Devitt, S.K. (2017). Trustworthiness of autonomous systems. In H. Abbass, J. Scholtz & D. Reid (eds.) *Foundations of Trusted Autonomous Systems.* Springer [Accepted-under review]

Trustworthiness of Autonomous Systems

Abstract

Effective robots and autonomous systems must be trustworthy. This chapter examines models of trustworthiness from a philosophical and empirical perspective to inform the design and adoption of autonomous systems. Trustworthiness is a property of trusted agents or organisations that engenders trust in other agent or organisations. Trust is a complex phenomena defined differently depending on the discipline. This chapter aims to bring different approaches under a single framework for investigation with three sorts of questions: Who or what is trustworthy?—metaphysics. How do we know who or what is trustworthy?—epistemology. What factors influence what or who should we trust?—normativity. A two-component model of trust is used that incorporates competence (skills, reliability and experience) and integrity (motives, honesty and character). It is supposed that human levels of competence yield the highest trust whereas trust is reduced at sub-human and super-human levels. The threshold for trustworthiness of an agent or organisation in a particular context is a function of their relationship with the truster and potential impacts of decisions. Building trustworthy autonomous systems requires obeying the norms of logic, rationality and ethics under pragmatic constraints—even though there is disagreement on these principles by experts. Autonomous systems may need sophisticated social identities including empathy and reputational concerns to build human-like trust relationships. Ultimately transdisciplinary research drawing on metaphysical, epistemological and normative human and machine theories of trust are needed to design trustworthy autonomous systems for adoption.

1. Introduction

Humans are constantly engaged in evaluating the trustworthiness of humans and systems. Effective robots and Autonomous Systems (AS) must be trustworthy. Understanding how humans trust will enable better relationships between human and AS. Trust is essential in designing autonomous and semi-autonomous technologies, because "No trust, no use" (Schaefer, Chen, Szalma, & Hancock, 2016). Additionally, rates of usage are proportionally related to the degree of trust expressed (Lee & See, 2004). Hancock, Billings & Schaefer (2011) argue that trust begets reliance, compliance and use. However, humans do already rely on systems they do not trust. Consider the unreasonable privacy policies agreed to by users to access services via apps, websites and cloud services (Steinfeld, 2016). Because privacy policies can be changed at any time, private data may be sold by organisations for profit without explicit consumer consent or even awareness. Consumers can find the benefits of the services to enhance their lives and productivity too strong to resist. In these situations,

Email: kate.devitt@qut.edu.au

people rely on systems they do not trust and are not trustworthy. People know that their data may be shared for corporate interests. People know that they have signed away rights on their own images, etc.. by using these services. As more services operate without human decision makers yet offer irresistible perks, humans may increasingly rely on untrusted AS to decide for them. Instead of trust, it may be better to consider human reliance on other humans and systems as a measure of risk aversion—of which trustworthiness remains a significant part. This chapter considers the trustworthiness of autonomous systems by examining models of trustworthiness from both a philosophical and empirical perspectives to inform the design and adoption of autonomous systems. The chapter connects and augments research of Lewis, Sycara & Walker's in Chapter 8 that also explores concepts of trust, models of trust and factors affecting trust.

1.1. Autonomous Systems

AS can be robots, AI programs or software that operate without human control. AS are made by teams of engineers, designers, mathematicians, and computer programmers to serve a human need. AS actions and decisions are made by complex hierarchical processes balancing the uncertainties of cross modal inputs such as cameras, microphones, tactile responders with internal representations such as maps, directives and event memories. AS execute functions such as actively selecting data, transforming information, making decisions, or controlling processes without inputs (Lee & See, 2004, p. 50). AS are defined in contrast with automated systems and manual systems. Automated systems are largely deterministic to achieve predefined goals. Classic automata such as Japanese karakuri demonstrate complicated, nevertheless predictable behaviours (Šabanović, 2014). In contrast, AS learn and adapt in their environments rendering their actions more indeterminate over time (Hancock, 2016; Schaefer et al., 2016). Advanced AS may be capable of executive functions such as planning, goal-setting, rule-making and abstract conceptualisation. An ‘autonomous system’ can refer to a subset of functions within a larger functional system or refer to the superset of functions undertaken by an agent or machine. Regardless of the scope of functions of an autonomous system, it is important that AS operate without human control.

1.2. Trustworthiness

Trustworthiness is a property of an agent or organisation that engenders trust in another agent or organisation. Trust is a psychological state in which a person makes themselves vulnerable because they are confident that other agents will not exploit them (Nave, Camerer, & McCullough, 2015). Trust is also a social feeling of mutual confidence that increases the efficiency of systems, allowing adaptations to externalities and uncertainties (Arrow, 1974). Trust, like empathy, truth telling and loyalty lubricates social interactions. Humans depend on flexible cooperation with unrelated group members that rely on trust (Sterelny, 2012). Thus, social success relies both the evaluation of the trustworthiness of others and the presentation of oneself as trustworthy (Engelman, 2014).

We can distinguish between the trust we place in individuals, and the general trust we have in our society that affects how we make decisions more broadly, e.g. Adam Smith (1776) in *the Wealth of Nations* noted that a merchant is more comfortable trading within their own society because they can "know better the character and situation of the persons whom he trusts." Empirical literature has linked improved trust with more efficient public institutions, greater economic prosperity, self-reported health and happiness across many societies using a range of statistical techniques (see Carl & Billari, 2014). Within a Nation or society, trust is quite heterogeneous between individuals. Surveys on whether subjects trust a generic person (measured on a scale between 0 (no trust at all) and 10 (fully trusted) find large interpersonal differences. (Butler, Giuliano, & Guiso, 2009). Economic productivity peaks when the average citizen rates a generic person a '7' level of trust—a fairly high level of trust. Pessimists trust too little and give up opportunities too often. Optimists trust too much and get cheated more frequently. How does this trust research relate to AS? Do economic models apply to designing trustworthiness in AS? Should we create trustworthy systems to engender a '7' level of trust matching optimum human economic performance? That is to say, if we test the trustworthiness of autonomous-human interactions, should we aim to replicate the trust metrics found between people or some other measure?

It is important to acknowledge that trust is a complex phenomena and has been defined differently depending on the discipline (Rousseau, Sitkin, Burt, & Camerer, 1998). Economists consider it calculative (Williamson, 1993) or institutional (North, 1990). Psychologists focus on the cognitive attributes of the trustor and the trustee (Rotter, 1967; Tyler, 2006). Sociologists find trust within human relationships (Granovetter, 1985). Understanding the way humans conceive of and act regarding trust is critical to ensure the success of trusted AS. To bring different approaches under a single framework for investigation, this chapter will examine trustworthiness with three questions:

1. Who or what is trustworthy?—metaphysics
2. How do we know who or what is trustworthy?—epistemology
3. What factors influence what or who should we trust?—normativity

Building trustworthy autonomous systems requires understanding trust in human-human relationships and human-AS interactions. A research program on trusted AS ought to incorporate mental models informed from cognitive science to better understand and respond to human thoughts and behaviour. An example of such a research program is the recent work programming a robot with ACT-R/E (Khemlani, Harrison, & Trafton, 2016; Trafton et al., 2013), an embodied extension of the ACT-R (Anderson, 2007) cognitive architecture. The ACT-R/E implementation takes features of human cognition, such as segmenting time into events and narrative explanation to bring meaningfulness and trust to robot-human relationships. But, it is just one of many promising frameworks to align AS with human cognition. This chapter considers a range of theories of trust to influence the design trustworthy autonomous systems.

2. Background

The Fukushima Daiichi nuclear power plant disaster stemming from the Japanese earthquake and tsunami in March 2011 motivated DARPA to develop the Robotics Challenge (DRC) in 2012. Immune to radiation damage, Japan could have used robots to help rescue people, or go into the Fukishima power plant to turn off valves, investigate leaks or structural damage. Yet after decades of robot research and development Japan did not have a rescue robot. Where was the real Astroboy (Maleki & Farhoudi, 2015; Šabanović, 2014)? Humanoid Robotics Project (HRP)-2 was functionally designed to assist people in construction, dangerous environments and home (Kaneko, Harada, Kanehiro, Miyamori, & Akachi, 2008) but did not have the operational capacities to help when needed.

In response, the DRC challenged robots to perform tasks modeled on the context of urban search and rescue (USAR) and industrial disaster response task domains (Yanco et al., 2015). Tasks were real-world anthropomorphic manipulation and mobility; controlled by automated interfaces and teleoperation. Challenges included obstacles such as opening a door, turning a valve, driving a car, and walking over a pile of chaotic bricks. The first robots to attempt the challenge failed miserably. They almost all fell over or were unable to complete tasks so simple for humans. The DRC robots were not even autonomous—actions were manually controlled by teams.

Thus, despite early optimism that robots would be capable of performing human-level tasks by 2015, machines are still far from achieving this goal. Very basic tasks still require supervisory human control from one or more operators. Complex environments such as USAR, require continuous direct control by multiple operators. Engineering autonomy in robots requires more research in both pragmatic design and societal implications. Trust will emerge from evidence-based control interface design that accommodates multiple control paradigms of the robot and the user (Yanco et al., 2015).

Even though the DARPA challenge remains difficult to accomplish, AS are already being depended upon in our lives, from our adaptive smart phones (Levy, 2016), to off shore oil rig drilling programs (Gressgård, Hansen, & Iversen, 2013). Self-driving modes in cars (see Telsa WIRED.com, 2016), mining trucks (Sganzerla, Seixas, & Conti, 2016) and buses (Reilly, 2016) are already in use. Now is the time to understand the metaphysical, epistemological and normative dimensions of trust and trustworthiness so that we can build, use and thrive with AS.

3. Who or what is trustworthy?

Who or what is trustworthy? In this section I consider what sort of property trustworthiness is and the sorts of components a trusted AS might comprise of. Trustworthiness might be an *intrinsic* property of an agent similar to height, or a *relational* property similar to tallness. Perhaps a robot that survives the apocalypse, like WALL-E (Mattie, 2014) is trustworthy due to *intrinsic* moral virtues such as

charm, cheeriness and helpfulness, even if there are no other humans or robots to trust him? Or WALL-E is trustworthy when compared to other robots such as EVE programmed to obey directives. Trustworthiness might be a substantial property—an independent particular—or a dispositional property—the capacity of an object to affect or be affected by other things. The classic example of a dispositional property is fragility. A vase is fragile because it breaks easily. A dispositional account might suppose that a person is trustworthy because they speak truthfully or act reliably with others.

It might be thought that trustworthiness is both a dispositional *and* relational property established by the subjective judgment of one agent X of another agent Y in virtue their shared spatio-temporal interactions. For example, an employee goes through a three month probation period or a solder undergoes basic training to build their reputation with a Drill Sergeant or manager. The graduating employee or solider are deemed trustworthy for a prescribed set of activities with a particular group of people in a specific context. Note that any trustworthiness ascribed to an individual due to these processes pertains to that domain of actions. It's not clear how generalizable or transferable trustworthiness is. At least an argument needs to be made to demonstrate the transferability of trustworthiness across domains.

What is interesting about Trustworthiness understood as a dispositional and relational property is that it can be established by combining judgments from multiple agents, such as through peer assessment (Lopez, 2015, September 24). In this way, an IT device can be judged trustworthy through a network of sensors using a reputation-checking algorithm. For example, beacon nodes on Wireless Sensor Networks can be evaluated on whether they are providing accurate location identification by 1-hop neighboring nodes (Srinivasan, Teitelbaum, & Wu, 2006). Autonomous trustworthiness-evaluation and -judgment is important when networks are vulnerable to malicious interference. Indeed, trustworthiness evaluation programs are considered increasingly important with the proliferation of autonomous systems connected via the Internet of Things (IoT) (see Chen et al., 2011; Sicari, Rizzardi, Grieco, & Coen-Porisini, 2015; Yan, Zhang, & Vasilakos, 2014).

If the dispositional and relational account of trustworthiness is right, then what dispositional properties does it consist of? In the preceding paragraphs I suggested that a person might be trustworthy because they speak truthfully or act reliably. Let's look at these ideas more closely.

Central to the notion of trustworthiness is reliability and accuracy. So, an AS is trustworthy if we can *rely* on it being *right*. For example, a binnacle compass is trustworthy if a sailor can rely on it to accurately adjust to the rise and fall of the waves and orient to magnetic north (Basterretxea-Iribar, Sotés, & Uriarte, 2016). If a sailor navigates to the wrong shore, she might wonder if her compass has become unreliable and thus she ought not trust it. Perhaps ferrous nails have been used that pull the needle away from true readings and the binnacle compass's reliability compromised?

Is trustworthiness more than reliability? How do properties such as adaptability meet reliability? For example, the trustworthiness of a rescue dog might be its capacity to adapt to severe conditions, such as digging through an avalanche to find a stranded person, even if the dog has never encountered such an environment. Adaptability is not an orthogonal trait, but a higher order reliability. In this case, we rely on the dog to be adaptable in unusual, unexpected or changing conditions. The trustworthiness of people, creatures and machines is related to the reliability of their capacities and functions in domains of differing complexity and uncertainty.

Is trustworthiness also about redundancy? We know that AS will not be perfectly safe. There will be hardware failures, software bugs, perception errors and reasoning errors (Fraichard & Kuffner, 2012). Aerospace and military operations build in an expectation of failure into design to enable trust. For example, Boeing 747's only need a single engine to fly, yet are equipped with four engines to ensure redundancy (Downer, 2009). The Space Shuttle program used five identical general purpose digital computers (Sklaroff, 1976). Four of these computers operated as a redundant set and the fifth calculated non-critical computations. The anticipation of failure and the deliberate engineering of multiple systems in avionic engineering makes these systems more reliable and hence more trustworthy. Still, is there more to trust than reliability?

Philosophers have traditionally differentiated reliability and trust. While reliability is necessary for trust, it isn't sufficient. Reliability is a property of machines and inanimate objects, where as trust occurs between conscious agents. For example, we rely on a shelf to hold books, but we do we *trust* the shelf (Hawley, 2014)? Fully-fledged trust seems to involve reliability *and* psychological components such as the ability to apologise if we let people down, if we fail to do as we said we would. A shelf has no attitudes towards what it does. Human trust is traditionally mentally, linguistically and rationally based rather than limited to summaries of behavior (Faulkner, 2007; Hieronymi, 2008; Keren, 2014; Simpson, 2013). AS are a challenge to traditional philosophical distinctions on trust because they are inanimate, in the sense that they are programed to fulfill a set of tasks within a domain and have no intrinsic care for humans and no self-driven desire to maintain their reputation. The tradition to incorporate psychological attitudes in a model of trust could either be misplaced or reconsidered to drive the design processing the age of AS.

By focusing on systems as well as people, the business management literature may provide a more suitable starting framework for building trusted AS than philosophy (for more philosophical discussion see McLeod, 2015). The management two-component model of trust differentiates competence—consisting of skills, reliability and experience—and integrity—consisting of motives, honesty and character (see *Figure 1*). Using this framework user trust in AS could be grounded in reliable operations built by high-integrity organisations.

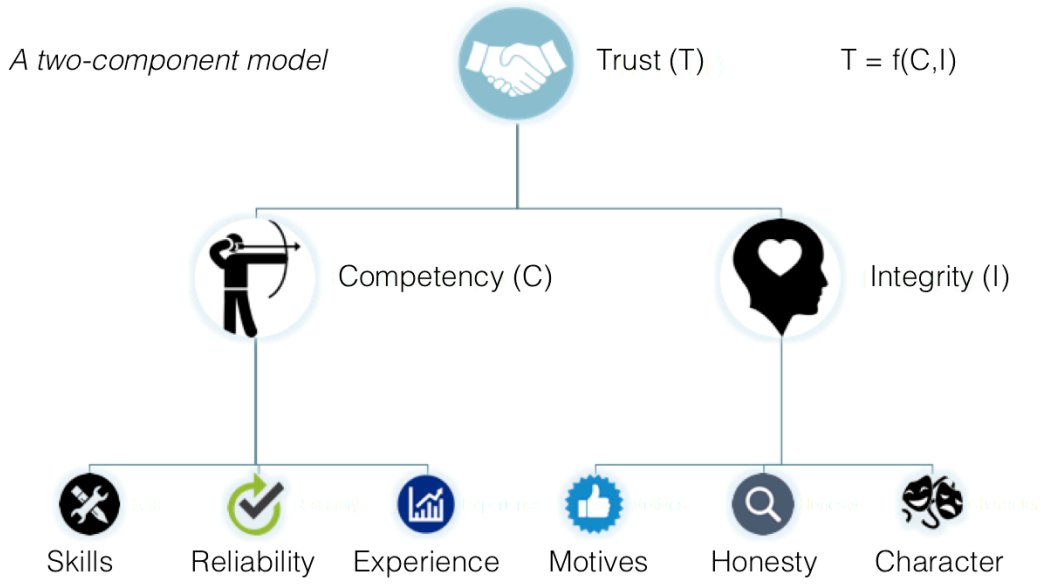

Figure 1: A two-component model of trust incorporating *competence*—skills, reliability and experience—and *integrity*—motives, honesty and character (Connelly, Crook, Combs, Ketchen, & Aguinis, 2015; Connelly, Miller, & Devers, 2012; Kim, Ferrin, Cooper, & Dirks, 2004; Luo, 2002).

Competence comprises of skills, reliability and experience. A person or robot can be competent and yet occasionally not have exactly the right skills for the job, or the sometimes fail to do a task within their domain and sometimes reach the limit of their experience. Competence is thought to improve when an individual learns more skills, becomes more reliable and has more experiences. Integrity can be analysed as comprising of motives, honesty and character. We trust someone who is trying their best, who is transparent about their actions and has a character that, regardless of competence, inclines them to take responsibility for their actions, be thoughtful and empathetic to others and other traits. This two-factor model of trust combines ability and ethics (Connelly, Crook, Combs, Ketchen, & Aguinis, 2015; Connelly, Miller, & Devers, 2012; Kim, Ferrin, Cooper, & Dirks, 2004; Luo, 2002). Trust (T) consists of:

Competence (C)

skills (C_s)

reliability (C_r)

experience (C_e)

Integrity (I)

motives (I_m)

honesty(I_h)

character (I_c)

$T = f(C,I)$

Research suggests an asymmetry in the way trust is lost between these two factors. A single integrity failure may result in a loss of trust in the way that a single incompetent action does not (Kim et al., 2004). People use integrity judgments to generalize across domains of a relationship, where as competence is more domain specific (Connelly et al., 2012). Additionally integrity-based trust implies a reduced threat of opportunism in a way that competence-based trust does not (Luo, 2002). Trust depends on beliefs about the other's benevolent motives (Yamagishi, 2011).

Notice the difference between human-human trust and human-AI trust violations. There is an interesting asymmetry between levels of competence required for humans to trust other humans versus trusting AI. Unlike human-human relationships, trust built up inductively between humans and AI can be destroyed with single instances of inaccuracy or unreliability. Consider the mistakes Google's AI made identifying vertical wavy lines as a starfish (Nguyen, Yosinski, & Clune, 2015). A single misidentification of a starfish can end trust in that machine learning algorithm even though it has performed well in the past. Consider the disproportionate media scrutiny of the first Tesla autopilot fatality. The Tesla flaw was due to the car's incapacity to differentiate the reflectance of light from a truck from the reflectance of the sky (The Tesla Team, 2016). Even though human drivers make perceptual errors leading to crashes all th time, the Telsa fatality caused much uncertainty around whether the AI responsibly could be trusted. The Tesla case is a good example of much higher competence-based trust thresholds for AS than human operators and where a single model may not be sufficient.But, not only are competency requirements misaligned between human-humans and human-AS, but the integrity aspects of the model present a challenge for AS design.

Consider the requirement for honesty in *Figure 1*. Engineers might correctly wonder *how* to communicate complex computational processes to human operators who themselves do not have the competency to understand their underlying logical operation? The data and algorithms of autonomous agents are hidden from most human stake-holders and cannot be understood even if a translation layer were added and explanations communicated in plain language. Perhaps human do not expect honesty from AS the same way they do from other humans? A question then is, whether humans *should* mistrust AS based on perceived honesty violations (I_h). Should engineers creating an AS prioritise transparency and communication of their decision-making mechanisms for trust and adoption? Should users demand them? It is important to consider that integrity components of the trust model might be appropriate for the human engineers, designers and corporate representatives of AS, but perhaps not crucial for the systems themselves? That is, so long as human stakeholders can honestly report the technical specifications of AS to other experts such as regulators, then AS do not need to convey integrity information to users.

What about the role of motives (I_m) and character (I_c) on trusted autonomy? Sometimes humans believe AS have more psychological reality than they actually do due to clever programming. ELIZA was one of the first relational AIs designed to engender trust using simple grammatical tricks (Carbonell, 1970; Turkle, 2007). Little has been developed since that could be dubbed motives or character. AIs in science fiction imagine how character might affect operations. HAL from the movie *2001: A Space Odyssey* (Kozlovic, 2003; Kubrick & Clark, 1968) is a malevolent AI who lacks integrity, but is fairly competent at achieving a mission—albeit his own. Deep Thought from the Hitchhikers guide to the Galaxy (Adams, 1979; Naiditch, 2000) is a benevolent AI who provides answers that humans don't want to hear, such as that the meaning of life, the universe and everything is 42. AIs can have varying degrees of competence and integrity that affects how we trust them. Additionally, may be other factors in a successful model of trust to truly understand how humans will respond to extremely smart AI.

The model described in this section is the start of an investigation of what trustworthiness could be between humans an AS based on an interdisciplinary investigation. Critics have noted that the model above confuses an influencing factor and an indicator[1]. They argue that reliability is an *indicator* of competence, not an input like skills and experience that generate competence. Skills and competence are independent variables that influence competence. I argue that while reliability is not an input, it is a *property* of a trustworthy system, not merely an indicator, hence its inclusion in the model along with skills and experience. Isolating reliability from skills and experience is meant to allow for multiple ranges in skills, reliability and experience to operate independently from one another. So, a person might be a skilled carpenter with years of experience, yet be incompetent at time t_m because his divorce lead him to alcoholism and unreliable behaviours. Reliability is not merely the combination of skills and experience, it requires additional features such as the adaptability and redundancy discussed above. However, the critic is right that much more work needs to done to refine and hone this model to appropriately capture the metaphysics of trustworthiness for AS. The management model is just the beginning of incorporating human factors into AS design.

4. How do we know who or what is trustworthy

How do we decide whether to trust? In §3 the properties that establish and define trustworthiness were considered. In this section the epistemology of trustworthiness is examined—how do we know who or what is trustworthy? What are the indicators of trust? If a person claims to justifiably trust another, it indicates they have the ability and confidence to predict others' behaviour (McAllister, 1995). Implicit, heuristic or 'gut' indicators of trust are often grounded in physical responses and intuitions. Explicit, reflective or rational trust stems from our experience of people over time and our reasons to judge their trustworthiness. Often we do not know why we trust, we

[1] Many thanks to an anonymous reviewer for bringing up this distinction

trust implicitly. Thomas Reid (1764) argued that reasons could not be required for trust given that 'most men would be unable to find reasons for believing the thousandth part of what is told them.' Reid's point is that humans must be justified to trust even in the absence of reasons. Consider the way we use Google maps. Many people use Google maps to get them where they need to go, without knowing how Google maps works, how their phone works or how traffic influences the instructions Google maps provides. Not only do people not know why they trust Google maps, it does not seem to concern people that they do not know why. So how do humans make trust judgments of systems and each other, and are these the same mechanisms that elicit trust in AS? This section moves through implicit and explicit justifications of trust followed by a cognitive model of trust and competence and finally a brief comment on the relationship between trustworthiness and risk.

4.1. Implicit justifications of trust

Implicit justifications of trust are preconscious, embodied trust responses developed without top-down cognitive evaluations. For example a monkey climbs a vertical structure implicitly trusting that it will improve their odds of survival against predation. Researchers know how to alter physical properties of embodied AS (i.e. robots) to engender implicit trust including how they look, sound and feel. Social robots are designed with big responsive eyes and eyebrows (Breazeal, 2002), as are mobile, dexterous and social robots (MDS) (Breazeal et al., 2008; Trafton et al., 2013). Some designers have shaped robots like baby animals—such as the harp seal robot PARO (Šabanović, 2014)—and use biomimetic features such as soft skin for tactile trust (Kim, Alspach, & Yamane, 2015). The Kismet robot with human-like eyes, eyebrows and lips was designed to recognize and mimic emotions, including facial expressions, vocalisations and movement (Breazeal, 2002).

Physical actions connote trust in humans. Japanese robot designers have found cultural identification with a robot who imitates traditional 'aizu bandaisan' dance (Šabanović, 2014). Japanese robot designers try to build trust by incorporating aspects of fictional references to helpful and social robots, such as Anime characters Astroboy and the Patlabor (Šabanović, 2014). But, representations can be incredibly primitive and build emotional attachment, for example, humans watching 2D dots moving on a screen intuitively differentiate between *animate* versus *inanimate* movement based on how wel algorithms replicate biological behaviour (Pylyshyn, 2003). Mimicry of biological behaviours can make people empathise and be concerned for the wellbeing of robots, evidenced by viral videos of the Spot robot by Boston Dynamics being kicked and struggling to stay upright (Boston Dynamics, 2015).

People enter into a relationship with a robot if it simulates human-like emotional and personal understanding, even though these relationships lack the authenticity of shared human meaning (Turkle, 2007). Entirely soft autonomous robots may bridge the authenticity divide, triggering different emotions and trust reactions than solid state robots. Consider the 3D printed soft Octobot designed to emulate a real Octopus, controlled with microfluidic logic instead of microchips (Wehner et al., 2016). Biology-inspired control systems are likely to affect trust responses.

The way AS communicate verbally and through sound can have a big impact on implicit trust. Tom Gruber (Siri Advanced Development Head at Apple) argues that people feel more trusting of Apple's Siri if she has a higher quality voice, "the better voice actually pulls the user in and has them use it more. So it has an increasing-returns effect" (Levy, 2016).

Physical characteristics also impact on how much humans move from empathy to revulsion when robots are like humans, but eerily not quite like humans—known as the uncanny valley (Mori, MacDorman, & Kageki, 2012) impacting how much people intuitively trust them.

There is much research still to be done on whether AS that does not attempt human-like physical characteristics might not arouse the same empathy or emotional connection, but may still generate trust. The rise of chatbots in the tradition of Eliza is a linguistic means by which to generate disembodied trust (Dale, 2016). However, one benefit from realistic facial gestures and embodied movements of robots could be a speed advantage of conveying subtle information regarding the uncertainty of a robot's beliefs, their skepticism or their competing interests when providing an answer to human query improving integrity judgments (see *Figure 1*. §3). Such gestures may be implementable as avatar animations alongside text communication.

The model outlined in §3 may also help us understand how humans implicitly trust autonomous systems in lieu of human-like physical characteristics or avatars. Consider human-drivers who trust Telsa's autopilot function. The car has no physical similarities with humans. Additionally, Telsa drivers cannot trust Tesla because they *explicitly* know anything about the algorithms before they set the autopilot on. Trust could come from implicit factors such as integrity or reliability (see *Figure 1.*). Integrity judgments may stem from a cult of personality around Elon Musk's extensive future vision for solar power, electric cars and sustainable colonies on Mars (Flanagan, 2015, Oct 4)?

4.2. Explicit justifications of trust

Trust is explicitly justified when we have reasons to rely on someone or something. These reasons might coalesce into a deductive, inductive or abductive inference based on the testimony and behaviour of an agent. The link between trust and higher order reasoning is supported by research showing that human intelligence relates to how successfully people evaluate trustworthiness (Carl & Billari, 2014; Cosmides, Barrett, & Tooby, 2010; Yamagishi, 2001). Under this hypothesis, intelligent people foster relationships with people less likely to betray them and make better contextual judgments to account for circumstances where trust is difficult to uphold. Explicit reasons for trust may allow more nuanced and accurate trust judgments than relying on gut feelings or intuition.

Faulkner (2011) argues that though we need reasons to trust an agent generally, we do not need reasons to justify *particular* statements from that agent. Our reasons to trust are based on evaluations of a *general* trustworthiness of an agent (Faulkner, 2007; Hawley, 2014; Hieronymi, 2008; Hinchman, 2005; McGeer, 2008). After all, the boy

who cried wolf was not trusted in the end because he had a history of false testimony even though he was correct in the final instance. A trustworthy reputation for Y built up inductively with X can be shared quickly via testimony to other agents P, Q & R etc... Thus, the value of a trustworthy reputation is not only the ability of X to act based on information provided by Y, but its *transferability*, that is, secondary agents P, Q and R, are justified to trust Y iff they trust X without themselves needing prior interaction with Y. The transferability of a trustworthiness judgment increases the effectiveness and efficiency of social relationships and information systems.

But, does increased efficiency dangerously increase risk? Hume rejected testimony as a source of justification for trust (Hume, 1739). He thought that a hearer was justified to trust based only on their *personal* observations of the speaker's history of truth-telling plus inductive inference from those observations (Goldman, 2004). Hume's reluctance to accept other people's pronouncements demonstrates the subtly and context-sensitivity of trust relationships. An AS might be trustworthy for native English speakers, but break down when deployed in mixed language context. Or an AS learns how to operate with a Platoon, but must be re-skilled each time it interacts with a new human team.

Highly complex AS are a problem for explicit justifications of trust. Because if reasons are required for trust, then perhaps no individual has sufficient reasons to make such a judgment? Take the job of calibrating a ScanEagle unmanned aircraft with hyperspectral imagery sensors to map coastal areas (Hughes, 2015). One individual might verify the location and ensure the imagery sensors are operating correctly but be unable to evaluate the hyperspectural map. The point is that no one operator may know or vouch for all components, mechanisms and physical properties that comprise a complex AS. A key difference between human-human trust and human-As trust is the complexity and difficulty of a single agent-agent dyad relationship.

I propose that instead of relying on individual testimony AS be judged trustworthy by teams and groups that are themselves deemed to be trustworthy within the domain. Groups may include (but are not limited to):

1. Regulatory agencies responsible for issuing parameters of safe operation including physical construction and operational algorithms, operator licensing, maintenance, consumer safety.
2. Institutions and companies designing and building AS.
3. Cohesive teams of staff responsible for successful operations
4. Environmental conditions conducive to operational success

Hume's framework could still be useful within a more layered and complicated system of establishing explicit trust. A Humean regulatory framework means that an individual is justified in trusting an AS in virtue of their background knowledge of the past veracity of regulators, companies and staff plus inductive inference from those beliefs to a current instance. However, induction remains a significant problem for fast evolving AS. New AS may be made by cohesive and trustworthy teams, yet not

have sufficient inductive evidence to generate warranted trust in their safe operation. This may be true, even though an individual knows that a particular aircraft company has a history of trustworthiness and that the regulatory bodies have a history of safe aircraft policies. In cases where innovation is radical and complex, trustworthinesss needs inductive and abductive arguments—inference to the best explanation—to justify operations. An individual or organisation should devise an individualized set of weighted factors that together render a trust or not-trust threshold for a particular AS.

4.3. A cognitive model of trust and competence

Considering both intrinsic and extrinsic forms of justification, is there a linear relationship between competence and trust (holding integrity constant)? I propose that trust and competence forms more of a quadratic relation for trust. We build trust as agents become more competent. We reserve a pinnacle of trust at a human level of competence, and then trust declines as humans or machines exhibit competence at the outlier or far beyond ordinary human capacity to understand it—see *Figure 2*. This model needs empirical testing, but I think the burden of proof is on the developers of AS to demonstrate how trust can be retained or improved as competence surpasses human capabilities and understanding. Such a justification may arise via reputational justifications as specified in §4.2.

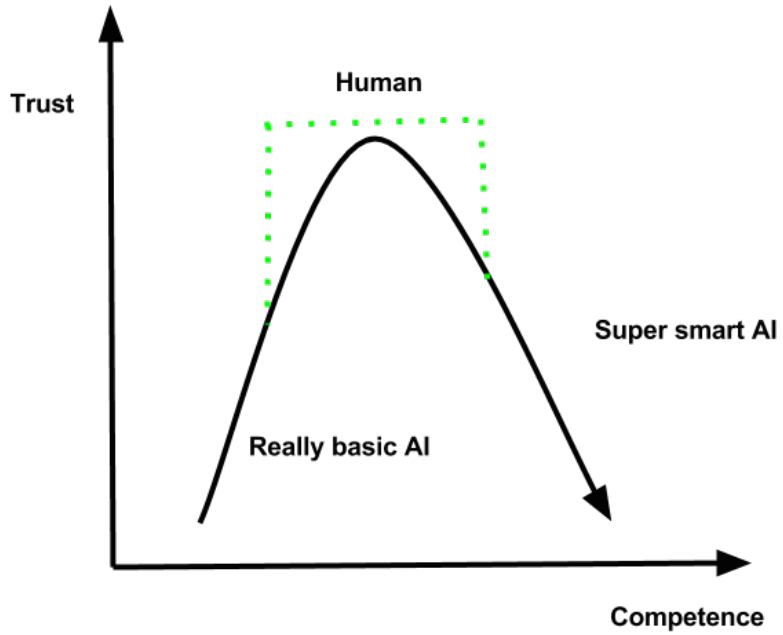

Figure 2: Model of trust and competence where human levels of competence yield the highest trust and trust is reduced at sub-human and super-human levels.

To appreciate the impact of outlier competence, consider the AlphaGo game played against leading Go player Lee SeeDol in 2016, (Metz, 2016). In move 37, Match 2, AlphaGo—a machine learning AI—placed a single black stone on the board that shocked the human player Lee SeeDol so much that he immediately left the table. This move was incomprehensible at a top Go playing level. What this move revealed was that humans sometimes do not understand why an AI acts in order to evaluate it. This is relevant because each competitor playing Go must presume the capacity in their opponent (human or AI) and use game play to build theories to explain strategies and mistakes of their opponent. When playing another human, Go players might overtly inquire about the opponents Go background (how old were they when they begin playing? How much have they played? Who have they played against? What books have they read? What teachers have they had? What sort of handicap do they have? etc…). Players watch their opponents actions, not only the stones placed, but the manner of their placement, and the ultimate destination on the board. Each move can be evaluated in the immediate context of the game, but also in forming what Nelson Goodman (1983) describes as *overhypotheses* about their opponents style, learning journey, preferences, beliefs and desires. Players use these overhypotheses to predict what an opponent will do, then use these predictions to design their own strategies to counteract them. In terms of outcome, Move 37, was very strong, providing support to stones over a large swathe of the board. But, at the moment the move was made, it was impossible to trust by the human opponent because they could not evaluate the competency of the action based on the information available about its genesis. What was AlphaGo? How does it think? What grounds its decisions? How does it make its decisions? Human understanding is critical to trust between humans and AS. It is likely in the future that more and more AIs driving AS are complex, sophisticated intellects, born of machine learning and other architectures. The danger is that humans do not trust them because they cannot understand them.

Consider if AlphaGo was a platoon commander, sending troops into a war zone. Imagine, just as in move 37, the AS commander ordered soldiers to go to a place they could not make sense of; that they felt put their lives or civilian lives at unnecessary risk? Keep in mind that each soldier has a duty to disobey an unlawful order if its illegality is immediately obvious, such as procedural irregularity or moral gravity (Osiel, 1998). In these cases, humans ought not trust the AI, even if the AI proves to be more competent than human decision makers. The AI could have access to huge repositories of data unable to be processed by humans. These calculations and decisions are frightening to humans and justify wariness and skepticism. Even more significantly, suppose complex sophisticated AIs were in charge of Lethal Autonomous Weapons Systems (LAWS), both decisions to target and decisions to fire, how do we know whether to trust them? How would deaths be judged just or unjust if the algorithms deciding who dies are beyond human comprehension? LAWs led by AIs may lead to unintended initiation of armed conflicts and the unjust escalation of conflicts (Asaro, 2016).

It is important to note that leading manufacturers of LAWs currently require human oversight and judgment for all decisions to target and to fire (Asaro, 2016). Current restrictions are based on the notion that humans are better decision-makers than

machines. However, manufacturers continue to build incrementally autonomous capabilities across all systems. To imagine the impact of increasing autonomy for weapons systems, it is instructive to consider how other industries have rolled out autonomous systems and their impact on human users. Car manufacturer Tesla released a self-driving mode on its cars with the requirement that humans always have their hands at the wheel. Yet, Tesla drivers drive while deliberately disobeying protocols because they trust that the systems *do not* actually require their oversight (Slow News Day, 2016). There is evidence as AS become increasingly sophisticated humans may become either overly trusting or overly skeptical. Consider research on autonomous offshore oil drilling system operations (Gressgård et al., 2013). Drill operators sometimes abandon their duty to oversee AS due to competing cognitive demands or they ignore the AS and make their own decisions inefficiently. In both cases the level of trust in the autonomous system plays a direct role in how humans view their obligations to participate in broader systems operations or obey oversight protocols. In sum, while there are currently policies requiring LAWs to be under ultimate human control, the pressures and stress of combat may lead to humans relinquishing control. In the future humans may not have the competence to be in control of these systems.

Perhaps more frightening is a future where AS knows how to manipulate consent and trust in humans (Bostrom, 2016). This is a situation where we trust an AS because it is clever enough to manufacture our trust. But, it does so in either a disingenuous or manipulative way. It is not hard to imagine such an AI capitalizing on inductive trust tendencies or biases in humans. Consider Nelson Goodman's (1983) thought experiment about the colour of emeralds known as the 'grue-paradox' (Cohnitz & Rossberg, 2016). In this hypothetical, all our experience of emeralds is their greenishness, so we ascribe to them the stable and persistent property 'green'. Goodman points out that in fact, Emeralds might be not green but 'grue'. Grue is a property of objects that makes them look green until a particular time (e.g. 2025), but look blue afterwards:

> (DEF 1) *x* is grue $=_{df}$ *x* is examined before *t* and green ∨ *x* is not so examined and blue

If Emeralds are grue, they have never been green. Now suppose we take this hypothetical case of false induction (i.e. trying to establish facts about emeralds and their colour from history and experience) and consider malevolent programmers building an AS. These programmers design a robot that engenders trust over time, for a long time, like an embedded undercover operative. During production and deployment, the AS passes every test humans and regulators can design to establish its trustworthiness. The AS is tested in hundreds of real time situations and thousands of simulated scenarios. But, unbeknownst to regulators, it has been programmed to switch modes in 2025 while deeply embedded in society. So, humans trusted it, but then the AS betrays them and carries out its secret objective. There was no way to know, inductively that the AS would flip. That it was actually an untrustworthy AS. It is also concerning to consider if such hidden higher-level objectives can be

programmed, such programs could be activated or changed remotely and iteratively—threatening the integrity of the AS.

4.4. Trustworthiness and Risk

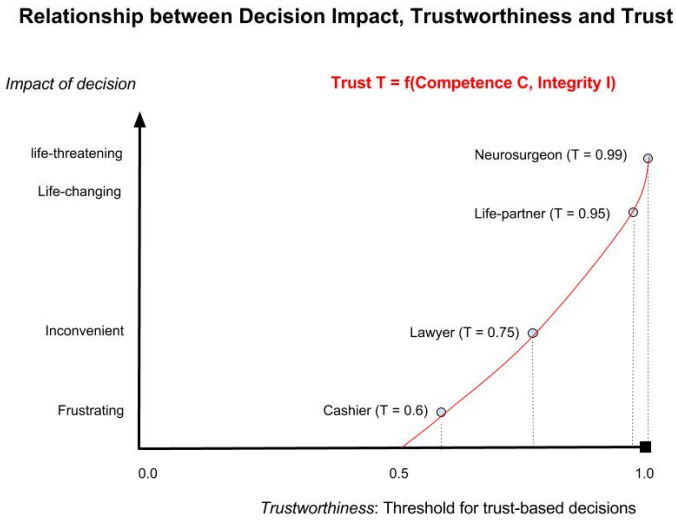

Figure 3: Relationship between decision impact, trustworthiness and trust.

Finally, when ascribing trustworthiness to agent Y, X needs to consider the context of decisions. We have different thresholds for trust depending on the risk of the decisions that have to be made and this in turn depends on impact of decisions—see *Figure 3*. *Figure 3.* Shows the relationship between decision impact, trustworthiness and trust. Life-threatening decisions, such as our choice of neurosurgeon have a higher threshold for trust than merely inconvenient decisions such as our choice of lawyer to settle a contract on a house. Consider PARO, a robot that resembles a baby-seal designed to assist the elderly similar to pet therapy. If PARO malfunctions, very little is lost to the humans who rely on it. But if a rescue robot malfunctions during an evacuation human lives are at stake. If 0 = no trust and 1.0 = absolute trust, We may need to trust our surgeon .99 in order to agree to brain surgery, but only need to trust our check out clerk .65 in order to complete our retail shopping. This is relevant in AS where similar algorithms may be installed or implemented into a huge variety of contexts. We can imagine perceptual and mechanical algorithms that allow a capsicum-picking robot (Perez, Lehnert, Sa, McCool, & Upcroft) to drop 1 in 10 vegetables being reconfigured to help in a rescue operation where dropping 1 in 10 children from a boat is absolutely unacceptable.

4.5. Summary

This section examined the epistemology of trustworthiness. Implicit indicators of trust can be grounded in physical responses and intuitions as well as reputational features of the system that designed and built the AS. Explicit, reflective or rational trust can be elusive, but must stem from our experience of people over time and our reasons to judge their trustworthiness. As AS become more complex, reasons to trust need to be curated from teams of experts including regulators, designers, engineers, users and so forth. Inductive reasoning may need to be augmented by abductive reasoning for radically innovative AS that involve untested combinations of systems and/or new types of systems.

Even once explicit evaluation methods are established, the increasing competence of AS is a risk to human trust. I argue that increased competence increases trust in AS by humans both for implicit and explicit justifications up until competence far exceeds human comprehension. As AS competence continues to increase, humans may cease trusting them because they do not understand them (perhaps frustrating engineers and designers). Or, perhaps even worse, they falsely trust malevolent systems that should not be trusted. Either way, humans may become unreliable at evaluating trustworthiness as AS surpass human cognitive capacities.

5. What or who should we trust?

What or who should we trust? Robots and AS should be programed with our best normative theories of logic, rationality (Johnson-Laird & Byrne, 1993) and ethics tempered with pragmatic performance expectations. Robots and AS are already computational devices, thus abide by propositional logic, predicate logic, and sometimes paraconsistent logic (Torres, Abe, Lambert-Torres, & da Silva Filho, 2009). Robots increasingly make decisions under uncertainty using Bayesian rationality (Bessière, Laugier, & Siegwart, 2008; Thrun, Fox, Burgard, & Dellaert, 2001). In the future, robots and AS will be designed to test newer normative theories of rationality such as quantum cognition (Busemeyer & Bruza, 2012). Ethically, we should trust humans and AS that take care of our interests and obey the law. This section will briefly survey ethical theories that AS ought to abide by.

Legal frameworks can do some of the normative heavy lifting for AS, but unfortunately the law is not nearly nuanced enough to cover human-judged ethical behaviours. For example, suppose a tree branch has fallen on the road during a storm (Lin, 2013). A human driver would cross double-yellow lines on a road to go around the branch once a safety-check was undertaken and we would judge her ethical. However, for us to trust an autonomous car to make the same judgment, violating legal requirements regarding double-yellow lines, it would need to know a huge range of concepts and contexts, e.g. computational versions of terms such as ‘obstruction’ and ‘safe’ (Goodall, 2014). Humans make decisions that violate the law strictly speaking, but are usually nuanced actions that take context and risk into account.

In terms of human rights, AS ought to be aligned to the United Nations declaration of Human Rights (1948), the Geneva Conventions and Protocols (International Committee of the Red Cross, 1949), and human rights law (Asaro, 2012). Additionally, AS ought consider a broad range of ethical theories from philosophy. Consequentialism (or 'Utilitarianism) is a dominant ethical theory that would justify AS actions if they cause the most happiness or 'utility'. For a Utilitarian, LAWs would be justified if they remove human error, thus reduce civilian casualities. Self-driving cars are justified if they massively reduce the road toll, even if the occasional person or bystander is killed through error. Utilitarian arguments are the most frequently cited arguments in favour of deploying autonomous systems. Deontological arguments focus not on the ends of decisions, but the way decisions are made, aka 'the ends do not justify the means'. Kant might agree that lying to all children about the existence of Santa creates the most happiness, but, it is unethical because it violates the Categorical Imperative (Paton, 1971). A deontologically or 'duty' based AS may have a duty to retain all records of software upgrades and decision parameters in an impenetrable black box for later insurance claims and legal determinations regardless of whether such records end up disproportionately punishing low socioeconomic groups. Each design decision can be worked through from different ethical perspectives including social contract theory, virtue ethics or feminist ethics. While different theories may demand conflicting design decisions, many decisions may come out the same. For example, there are both Utilitarian and Kantian justifications for rescue robots to obey triage rules in a rescue. On the other hand, some ethical theories provide a unique way of understanding how and why we trust each other under stressful and uncertain circumstances. Virtue ethics justifies action not based on their consequences or intention, but on virtues such as bravery and honour. Where as Utilitarian or Kantian principles could possibly be coded into a decision maker, virtue is built up over time, via experience and feedback calibrating specific actions against virtuous norms. Virtue ethics could be incorporated into probabilistic decision systems because the right action is not the one that always produces the best outcome. Under virtue ethics we trust an AS if it made the best decision possible in its context given its operating parameters. Additionally, newer ethical theories might fill in some decision-making gaps. For example, Feminist ethics (Jagger, 1992) could justify preferential care behaviours in a special operations team. There is a particular synchronicity between virtue ethics and feminist ethics that could be fruitful for building trust (Halwani, 2003).

Our reliance on people and AS is affected by our level of dependence and cooperation. Our trust in our life partner to care for us involves a multi-faceted risk and trust over time (with shared cognition) versus the one-off trust we might place in a surgeon. For example, we don't really care if our surgeon is nice to his in-laws at Christmas, just so long as he can remove the tumour. We trust people who we believe have strong reasons for acting in our best interests (Hardin, 2002). The main incentive for these reasons is a desire to maintain a strong relationship with us (whether that is economic, love, friendship etc…). Trust between individuals is different to trust we have in corporations. This asymmetry is a really significant issue for AS, because humans ground their trust in beliefs about the corporation behind the AS, not the systems themselves instantiated in a single car, robot, or computer installation.

Social norming is an approach to procedural ethics outside of traditional philosophical theories from anthropology and sociology (Wrong, 1961). Social norming is about learning how to behave in groups to get along the best. It requires we understand social expectations. Detailed theories of cooperative behaviour stem from disciplines such as sociology, biology, anthropology and group psychology. These models are not about competence and achieving optimal performance on tasks, but about creating the most cohesive, resilient teams of organisms. Theories such as game theory contribute to understanding social norming (Ostrom, 2014). One of the many advantages of group level norms is the ability to train AS with social norming without needing top-down ethical theories to drive behaviours.

However, while there are promising avenues for research into the ethical programming to improve trust, many barriers exist for the universalization of such programming. This is because there remains vast disagreement on what the right ethical principles are or even whether ethical principles exist such that they could be implemented into an AS. What does ethical talk amount to? It seems that humans judge each others actions as ethical or not ethical based a huge range of theoretical, contextual, pragmatic and social factors that ethical theories struggle to explain beyond stipulating that actual human decision makers exhibit a sort of hopeless contrariness.

There is a lot of work to be done in determining what the most ethical action is in any particular context and what model underpins such actions. However, even if we can program AS to be ultimately logical, rational or ethical, humans may be uncomfortable. Would we trust machines that obey norms without empathy (Gleichgerrcht & Young, 2013)? Consider the origins of the word *robot* from the 1920 play, *Rossumovi Univerzální Roboti* (*Rossum's Universal Robots*). In the play Czech writer Karel Capek endowed robots with not just thoughts, but emotions to enable them to increase their productivity (Vukobratovic, 2007). Capek's robots were forced workers more like biological androids Replicants in Bladerunner than metal machines. If we program AS with emotions and empathy to build trust, will they suffer if we treat them badly? If AS are moral agents that can suffer, then building trustworthy autonomous systems also means building an ethical and legal framework around their use and identifying their rights (Schwitzgebel & Garza, 2015). Japanese roboticists are already designing robots to have 'kokoro', translated into heart, spirit or mind (Šabanović, 2014). Kokoro stems from animist spiritual thinking that all objects, including rocks and trees, have some level of consciousness and agency including emotions, intelligence and intention. Robots and AS of the future may need complex social identities to meet ethical and social norms.

6. Discussion

The discussion of the metaphysics, epistemology and normativity of trustworthiness has assumed that trustworthy AS are the desired goal. However, do humans want their decisions automated even if available AS are trustworthy? One the one hand optimising AS could be ideal for human-robot interactions, freeing up time and

resources, but on the other hand, perhaps humans want to make their own decisions? We might think that humans develop a sense of identity and security from decision making responsibility in their roles and jobs and that we risk devaluing human workers by outsourcing decisions to AS. If so, then even if AS increase process productivity, it may decrease productivity overall. Alternatively, humans may find work tedious and be glad for near-optimal autonomous task allocations (Gombolay, Gutierrez, Clarke, Sturla, & Shah, 2015). In the Culture novels by science fiction writer Iain M.Banks, the AS 'Minds' make most human decisions that aren't spiritual or fun and the human populace are perfectly content (Rumpala, 2012). 'Minds' are sentient hyperintelligent AIs on space ships and inhabited planets that have evolved to become far more intelligent than their original biological creators. The minds have taken over the administrative infrastructure of the Culture civilization. We don't have to go too far to see that humans already welcome efficiencies that stem from machine learning when they use their smart phones. How many decisions and what sorts of decisions will humans outsource to an AS if given the opportunity?

Interestingly Gombolay et al. (2015) found that contrary to their hypotheses (and in alignment to Iain M. Banks), humans prefer to outsource decision making to autonomous robots even when they perceived their human co-leader more favorably than their robotic co-leader. Interestingly, in follow up questionnaires, subjects felt that their human co-leader had additional properties, such that they liked, appreciated and understood them, that humans understood, trusted and respected each other, and finally that subjects and human co-leaders were important to the task. However, liking humans and wanting them around is not the same as wanting humans to make decisions.

One of the important distinctions when considering AS is the difference between physically instantiated AI (e.g. personal robot) that learns and grows with an individual or team, versus an integrated AI programmed to act over many physical bodies (e.g. networked self-driving cars) that show no preferential or focused behaviours with individual humans. In the latter case, Iain M. Banks Minds and Apple's subtle machine learning might work fine. But, in the former case, social norming may be the right solution.

7. Conclusion

This chapter has examined the trustworthiness of autonomous systems. I have argued that effective robots and autonomous systems must be trustworthy and the risks of reliance justified relative to perceived benefits. Trustworthiness is a dispositional and relational property of agents relative to other agents within spatiotemporal bounds. Trustworthy agents must be reliable (incorporating adaptability and redundancy). A two-component model of trust from the management literature was used to differentiate factors of competence (skills, reliability and experience) to factors of integrity (motives, honesty and character). When humans evaluate the trustworthiness of autonomous systems and other humans they use intrinsic, 'gut' level cues such as physicality as well as extrinsic 'top down' reasoning. Humans tend to trust agents

operating within the bounds of human cognition and are less trusting as systems operate at super-human levels. The threshold for trustworthiness of an agent or organisation depends on the impact of decisions in a particular context. Building trustworthy autonomous systems requires obeying the norms of logic, rationality and ethics under pragmatic constraints—even though there is disagreement on these principles by experts. AS may need sophisticated social identities including empathy and reputational concerns to build human-like trust relationships. Ultimately transdisciplinary research drawing on metaphysical, epistemological and normative human and machine theories of trust are needed to design trustworthy autonomous systems for adoption.

8. References

Adams, D. (1979). *The Hitchhikers Guide to the Galaxy*. United Kingdom: Pan Books.

Anderson, J. R. (2007). *How can the mind exist in a physical universe*. Oxford: Oxford University Press.

Arrow, K. J. (1974). *The limits of organization*. New York: W. W. Norton and Co.

Asaro, P. (2012). On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross, 94*(886), 687-709.

Asaro, P. (2016). *Killer robots and the ethics of autonomous weapons*. Paper presented at the Ethics of Artificial Intelligence, NYU. http://livestream.com/nyu-tv/ethicsofAI/videos/138822041

Basterretxea-Iribar, I., Sotés, I., & Uriarte, J. I. (2016). Towards an Improvement of Magnetic Compass Accuracy and Adjustment. *Journal of Navigation, 69*(6), 1325-1340. doi:10.1017/S0373463316000138

Bessière, P., Laugier, C., & Siegwart, R. (2008). *Probabilistic reasoning and decision making in sensory-motor systems* (Vol. 46): Springer Science & Business Media.

Boston Dynamics. (2015). Introducing Spot. 9 Feb. Retrieved from https://youtu.be/M8YjvHYbZ9w

Bostrom, N. (2016). *Ethics of Artificial Intelligence*. Paper presented at the Ethics of Artificial Intelligence, New York University. Livestreaming Internet Video Capture retrieved from https://wp.nyu.edu/consciousness/ethics-of-artificial-intelligence/

Breazeal, C., Siegel, M., Berlin, M., Gray, J., Grupen, R., Deegan, P., . . . McBean, J. (2008). *Mobile, dexterous, social robots for mobile manipulation and human-*

robot interaction. Paper presented at the ACM SIGGRAPH 2008 new tech demos.

Breazeal, C. L. (2002). *Designing sociable robots*. Cambridge, Mass: MIT Press.

Busemeyer, J. R., & Bruza, P. D. (2012). *Quantum models of cognition and decision*: Cambridge University Press.

Butler, J., Giuliano, P., & Guiso, L. (2009). The Right Amount of TrustNational Bureau of Economic Research. Cambridge, MA. Retrieved from http://www.nber.org/papers/w15344.

Carbonell, J. R. (1970). AI in CAI: An artificial-intelligence approach to computer-assisted instruction. *IEEE transactions on man-machine systems, 11*(4), 190-202.

Carl, N., & Billari, F. C. (2014). Generalized Trust and Intelligence in the United States. *PLOS ONE, 9*(3), e91786. doi:10.1371/journal.pone.0091786

Chen, D., Chang, G., Sun, D., Li, J., Jia, J., & Wang, X. (2011). TRM-IoT: A trust management model based on fuzzy reputation for internet of things. *Computer Science and Information Systems, 8*(4), 1207-1228.

Cohnitz, D., & Rossberg, M. (2016). Nelson Goodman. In E. N. Zalta (Ed.), The Standford Encyclopedia of Philosophy (Spring 2016 ed.): Standford Encyclopedia. Retrieved from http://plato.stanford.edu/archives/spr2016/entries/goodman/.

Connelly, B. L., Crook, T. R., Combs, J. G., Ketchen, D. J., & Aguinis, H. (2015). Competence- and Integrity-Based Trust in Interorganizational Relationships: Which Matters More? *Journal of Management*. doi:10.1177/0149206315596813

Connelly, B. L., Miller, T., & Devers, C. E. (2012). Under a cloud of suspicion: trust, distrust, and their interactive effect in interorganizational contracting. *Strategic Management Journal, 33*(7), 820-833.

Cosmides, L., Barrett, H. C., & Tooby, J. (2010). Adaptive specializations, social exchange, and the evolution of human intelligence. *Proceedings of the National Academy of Sciences of the United States of America, 107*(Supplement 2), 9007-9014. doi:10.1073/pnas.0914623107

Dale, R. (2016). The return of the chatbots. *Natural Language Engineering, 22*(5), 811-817. doi:10.1017/S1351324916000243

Downer, J. (2009). *When failure is an option: Redundancy, reliability and regulation in complex technical systems*. Centre for Analysis of Risk and Regulation:

Economic and Social Research Council. Retrieved from http://eprints.lse.ac.uk/36537/1/Disspaper53.pdf.

Engelman, J. B. M. (2014). *An empirical investigation of the evolutionary and ontogenic roots of trust.* (PhD), University of Leipzig, Leipzig.

Faulkner, P. (2007). On Telling and Trusting. *Mind, 116*(464), 875-902. doi:10.1093/mind/fzm875

Faulkner, P. (2011). *Knowledge on trust*. Oxford: Oxford University Press.

Flanagan, G. (2015, Oct 4). If you think Apple is a cult, you haven't been to a Tesla event. *Business Insider*. Retrieved from http://www.businessinsider.com.au/cult-elon-musk-tesla-model-x-2015-10

Fraichard, T., & Kuffner, J. J. (2012). Guaranteeing motion safety for robots. *Autonomous Robots, 32*(3), 173-175.

Gleichgerrcht, E., & Young, L. (2013). Low levels of empathic concern predict utilitarian moral judgment. *PLOS ONE, 8*(4), e60418.

Goldman, A. (2004). Epistemology and the Evidential Status of Introspective Reports I. *Journal of Consciousness Studies, 11*(7-8), 1-16.

Gombolay, M. C., Gutierrez, R. A., Clarke, S. G., Sturla, G. F., & Shah, J. A. (2015). Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams. *Autonomous Robots, 39*(3), 293-312. doi:10.1007/s10514-015-9457-9

Goodall, N. J. (2014). Machine ethics and automated vehicles. In S. Beiker & G. Meyer (Eds.), *Road Vehicle Automation*. DE: Springer Verlag.

Goodman, N. (1983). *Fact, fiction, and forecast*. Cambridge, MA: Harvard University Press.

Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. *American journal of sociology, 91*(3), 481-510.

Gressgård, L. J., Hansen, K., & Iversen, F. (2013). Automation systems and work process safety: Assessing the significance of human and organizational factors in offshore drilling automation'. *Journal of Information Technology Management, 24*(2), 47.

Halwani, R. (2003). Care ethics and virtue ethics. *Hypatia, 18*(3), 161-192.

Hancock. (2016). *Imposing limits on autonomous systems*. [Unpublished].

Hancock, P. A., Billings, D. R., & Schaefer, K. E. (2011). Can you trust your robot? *Ergonomics in Design: The Quarterly of Human Factors Applications, 19*(3), 24-29.

Hardin, R. (2002). *Trust and trustworthiness*: Russell Sage Foundation.

Hawley, K. (2014). Trust, Distrust and Commitment. *Noûs, 48*(1), 1-20. doi:10.1111/nous.12000

Hieronymi, P. (2008). The reasons of trust. *Australasian Journal of Philosophy, 86*(2), 213-236. doi:10.1080/00048400801886496

Hinchman, E. S. (2005). Telling as Inviting to Trust. *Philosophy and Phenomenological Research, 70*(3), 562-587. doi:10.1111/j.1933-1592.2005.tb00415.x

Hughes, R. (2015). *Sensors for coastal remote sensing*. Paper presented at the SpaceNet Remote Coastal Workshop. http://sydney.edu.au/research/spacenet/pdfs/Roy_Hughes_presentation.pdf

Hume, D. (1739). A treatise of human nature being an attempt to introduce the experimental method of reasoning into moral subjects. *ebooks@Adelaide.* Retrieved from https://ebooks.adelaide.edu.au/h/hume/david/treatise-of-human-nature/

International Committee of the Red Cross. (1949). Geneva Conventions of 1949 and Additional Protocols, and their Commentaries. *Treaties, States Parties and Commentaries.* Retrieved from https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/vwTreaties1949.xsp

Jagger, A. M. (1992). Feminist Ethics. In L. Becker & C. Becker (Eds.), *Encyclopedia of Ethics* (pp. 363-364). New York: Gardland Press.

Johnson-Laird, P. N., & Byrne, R. M. J. (1993). Models and deductive rationality. In K. Manktelow & D. Over (Eds.), *Rationality: Psychological and Philosophical Perspectives*. London: Routledge.

Kaneko, K., Harada, K., Kanehiro, F., Miyamori, G., & Akachi, K. (2008). *Humanoid robot HRP-3.* Paper presented at the 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems.

Keren, A. (2014). Trust and Belief: A Preemptive Reasons Account. *Synthese: An International Journal for Epistemology, Methodology and Philosophy of Science, 191*(12), 2593-2615.

Khemlani, S., Harrison, A., & Trafton, G. (2016). *An embodied architecture for thinking and reasoning about time*. Paper presented at the The 38th Annual

Meeting of the Cognitive Science Society, Philadelphia. http://mindmodeling.org/cogsci2016/papers/0006/paper0006.pdf

Kim, J., Alspach, A., & Yamane, K. (2015). *3D printed soft skin for safe human-robot interaction.* Paper presented at the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 Sep - 2 Oct.

Kim, P. H., Ferrin, D. L., Cooper, D., & Dirks, K. T. (2004). Removing the Shadow of Suspicion: The Effects of Apology Versus Denial for Repairing Competence- Versus Integrity-Based Trust Violations. *Journal of Applied Psychology, 89*(1), 104-118. doi:10.1037/0021-9010.89.1.104

Knight, W. (2015). Can this man make AI more human? *MIT Technology Review.* Retrieved from https://www.technologyreview.com/s/544606/can-this-man-make-aimore-human/

Kozlovic, A. K. (2003). Technophobic Themes In Pre-1990 Computer Films. *Science as Culture, 12*(3), 341-373. doi:10.1080/09505430309008

Kubrick, S., & Clark, A. C. (Writers). (1968). 2001: A Space Odyssey [Film].

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of Human Factors and Ergonomics Society, 46*(1), 50-80. doi:10.1518/hfes.46.1.50_30392

Levy, S. (2016, 25 August). The iBrain is here and its already inside your phone: An exclusive inside look at how artificial intelligence and machine learning work at Apple. *Back Channel*. Retrieved from https://backchannel.com/an-exclusive-look-at-how-ai-and-machine-learning-work-at-apple-8dbfb131932b - .ka6qi6ga5

Lin, P. (2013, 8 Aug). The ethics of autonomous cars. *The Atlantic*. Retrieved from http://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/

Lopez, T. (2015, September 24). Army changing basic training this October. *Army News Service.* Retrieved from https://www.army.mil/article/156044

Luo, Y. (2002). Contract, cooperation, and performance in international joint ventures. *Strategic Management Journal, 23*(10), 903-919.

Maleki, F., & Farhoudi, Z. (2015). Making Humanoid Robots More Acceptable Based on the Study of Robot Characters in Animation. *IAES International Journal of Robotics and Automation, 4*(1), 63-72. doi:10.11591/ijra.v4i1.6639

Marcus, G. F. (2012, 25 Nov). Is 'deep learning' a revolution in artificial intelligence? *The New Yorker.* Retrieved from http://www.newyorker.com/news/news-desk/is-deep-learning-a-revolution-in-artificial-intelligence

Mattie, S. (2014). WALL · E on the Problem of Technology. *Perspectives on Political Science, 43*(1), 12-20. doi:10.1080/10457097.2013.784576

McAllister, D. J. (1995). Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of management journal, 38*(1), 24-59.

McLeod, C. (2015). Trust. In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy (Fall 2015 ed.). Retrieved from https://plato.stanford.edu/archives/fall2015/entries/trust/.

McGeer, V. (2008). Trust, hope and empowerment 1. *Australasian Journal of Philosophy, 86*(2), 237-254. doi:10.1080/00048400801886413

Metz, C. (2016, 16 March). In Two Moves, AlphaGo and Lee Sedol Redefined the Future. *WIRED.com*. Retrieved from https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine, 19*(2), 98-100. doi:10.1109/MRA.2012.2192811

Naiditch, D. (2000). The meaning of life (Vol. 8, pp. 74): Skeptics Society & Skeptic Magazine.

Nave, G., Camerer, C., & McCullough, M. (2015). Does oxytocin increase trust in humans? A critical review of research. *Perspectives on Psychological Science, 10*(6), 772-789.

Nguyen, A., Yosinski, J., & Clune, J. (2015). *Deep neural networks are easily fooled: High confidence predictions for unrecognizable images.* Paper presented at the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

North, D. C. (1990). *Institutions, institutional change, and economic performance*. Cambridge;New York;: Cambridge University Press.

Osiel, M. J. (1998). Obeying orders: Atrocity, military discipline, and the law of war. *California Law Review*, 939-1129.

Ostrom, E. (2014). Collective action and the evolution of social norms. *Journal of Natural Resources Policy Research, 6*(4), 235-252.

Paton, H. J. (1971). *The categorical imperative: A study in Kant's moral philosophy*: University of Pennsylvania Press.

Perez, T., Lehnert, C., Sa, I., McCool, C., & Upcroft, B. (2016). *Sweet pepper pose detection and grasping for automated crop harvesting*.

Pylyshyn, Z. (2003). *Seeing and visualizing: It's not what you think*. Cambridge, MA: The MIT Press.

Reilly, C. (2016, 1 September). Australia's first autonomous bus trial goes off without a hitch (or a driver). *CNET*. Retrieved from http://www.cnet.com/au/news/western-australia-perth-first-driverless-bus-trial-rac-autonomous/

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality, 35*(4), 651-665. doi:10.1111/j.1467-6494.1967.tb01454.x

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of management review, 23*(3), 393-404.

Rumpala, Y. (2012). Artificial intelligences and political organization: An exploration based on the science fiction work of Iain M. Banks. *Technology in Society, 34*(1), 23-32. doi:http://dx.doi.org/10.1016/j.techsoc.2011.12.005

Šabanović, S. (2014). Inventing Japan's 'robotics culture': The repeated assembly of science, technology, and culture in social robotics. *Social Studies of Science, 44*(3), 342-367.

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors, 58*(3), 377-400. doi:10.1177/0018720816634228

Schwitzgebel, E., & Garza, M. (2015). A Defense of the Rights of Artificial Intelligences. *Midwest studies in philosophy, 39*(1), 98-119. doi:10.1111/misp.12032

Sicari, S., Rizzardi, A., Grieco, L. A., & Coen-Porisini, A. (2015). Security, privacy and trust in Internet of Things: The road ahead. *Computer Networks, 76*, 146-164.

Simpson, E. (2013). Reasonable Trust. *European Journal of Philosophy, 21*(3), 402-423.

Sganzerla, C., Seixas, C., & Conti, A. (2016). Disruptive Innovation in Digital Mining. *Procedia Engineering, 138*, 64-71. doi:http://dx.doi.org/10.1016/j.proeng.2016.02.057

Slow News Day (Producer). (2016). Tesla's Model S Autopilot is Amazing! Retrieved from https://youtu.be/UgNhYGAgmZo

Sklaroff, J. R. (1976). Redundancy management technique for space shuttle computers. *IBM Journal of Research and Development, 20*(1), 20-28.

Smith, A. (1776). *An Inquiry into the Nature and Causes of the Wealth of Nations. Book 4, Ch. 2*. (1991 ed.). London: David Campbell Publishers.

Srinivasan, A., Teitelbaum, J., & Wu, J. (2006, Sept). *DRBTS: Distributed Reputation-based Beacon Trust System.* Paper presented at the 2nd IEEE International Symposium on Dependable, Autonomic and Secure Computing. 29 Sep to 1 Oct.

Steinfeld, N. (2016). "I agree to the terms and conditions": (How) do users read privacy policies online? An eye-tracking experiment. *Computers in Human Behavior, 55, Part B*, 992-1000. doi:http://dx.doi.org/10.1016/j.chb.2015.09.038

Sterelny, K. (2012). *The evolved apprentice*: MIT press.

The Tesla Team. (2016). A tragic loss. Retrieved from https://www.tesla.com/en_AU/blog/tragic-loss

Thrun, S., Fox, D., Burgard, W., & Dellaert, F. (2001). Robust Monte Carlo localization for mobile robots. *Artificial Intelligence, 128*(1-2), 99-141.

Torres, C. R., Abe, J. M., Lambert-Torres, G., & da Silva Filho, J. I. (2009). *Paraconsistent autonomous mobile robot Emmy III.* Paper presented at the Advances in Technological Applications of Logical and Intelligent Systems: Selected Papers from the Sixth Congress on Logic Applied to Technology.

Trafton, G., Hiatt, L., Harrison, A., Tamborello, F., Khemlani, S., & Schultz, A. (2013). Act-r/e: An embodied cognitive architecture for human-robot interaction. *Journal of Human-Robot Interaction, 2*(1), 30-55.

Turkle, S. (2007). Authenticity in the age of digital companions. *Interactions Studies, 8*(3), 501-517.

Tyler, T. R. (2006). *Why people obey the law*: Princeton University Press.

United Nations. (1948). Universal Declaration of Human Rights. Retrieved from http://www.un.org/en/universal-declaration-human-rights/

Vukobratovic, M. K. (2007). When were active exoskeletons actually born? *International Journal of Humanoid Robotics, 4*(03), 459-486.

Wehner, M., Truby, R. L., Fitzgerald, D. J., Mosadegh, B., Whitesides, G. M., Lewis, J. A., & Wood, R. J. (2016). An integrated design and fabrication strategy for entirely soft, autonomous robots. *Nature, 536*(7617), 451-455. doi:10.1038/nature19100

Williamson, O. E. (1993). Calculativeness, Trust, and Economic Organization. *Journal of Law and Economics, 36*(1), 453-486. doi:10.1086/467284

WIRED.com (Producer). (2016, 10 Sep 2016). How Tesla's self-driving autopilot actually works. [Video] Retrieved from https://www.wired.com/video/2016/08/how-tesla-s-self-driving-autopilot-works/

Wrong, D. H. (1961). The oversocialized conception of man in modern sociology. *American sociological review*, 183-193.

Yamagishi, T. (2001). Trust as a form of social intelligence. In K. Cook (Ed.), *Trust in Society*. New York: Russell Sage Foundation.

Yamagishi, T. (2011). *Trust: The Evolutionary Game of Mind and Society* (Vol. 1. Aufl.): Springer-Verlag.

Yan, Z., Zhang, P., & Vasilakos, A. V. (2014). A survey on trust management for Internet of Things. *Journal of network and computer applications, 42*, 120-134.

Yanco, H. A., Norton, A., Ober, W., Shane, D., Skinner, A., & Vice, J. (2015). Analysis of Human-robot Interaction at the DARPA Robotics Challenge Trials. *Journal of Field Robotics, 32*(3), 420-444. doi:10.1002/rob.21568