrowth of creating "user friendly", approachable, and generally easy-to-use software. It is certainly hard to see any virtue in creating repulsive, hostile, or hard-to-use systems.

Our fiction is full of lovable robots and artificial people from Tik-Tok of Oz to Astro Boy, the Jetsons' Rosie, Doctor Who's K-9, David and Teddy of "A.I.", Baymax of "Big Hero 6", and Wall-E among so many others. Then, of course, there are the sexy robots from Maria of "Metropolis", to Rhoda of "My Living Doll", to Ava of "Ex Machina". Many of Hollywood's sexiest actresses have played robots, "gynoids", and "fembots".

However, upon closer consideration, being emotionally appealing, engaging, and even seductive may not be such positive characteristics as they seem at first. Rather, it seems wiser for systems to maintain a bit of emotional distance of the sort that we associate with a "professional demeanor", for a variety of reasons, depending upon the exact role that it performs. Matthias Scheutz[61] and Blay Whitby[62] each discuss some of the negative aspects of too much emotional attachment in their chapters of "Robot Ethics".

[61] Matthias Scheutz, "The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots", *Robot Ethics: The Ethical and Social Implications of Robotics*, The MIT Press, December 2011, https://mitpress.mit.edu/books/robot-ethics accessed July 15, 2016

[62] Blay Whitby, "Do You Want a Robot Lover? The Ethics of Caring Technologies", *ibid.*

In part, this amounts to nothing more than applying rules that already guide interactions among humans. For example: military and other hierarchical organizations have non-fraternization regulations; medical ethics discourages doctors from treating family members and friends; most care-giving professions have rules against romantic, sexual, and other intimate relationships between care-givers and their charges.

Children offer a particularly sensitive area. On the one hand, it would be highly undesirable for interactions between a child and a personified system to compete with real people for a child's affection and attention. Children need to develop people skills, and that means dealing with real people. An automated system might be able to offer some support to a child in that endeavor, but it should not supplant other people. On the other hand, it it is quite conceivable that very shy children or those on the autism spectrum might be able to use the simpler and less threatening relationships with personified systems as a stepping stone to more complex relationships with other people.

This applies not merely to children but to anyone that these systems interact with. Except in very special circumstances—autistic children who need an "almost human" to practice on, a prisoner denied the company of folks from the outside world, and so on—artificial people are poor substitutes for real ones, and they should not replace genuine emotional bonds.

Even in these exceptions there is a risk. For instance, when cautioning against substituting emotional attachments to automata in *Robot Ethics*, Blay Whitby wrote:

> *For example, the category of people who most obviously are considered unfit for human society is convicted violent criminals. They would seem an ideal target market for robot lovers. Interestingly, they are a market rarely, if ever, mentioned by the enthusiasts for technology.*

But suppose that violent convicts were permitted robot lovers. Given that people have been known to see Roombas as personified, and develop deep emotional ties to bomb disposal robots, is there not a real possibility of such convicts developing strong ties to their particular automata? Once they are released, could such ties make them more willing to commit crimes whose punishment is merely to return to that private and intimate relationship? While most might not, still the possibility of such a phenomenon, and its effects on recidivism, must be admitted.

As we noted in the discussions of discretion, it is important that intimate and confidential information not be traded or given away by the system. One of the best ways to protect against that is to not encourage the human user to confide too much in the system. Likewise, a candid system should not pretend to be what it is not, and so propriety can be seen in part as merely as an act of candor, not pretending that it can return the human's feelings in kind.

All of this leads us to the concept of "propriety". A personified system should strike a balance—use an appropriate degree of emotional involvement. It should strive to be neither so cold and distant as to be unapproachable and difficult to interact with, nor to give a deceptive appearance of mutual emotional involvement that cannot be delivered. It should maintain a pleasant, polite, and positive relationship with its user, but tempered with the cool and professional demeanor of a human butler or valet, a therapist or care giver, a nanny or teacher, rather than trying to be the user's friend, lover, or confidant.

Trustworthiness

Pulling these virtues together, we can build a picture of trustworthiness. A good personified assistant should behave professionally, fitting into its environment appropriately, should have explicit loyalties, be candid about them and act upon them with discretion and propriety. Each of these can be viewed as duties. Thus, one can see how this dictum might lend itself to an implementation modeled after the *prima facie* duties of the Andersons[63] and the notion of an explicit deciding principle to guide the tradeoffs between them.

Candor requires some mechanism by which the system can, among other things, explain its decision-making process and principles. Several of the systems explored in the last few chapters provide mechanisms to enable this. Bringsjord[64] has demonstrated DCEC-to-English as well as English-to-DCEC translation. The Andersons' GenEth system for deriving principles from cases studies also results in explicit rules that can be laid out using the language supplied to describe the test cases. Finally Winfield's consequence engine[65] can cite specific consequences that caused actions to be rejected. The trick with all of these will be simplifying the explanations down to a level that is commensurate with true candor.

Aristotle's view of virtues is that they represented a mean between two extremes, and that is a theme that comes up here as well. A candid explanation, for instance must be balanced between overwhelming detail that could mask falsehoods and unpleasant truths, and a brevity that carries insufficient detail needed for understanding. Participating discretely in the information economy again balances disclosing too much and too little, and propriety requires balance between coldness and an illusion of impossible human connection.

[63] *see "Which Ethics? — Deontology"*

[64] *see "Which Ethics? — Deontology" again*

[65] *see "Which Ethics? — Consequentialism"*

Conclusions

In the course of this work, I've drawn a number of conclusions:

1. Even at fairly low levels of sophistication, current AI techniques are capable of making decisions based upon ethical considerations. Both deontological and consequentialist approaches have been demonstrated.
2. Creating a mechanism that is responsible for making the normative value judgements that is separate from the main control system is a workable model. This module can either filter out unacceptable future actions from the list of those to be chosen amongst or prioritize that list.
3. Perhaps the most promising normative ethical context for such systems is a flexible system such as W.D. Ross's *prima facie* duties, which allows for a hierarchy of principles and a decision mechanism for prioritizing them.
4. AIs can successfully use the solutions of trained ethicists to ethical dilemmas as a basis for learning to recognize principles.
5. Virtues could serve as an organizing principle for simplifying the complexity of calculating the suitability of an ever-growing number of potential actions against a growing number of principles.
6. Virtues can also serve as guiding principles during the creation of the programs of less autonomous personified systems and more logicist AIs.
7. I recommend a set of four "functional" and four "ethical" virtues under an umbrella of Trustworthiness:
    - Functional virtues
        - Helpfulness
        - Obedience
        - Friendliness
        - Courtesy
    - The ethical virtues of "Trustworthy Systems"
        - Loyalty
        - Candor
        - Discretion
        - Propriety
8. I've provided a number of examples of how these virtues can be operationalized.
9. All of this opens major opportunities for future research.