

Chapter 3 – Values in Machine Age

Chapter 3 – Values in Machine Age	1
An Introduction to Values.....	2
What is a value?.....	2
Intrinsic versus extrinsic values.....	3
Intrinsic Values in Philosophy and Psychology.....	3
Nonmoral Values versus Moral Values.....	4
Values versus Virtues.....	4
Necessary Values for Human Growth.....	5
Ethical Knowledge in Machine Age.....	7
What is Knowledge?.....	7
Structuring ethical challenges in IT driven knowledge creation.....	8
Ethical Challenges in Data Collection.....	8
Informed consent.....	11
User Control in Ubiquitous Computing.....	12
Ethical Challenges in Information Aggregation and Knowledge Creation.....	13
Data Quality.....	13
Truth.....	15
Transparency.....	16
Ethical Challenges in the Design of Knowledge Access.....	17
Access to Knowledge.....	18
Objectivity or Filter Bubbles.....	19
Ethical Uses of Information and Knowledge.....	21
Contextual Integrity	21
Privacy Harms.....	22
Computer Bias and Fairness.....	22
Summing Up: Ethical Knowledge Management.....	23
Freedom and Liberty in the Machine Age.....	26
Negative Liberty and Machines.....	26
Positive Liberty and Machines.....	27
Technology Paternalism and Controllability.....	28
Autonomy vis-à-vis Machines.....	30
Attention Sensitive Machines.....	33
Attention-sensitive Interruption.....	33
Summing up: Freedom and Liberty in the Machine Age.....	34
Health and Strength in the Machine Age.....	37
Machines’ direct impact on physical health and strength.....	37
Machines’ long-term effects on physical health and strength.....	37
Machines’ direct effects on mental health and strength.....	38
Machines’ indirect effect on mental health.....	38
Mental health challenges in response to computer use on the job.....	39
Machines’ indirect effect on physical health.....	40
.....	41
On Security and Safety in the Machine Age.....	42
Safety versus Security.....	42
Safety, Cyberwar and Cybercrime.....	43
Security Principles in Machine Engineering.....	43
Information Security Goals	44
Auditability.....	45
Privacy in terms of Security vs. Security in terms of Privacy.....	45

Accountability.....	45
Privacy and Surveillance.....	47
The issue of surveillance.....	47
The Pros and Cons of Surveillance.....	48
Reaching Golden Means in Mass-Surveillance?.....	49
Trust and Confidence in the Machine Age.....	51
What is trust?.....	51
Trust Mechanisms in Machines.....	52
How computer scientists understand trust.....	53
Reputation Systems.....	53
Belonging and Friendship in the Machine Age.....	56
What is Philia (friendship)?.....	56
How can machines influence the various dimensions of friendship?.....	57
Human-to-human friendship in first generation media environments and social networks.....	58
Shared life and learning in virtual worlds.....	59
Empathy in virtual worlds.....	60
Intimacy and Disinhibition in Online Environments.....	63
Intimacy with Artificial Beings.....	64
Final thoughts on friendship in the machine age.....	66
Dignity and Respect in the Machine Age	68
Dignity and Respect.....	68
Respectful Machines.....	69
Polite Machines.....	70
Ownership in the Machine Age.....	71
Coding Freedom: Open and Free Software.....	73
Patents and Copyrights.....	74
How the Privacy Chameleon is Woven into the Value Fabric.....	77
Privacy in Ethical Knowledge Creation.....	77
Privacy and Freedom.....	78
Privacy Trade-offs at all Levels of the Value Pyramid.....	78
Summing Up: Values in the Machines Age.....	80

An Introduction to Values

*“Le mieux est l'ennemi du bien.”
(The better is the enemy of the good)
(Voltaire, La Bégueule, 1772)*

In his famous book “The principle of responsibility,” Hans Jonas ([Jonas 1979](#)) wrote that we must always consider the potential risks of technological innovation: “The recognition of the *malum* is so much easier than that of the *bonum*...”, he wrote. “...primitively speaking, the prophecy of doom receives more attention than the prophecy of salvation (p.70).” With this perspective and emphasis of the negative potentials of technology Jonas influenced a long line of technology assessment work that focused on the negative consequences or potential risks of technical advancements. Less thought has been invested into building positive visions for the future, visions in which we constructively avoid the *malum* while focusing on the good.

To make rational decisions about technology investments, we do have to consider potential risks, but negative thinking rarely motivates people to do any better. In contrast, engineers are typically driven by their desire to build things they enjoy for themselves or find useful. They want to create value, not destroy it. So a better way to frame ethical system design is to embrace a desire to create value through technology. Hans Jonas himself wrote: “...it is not the moral law that motivates moral behavior, but the appeal of the good-per-se in the world” (([Jonas 1979](#)), p. 162). We need to concentrate on how technology can benefit society while addressing its risks along the way.

So what would be ‘good’ information technology? Where does value come from? And how can value be threatened? Reading the five future IT scenarios, you will have noticed that many values can be promoted or threatened by new IT devices and services. These values include privacy, security, freedom, trust, attention, transparency, and many others. On the negative side, future IT has the potential to brutally undermine most of the values we currently cherish. Think of the Alpha1 humanoid robots used by the police force against civilians, the soft robots spying on villagers, Big Data analysis of outdoor times preventing university access, and so on. I could have focused only on this kind of dark scenario, but again, I favored a balanced approach and also showed how IT can promote learning and health (The Wise Figure), coach us good ethical conduct (Arthur Agent) or support our convenience (robots that carry people).

But what is a value actually? Might we not argue that some of the noted issues and effects of IT, such as convenience, seem more like functionality than a value? Is health not a state of being rather than a value? Is control or transparency logically the same as liberty or knowledge? And are some values more important than others?

These questions show that we have to clarify what a value is. In this chapter, I outline the term “value” is understood in philosophy. I will discern intrinsic and extrinsic values as well as nonmoral and moral values. I will briefly discuss the role of values in moral philosophy and how values relate to and differ from virtues. Finally, I will choose those values that seem to be the most vital for human eudemonia (flourishing) and discuss those in detail.

What is a value?

The term “value” comes from “treasuring” something. It implies a degree of worthiness. It is derived from the Latin word “valere,” which means “to be strong” or “to be worth.” A value hence denotes something that is perceived as “good.” Clyde Kluckhohn defines a value as follows: “A value is a conception, explicit or implicit, distinctive of an individual or characteristic of a group, of the desirable which influences the selection from available modes, means and ends of action...A value is not just a preference but is a preference which is felt and/or considered to be justified – ‘morally’ or by reasoning or by aesthetic judgments, usually by two or all three of these” ([Kluckhohn 1962](#)), p. 395 et seq.).

A value is not equivalent with “the absolute good.” In fact, a value implies a threshold level while “the absolute good” is beyond any thresholds ([Shilton 2013](#)). The threshold level of how strongly something is valued depends on the culture of a group or a society at a specific time. Nietzsche (1844-1900), for instance, discussed the *ascetic* ideal that reigned in 19th century Germany ([Nietzsche 1887](#)). As a result of this ascetic ideal, he observed that charity, humility and obedience were important values in his society. In contrast, we can observe today’s capitalist societies, where economic success informs the dominant ideology and where almost opposite values like competition, pride and autonomy are favored. Charity, humility and obedience do still exist as values, which shows that values persist. But their importance fluctuates over the course of history and depends on the ideals of a society. As we move into the machine age, we must consider our current ideals. Our ideals will influence how we regard values such as privacy or freedom and how much importance we grant to them. Value ethics involves asking the question of “what is desirable, good or worthwhile in life? What is the good life as distinct from a morally good life? What values should we pursue for ourselves and others?” (([Frankena 1973](#)), p. 79).

Some values are considered so important over time by some societies that they become *rights* and enter countries’ legal systems. This is the case, for example, for human freedom and dignity. Other values transcend individual countries’ legal systems and become international conventions, encouraging societies to cooperate on the basis of common values. Such conventions are particularly important in times of significant globalization, like today. For example, the right to a private life has entered the European Convention of Human Rights (Art. 8 ([Council of Europe 1950](#))). Further examples are freedom of thought (Art. 18), freedom of peaceful association (Art. 20) and the right to be protected from unemployment (Art. 23), all of which have entered the Universal Declaration of Human Rights ([UN General Assembly 1948](#)). In this book, I will not speak of rights, because I assume that anything that has been recognized as a legal right today is also considered to be a current value (at least by all those countries and cultures that signed the agreements).

Intrinsic versus extrinsic values

Scholars make a fundamental distinction between *intrinsic* (final) values and *extrinsic* (instrumental) values. An *intrinsic* value is something that is valuable “in itself,” or “in its own right” (([Zimmerman 2010](#)), p. 3). When someone asks “**What is (the value x) good for?**”, then for an intrinsic value, the answer goes beyond the mundane. Happiness is such an intrinsic value. “What is happiness good for?” The answer is that happiness is simply there as an ultimate goal of human kind. Scheler (1874-1928) argued that some values are simply

given a priori and are anchored in each person’s *ordo amoris*, an “order, or logic, of the heart” that is not congruent with the logic of reason ([Frings 1966](#)).

Complementary to intrinsic values, scholars recognize *extrinsic* or *instrumental* values. Instrumental values lead causally to intrinsic values. They are not good for their own sake, but they relate to and enable something else that is good. Philosophy scholars therefore say that an extrinsic value is *derivatively* good. It derives its value from the fact that it leads to a higher (intrinsic) good (p. 4 in ([Zimmerman 2010](#))). For example, in the stories in chapter 3, many of the IT applications create convenience. This convenience is an extrinsic practical value because it can increase happiness. An old lady that cannot walk through a mall by herself will find shopping to be much more pleasurable if she is accompanied by a robot that carries her bags and even carries her around when she is physically unable.

Value theory often questions how many values there are. Scholars agree that there are many extrinsic instrumental values, but how many intrinsic values are there? Monists believe that there is only one final value or “super value” to which all other values relate or are instrumental. Epikur (341 BC – 270 BC) and Jeremy Bentham (1748-1823) are famous proponents of this view. They held the view that only the value of human happiness finally counts. This view on human nature has also been called “psychological hedonism” (([Frankena 1973](#)), p. 83). However, most philosophers who have written about intrinsic values have not been monists or even monistic hedonists. Instead, they outline other values besides happiness that have intrinsic value. Frankena (1973) identifies many intrinsic values such as knowledge, beauty, health, truth, power and harmony have all been considered as intrinsic values. He outlines that many philosophers regard the “presence of some kind of degree of excellence” as a characteristic for an intrinsic value.

Intrinsic Values in Philosophy and Psychology

Philosophy is not the only discipline to study values. Psychologists study human values to understand human behavior and motivation. Milton Rokeach developed an extensive value catalogue that has been tested throughout the world (1973) (see table x). In his work on values, Rokeach held five assumptions: “(1) the total number of values that a person possesses is relatively small; (2) all men everywhere possess the same values to different degrees; (3) values are organized into value systems; (4) the antecedents of human values can be traced to culture, society and its institutions, and personality; (5) the consequences of human values will be manifested in virtually all phenomena that social scientists might consider worth investigating and understanding” (([Rokeach 1973](#)), p. 1).

Rokeach’s stance that all men everywhere possess the same values has been challenged by proponents of “ethical relativism.” Ethical relativists believe that there are no universally valid norms and values. Instead, they argue that different cultures, beliefs and practices lead to different values and that all of them should be tolerated. In today’s global and postmodern world, this respect for other cultures and their doings is a very important and timely perspective. Yet, as Charles Ess argues, ethical relativists establish their own global value and that is *tolerance* for other cultures and individuals. Thereby, they indirectly admit that some values may be universal ([Ess 2013](#)). Some scholars have also warned that ethical relativism could lead to “moral isolationism” ([Midgley 1981](#)); if everyone is allowed to do as they please, the willingness and necessity to cooperate falters. Charles Ess warns that such developments can lead to a “paralysis of moral judgment” (([Ess 2013](#)), p. 217). Relativism would require us to accept many cultural and individual practices and preferences in the name

of tolerance that run counter to our intuitive and emotional judgment. As a result, we would be unable to develop true common ground for joint decision-making.

The extreme opposite of ethical relativism is ethical absolutism. Ethical absolutists believe that there are universally valid values that define what is right and good for everyone, everywhere and at all times. Extreme religious communities sometimes tend to argue along these lines.

The middle ground between these two extreme positions is ethical pluralism. Ethical pluralists agree with ethical absolutists that some values are universal. But, embracing ethical relativism, they argue that the degree to which such values are important in a society differ between cultures. They may also differ between individuals depending on where and how those individuals live in a society. For example, the value to belong to a family is probably universally felt in all societies and in most individuals. Yet the degree to which this belonging to a family is important for a person and determines his or her lifestyle differs between cultures and between social subgroups.

In this book, I embrace ethical pluralism. There are universal values that all cultures and individuals can agree on and strive for and it is this set of values that should be respected by our globally distributed IT systems. These universal values may still be of different importance from one country and subculture to another. All users should have the choice though to tweak and set their machines in a way they need it to have their particular value emphasis respected. Take the example of the privacy value: We know that there is a perception of privacy around the world. Machines can be set to respect this value. Yet, every user should be allowed to change the machine settings to be more or less open according to his or her individual preferences.

In order to identify the universal values that count for us globally I draw from knowledge about intrinsic values accumulated over the past 2500 years of scholarship in philosophy. And I then combine this philosophical knowledge with insights gained in psychology. As mentioned above, psychologists have studied values and come up with their own proposals of what is important for people. Even though psychologists pursue a different scientific method than philosophers, there is considerable overlap between the two disciplines when it comes to values. Table x aligns and contrasts the core values listed by philosopher William Frankena (p. 88 in [Frankena 1973](#)) and psychologist Milton Rokeach ([Rokeach 1973](#)).

A summary of values in philosophy according to William Frankena (1973)	Human values identified and measured in psychology (Rokeach, 1973)
Life, consciousness, and activity	Comfortable life (prosperous life)
Health and strength	n.a.
Pleasures and satisfactions	Pleasure (an enjoyable, leisurely life)
Happiness, beatitude, contentment	Happiness (contentedness)
Truth	n.a.
Knowledge and true opinion of various kinds, understanding, wisdom	Wisdom (mature understanding of life)
Mutual affection, love, friendship, cooperation	True friendship (close companionship); mature love (sexual and spiritual intimacy)
Harmony and proportion in one's own life	Inner harmony (freedom from inner conflict)
Power and experiences of achievement	Self-respect (self-esteem)
Self-expression	A sense of accomplishment (lasting contribution)
Freedom	Freedom (independence, free choice)
Peace, security	National security (protection from attack); family security (taking care for loved ones); a world at peace (free of war)
Adventure and novelty	Exciting life (a stimulating active life)
Good reputation, honor, esteem	Social recognition (respect, admiration)
Beauty, harmony, proportion of objects contemplated; Aesthetic experience	A world of beauty (beauty of nature and the arts)
Morally good dispositions or virtues	n.a.
Just distribution of goods and evils	Equality (brotherhood, equal opportunity for all)
n.a.	Salvation (belief in God, eternal life)

Table x Non-hierarchical collection of intrinsic values as summarized in philosophy and psychology

Philosopher William Frankena argued that everything that has value could somehow be related to his list of intrinsic values. Recently, some scholars have argued that new intrinsic values have emerged and should be added, for example, the ecological value of “natural environment” or “untouched wilderness” ([Zimmerman 2010](#)). In the face of current IT innovations, and considering that privacy reappears as a dominant value throughout our scenarios, we might consider adding privacy to the list as well. However, as I will show, privacy is not an intrinsic value. Instead it is an extrinsic value that is highly instrumental to the intrinsic values of knowledge, freedom, security or self- esteem.

Producing finite lists of values can be problematic because it risks excluding relevant concepts. “We should give up the attempt once and for all to make atomic lists of drivers and needs” (p.25), wrote Abraham Maslow (1970), who is known for having established one of the most popular lists of human values. Maslow’s main criticism of value lists is that values in themselves can be broken down into subcomponents: We can have multiple extrinsic values that cater to multiple intrinsic values. “If we wished, we could have such a list of drivers contain anywhere from one to one million drives, depending entirely on the specificity of

analysis” (([Maslow 1970](#)), p. 26). That said, I still believe that lists structure our thinking. IT investments and projects can use the list of intrinsic values as a starting point to creatively reflect on what to cater to and how.

Nonmoral Values versus Moral Values

The values I have described so far are *nonmoral* values. Nonmoral values are properties, states of affairs or facts that we consider good or desirable in our society. Nonmoral values are important because they give a frame and identity to our lives, telling us what is good and worthwhile to strive for. At the same time, nonmoral values are not morally obligatory. They don’t force us to act in a certain way (unless they have entered the legal system). In contrast, *moral values* imply an expectation of how people should behave *relative* to others. They exist as a response to the presence and needs of others (([Krobath 2009](#)) p. 178). They tell us what *ought* to be. Examples of moral values are honesty or fidelity, respect, and responsibility. The most well known moral rules of behavior are those embedded in our religious systems, such as the Ten Commandments guiding Christian and Jewish tradition or the rules embedded in the Quran. For example, “you shall not lie” is a rule that corresponds to the moral value of honesty.

Values versus Virtues

We have seen that ethics recognizes the normative character of nonmoral and moral values. However, some scholars criticize ‘value ethics’ in general. Karl R. Popper (1902 – 1994) said, “much of what is written about values are empty words” (p. 282 in Popper 1979, cited in ([Krobath 2009](#)) p. 13). Some philosophers don’t accept values as having a normative status. Kurt Baier, for example, wrote “The assessment of the value of a thing does not, by itself imply that one should do anything” (p. 53 in Baier, 1969). The German philosopher Martin Heidegger was radically against value ethics, saying that “Thinking in values is the biggest blasphemy, that can be thought of in the face of being” ([Heidegger 2004](#)) cited in ([Krobath 2009](#)) p. x).

This criticism is not unfounded. Values are empty shells if people don’t act on them. I would even add that a major threat to value-based ethics is its potential of abuse by non-virtuous actors. Non-virtuous but powerful actors often claim values they really don’t pursue. They also establish values in a society that are unethical in the end. For example, the Nazi regime in Germany established the value of being of Aryan decent and murdered and prosecuted those parts of the population that were not. Another example is Darwin’s principle of the “survival of the fittest” that is propagated by some members of the elite today as a social value. The arrogance inherent in this value can lead to bitter discrimination against the handicapped, less intelligent or less wealthy. Finally, “transhumanists” in the IT world argue that humans are suboptimal biological systems as compared to digital machines. Transhumanists go on to establish an ideal of superior machines and inferior humans, a philosophy that could lead to a shift in how machines or the owners of a machine infrastructure would treat the rest of human society.

The success of a value-based approach therefore depends on a broad social, historical and philosophical consensus on what constitutes a value or what ideals form an epoch. At any time, the relative importance of values depends on the virtue and wisdom of top decision-makers: High courts, politicians, journalists, entrepreneurs, managers, bloggers, artists,

scientists and IT engineers, to name a few. These figures are in a position to decide whether our IT systems live up to the relevant values of the time. Their courage, generosity, high-mindedness, healthy ambition, truthfulness and perception of justice determine whether machines will promote or undermine our values.

What does it mean to be an eudemonistic leader who fosters the right values to the right extent? A leader who is interested in human growth and advancement beyond his personal bottom line? Leading management scholars like Ikujiro Nonaka have called for re-embracing Aristotle’s thinking to understand what’s expected from wise leadership ([Nonaka et al. 2011](#)). Aristotle (384 – 322 BC) extensively discussed how we reach eudemonia, how we make human flourishing possible by developing an ethical habitus. The virtues Aristotle identified in his “Nicomachean Ethics” are summarized in table x. **Note that Aristotle believed in holding to the principle of the Golden Mean. This means that all virtues derive parts of their rightness by being in the middle between two extreme forms of behavior. Table x therefore includes not only the virtue itself, but also the extreme forms of behavior that people should avoid.**

Throughout this book I will be coming back to some of these virtues. I will describe their importance for wise leadership for instance (chapter x). I will also mention some of them as they become important throughout the IT system design process.

Aristotle's list of virtues as outlined in the ‘Nicomachean Ethics’	
generally	courage (andreia) – acting neither foolhardily nor cowardly
	temperance (sophrosyne) – serenity and calmness vis-à-vis life – find the right balance between desire and indiscipline
in relation to money and property	generosity (eleutheriotes) – sharing goods appropriately in relation to what one possesses – find the right balance between wastefulness and stinginess
	high-mindedness (megaloprépeia) find the right balance between being too grand and too narrow minded/petty
concerning reputation and honor	inflatedness (megalopsychia) – appropriate self-confidence – find the right balance between faintheartedness and inflatedness
	healthy ambition (philotimia) – find the right balance between too little and too much ambition – controlling the urge to be superior
	gentleness (praotes) – not too much, not too little
in communication with others	veracity/truthfulness (aletheia) – find the right balance between showing off and irony – that what <i>is</i> can be recognized as such
	humor (eutrapelia) – pleasantness of conversation/ready wit
	kindness (philia) – brotherly love – neither too smarmy nor too hardheaded
in political life	justice (dikaiosyne) - righteousness

table x: Compilation of Aristotelian values provided by ([Krobath 2009](#)) (p.x)

Necessary Values for Human Growth

Computer ethics scholar Katie Shilton ([Shilton 2013](#); [Shilton et al. 2012](#)) accompanied IT

project teams over many years and tried to discern the reigning values. She was constantly confronted with the “paralysis” phenomenon of “whose values” should actually govern the project. Based on this experience, she developed a framework to describe the various forms in which values can be discussed in technology design ([Shilton et al. 2012](#)): The first dimension is *agency*: Who holds a value – a subject or a machine? Is the value stable or likely to change? Where does the value originate – from a cultural background or from an engineer’s preference? The second dimension along which values can be discussed is the *unit*: Is the value held by an individual or by a collective? By a user or by an engineer? The third dimension is the *assemblage*, which asks the final diversity of values catered to. Often people don’t agree on the importance of a respective value, or they prefer to emphasize different kinds of values in the design of a system; the resulting system is an assemblage of these views.

Shilton’s work shows that project teams suffer from disorientation as to what values to embed in technology. Agency, unit and assemblage question the source and justification of values held by different members of a project team. The hierarchy of values seems to be subjective for each team member. To prevent the hierarchy of values from seeming subjective to engineers or project leaders, it is important to create common ground on values. The starting point of such a common ground could be the intrinsic value categories that have been recognized as vital by both philosophy and psychology, as summarized in table x.

In fact, the engineering community has already approached some of these value categories. The value of *beauty*, for example, entered computer science after a long battle with those who believed that “form follows function.” Human computer interaction scholars have found ways to build systems in a usable way and create a positive user experience that based on our knowledge of aesthetics ([Nielsen 1993](#); [Norman 1988](#)). Emotional and affective computing approaches also work with our perceptions of beauty ([Zhang 2005](#)). The triumphal march of the beauty value in machine design serves as an encouraging example of how values can be embraced and embedded into machines.

An open question is whether some of the 18 intrinsic value categories listed in table x are more important than others. Can we identify any priority or hierarchy among them? One way to do so is to combine this list of cross-disciplinary intrinsic value categories with the needs that humanistic psychology has identified as particularly important for human eudemonia. As I mentioned, Abraham Maslow found some values to be triggered by humans’ *basic* needs ([Maslow 1970](#)). These basic needs are fundamental to human life, and the values they correspond to should therefore be prioritized in the design of IT systems.

Note that needs and values are not the same, but they are directly related. A need is a necessity or a strong want for something, and as long as this object of desire is unfulfilled, our valuation of it is particularly high. For example, when humans are hungry, they don’t value their safety, sense of belonging or esteem as much as they normally would: they just want food. Once the hunger is satisfied, it becomes unimportant (is less valued) in the current dynamics of the individual. Then, people turn to the next higher need they value. In many instances, I may value things, but I don’t need them; for instance, I may value property or fame. The needs described in Maslow’s pyramid are more fundamental. All of these values are needed for human flourishing, which is why he described them as “basic” at all levels of the pyramid.

Maslow’s hierarchy allows us to prioritize some values. We know that if the lower values are destroyed, the higher ones cannot materialize either. For example, we know that people need

some self-esteem. Yet this desire is less valued as long as people’s basic needs for health, food and safety are not (at least partially) fulfilled. So we need to ensure that the lower values such as health are ensured to not cripple human flourishing at higher levels.

Finally, Maslow identified two further values in addition to the five layers of needs in the pyramid. These values are knowledge and freedom, which he considered preconditions for the satisfaction of basic needs. We defend knowledge and freedom, he argued, “because without them the basic satisfactions are quite impossible, or at least, severely endangered” (p. 47).

Concentrating on Maslow’s work, I reduce the list of 18 intrinsic human value categories (table x) to just seven, which are summarized in figure x: knowledge and freedom as preconditions for human growth, health, safety, friendship, self-esteem and self-actualization. In the following I will outline how these intrinsic value categories are affected, created, fostered or potentially undermined by IT systems. I will not write about self-actualization, because the way in which this concept is understood by Maslow, I hardly believe that we can build it into machines ([Maslow 1970](#)). The remaining concentration on just six final intrinsic values is useful because it allows us to focus IT projects on what are indubitably important values across cultures. It also reduces the complexity of this book.

Looking at figure x, some computer ethics scholars will wonder why many of the values that are most commonly discussed in the discipline are not mentioned in the pyramid. For example, privacy, autonomy and trust are not depicted here, even though they are frequently debated values in the computer ethics literature. However, note again that the values shown in this synthesis pyramid are really value *categories* and, moreover, only categories of *intrinsic* character. This means that many extrinsic values help people to achieve these higher order constructs. For example, the need for safety and security cannot be achieved without trust. Trust is instrumental for creating a perception of safety. Therefore the section below on safety and security needs will integrate a subsection on the trust construct. So, in the following, I will not only define and discuss the intrinsic value categories noted in the value pyramid in figure x. I will also present in full detail core extrinsic values that cater to these. The summary of the concepts finally covered is shown in figure x at the end of this chapter.

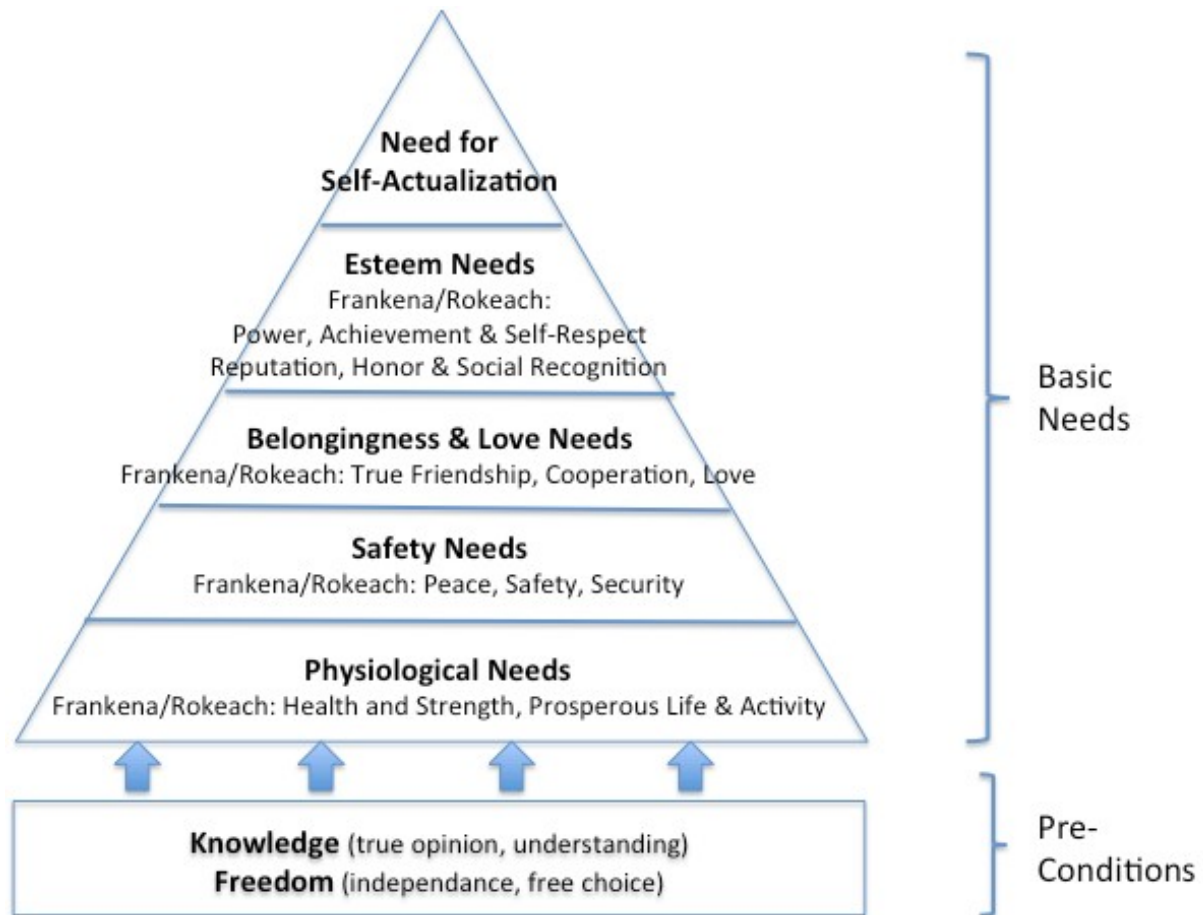


Figure x: Frankena’s and Rokeach’s list of values combined with Maslow’s hierarchy of human motivation and flourishing.

Exercise:

- Take the values that were identified in the scenarios of chapter 3 and align them with the values summarized in Maslow’s pyramid in figure x. Can you discern a hierarchy of extrinsic values based on this analysis?
- Study table x on the intrinsic values that the American philosopher William Frankena summarized. Compare this list with the list of values accumulated by psychologist Rokeach. Are there any differences between these two lists? Wherein do they reside?

Ethical Knowledge in Machine Age

*“He that knows nothing,
doubts nothing”
(1611, Cotgrave, “Rien”)*

Mens’ relationship with knowledge (episteme) has been an ambiguous one for millennia. There seems to be a deep fear that wanting to know too much can be dangerous; or at least knowing too much about what Ginzburg (1976) has coined “high knowledge”; that is insight into the secrets of nature (“arcana naturae”), God (“arcana imperii”) and power (“arcana imperii”). The bible tells us the story of Adam who followed his curiosity and ate from the tree of knowledge of good and evil. And as he did so, humanity was tossed from paradise (figure x). In his Epistle to the Romans, St. Paul warned the Romans “noli altum sapere, sed time” which has been translated and interpreted as an appeal to not know too much. But at the same time, already Aristotle noted that “all men by nature desire to know” (([Krobath 2009](#)), p. 215). Maslow talks about “the reality of the cognitive needs” (([Maslow 1970](#)), p. 49). And Kant called out provocatively, “Sapere aude”, “Dare to know”. History shows that the value to accumulate knowledge is a contested one.



Figure x: Adam und Eva Photo: Lucas Cranach The Elder, (Courtauld Institute Gallery)

What is Knowledge?

Philosophers ascribe the established definition of knowledge to Plato who saw it as “a justified, true belief” ([Ichikawa 2012](#)). In this definition three core components of knowledge become apparent. First, knowledge needs substantiation and justification so that the one knowing can be sure that what he knows is in fact knowledge and not just an attitude or a fake. Aristotle said: “we know something when we know the reason why something is the

way it is and can be sure that it cannot be otherwise.” ((Krobath 2009), p. 200). Second, knowledge needs to be present in a knowing subject; a person who “believes” in the knowledge artifact and for whom it is relevant. In other words: Knowledge needs a beholder. And third, knowledge needs truth: “we can say that truth is a condition of knowledge; that is, if a belief is not true, it cannot constitute knowledge. Accordingly, if there is no such thing as truth, then there can be no knowledge.” ((Krobath 2009), p. 213)

With the explosion of IT capabilities and an unprecedented capacity to collect data and information, analyze it, store it and combine it the term “knowledge” has gained tremendously importance. Scholars write about “The Knowledge creating company” (Nonaka et al. 1995) politicians propagate “The knowledge society” (Stehr 1994) and the creation of “Knowledge commons” (Hess et al. 2006). IT folks market databases that promise to be “Knowledge Management Systems”. “Big Data” sets lead scholars to talk about an “industrialization of knowledge” which is supposed to result from a confluence of (i) big data generation and collection, (ii) data processing and analysis, and (iii) data-driven decision making and control (OECD 2014 *forthcoming*).

The promise of “knowledge” created by IT systems is compelling. The traditional understanding and connotation of the term “knowledge” carries weight in people’s mind, because it stands for believable, justified and true phenomena. We have to be very careful though to not overstrain the term “knowledge” when we use it in the IT world as a synonym for all sorts of data processing. In many cases, IT investments have gone astray in past years when IT managers believed that they could really procure “knowledge” when buying a “Knowledge Management System” or a means for “Knowledge Discovery”, where really they only got a database or a visualization tool. So one first ethical question when it comes to a discussion of “knowledge” in an IT context is to ask about the conditions under which we are actually allowed to use the term in such a way to not mislead investors and users. Deceptive wording (i.e. in advertising) is a well-known problem in marketing. Based on an analysis of “deception by implication” in marketing communication (Hastak et al. 2011) I would argue that there is a risk that the term „knowledge“ has in some cases been misleading in an IT context. A „semantic confusion“ is often created among recipients who associate „knowledge“ with a true and justified belief, a property that many IT systems do not necessarily deliver.¹

That said, the scientific community makes a very clear distinction between data, information and knowledge that is particularly important when analyzing machine capability. Meyer (2007) summarizes the established view: *Data* is observed symbols, for example raw sensor data, sensor meta-data, a birth date, a name etc. *Information* is interpreted symbols and symbol structures, i.e. aggregated and “cleansed” sensor data or structured data sets. Floridi goes a step further and defines information as “well-formed meaningful data that is truthful” (Floridi 2005). *Knowledge* is then interpreted symbol structures or patterns that are used within a decision process (Meyer 2007). Knowledge is created, for example, when data scientists take the aggregated and cleansed sensor data and put it to statistical analysis to extract (if possible) causal models or when they develop higher level indices that can

¹ I am aware that the IT world has started to embrace its own definition of what knowledge is. For example, Hess and Ostrom (2007) write “knowledge [...] refers to all intelligible ideas, information, and data in whatever form in which it is expressed or obtained” (Hess, C., and Ostrom, E. 2006 *Understanding the Knowledge Commons* Cambridge, US, MIT Press.. Daniel Bell, cited in Cleveland (1982), defined information as “data processing in the broadest sense” and knowledge as “an organized set of statements of facts or ideas [...] communicated to others” (cited in OECD 2014).

support decision-making. Knowledge can re-enter the information base for further knowledge elaboration (for example, statistical factors re-entering a database for further analysis). Figure x illustrates the distinction of terms.

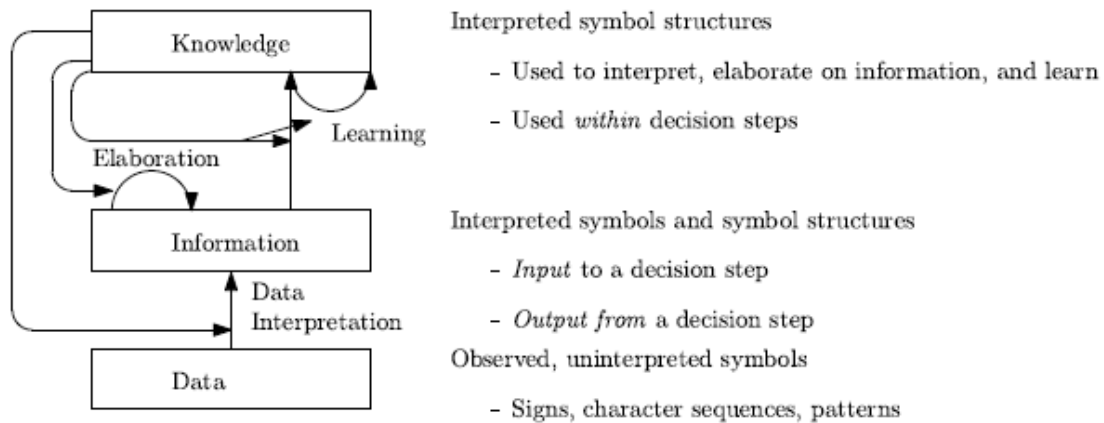


Figure x: The distinction of data, information and knowledge (taken from p. 6 ([Meyer 2007](#)) who adopted and refined the model of ([Aamodt et al. 1995](#)))

Structuring ethical challenges in IT driven knowledge creation

The scenario descriptions in chapter 3 show that future societies bear tremendous potential for us to become more knowledgeable through our machines. We will probably have ubiquitous access to knowledge and information when we need it and where we need it. Potentially, we have agents like Sophia's Arthur or the Wise Figure that search and filter information for us and make us see phenomena we cannot perceive naturally (such thermodynamic and magnetic information). They may aggregate and interpret information to some extent and then coach us and help us to learn at our individual level of knowledge capability. Ideally our machines will have instant and low-cost access to large parts of 'earth-knowledge' as well as 'earth-information'.

The amount of information to know about is exploding. I mention in the scenarios that people have gotten passive towards new information or also elapse some of it. They may stop to know. The latter loss must again be overcome by machines, such as Hal (in the robot scenario), who confronts Carly actively with the information she needs for her job. But then the question is what is an ethically correct way to select information that is relevant for a person and how can we ensure that such an information selection process by a machine does not integrate any bias? Who should be allowed to select what we should know? And how transparent does this process need to be?

The collection, aggregation, interpretation, access and use of information and knowledge can be depicted as a process structure (figure x). At each stage of this process we can observe distinct ethical challenges. More precisely, we have ethical challenges on two levels: One is looking into *whether* we should do something (propositional ethical questions): Is it ethically

legitimate *that* we collect certain data, aggregate, interpret it, access and use it? The other level of ethical challenge is of procedural nature. This puts the spotlight on *how* we actually create knowledge. How do we collect data? How do we aggregate it and make it accessible? How do we use what we know?

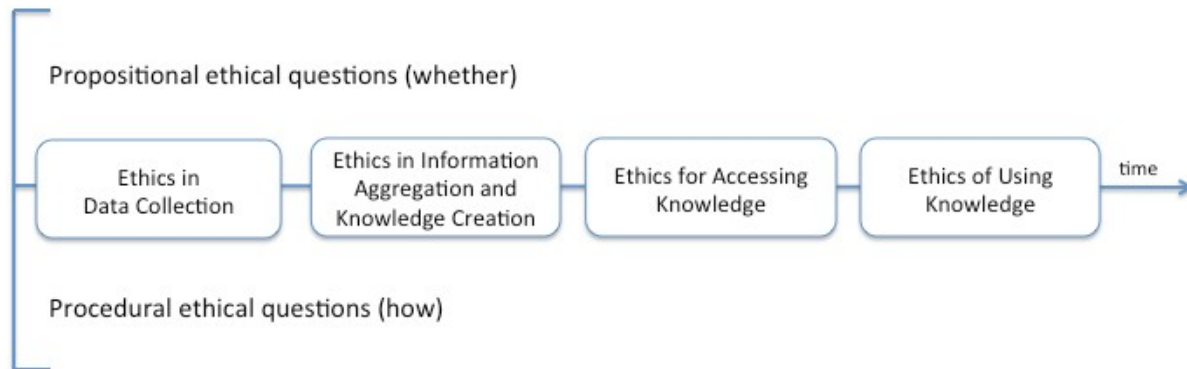


Figure x: A process structure for ethical challenges in knowledge management

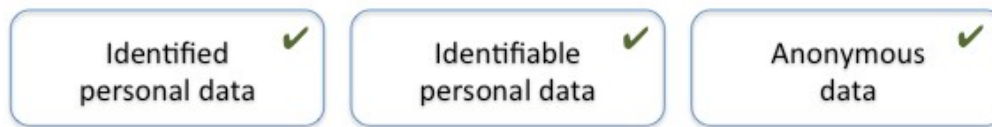
Ethical Challenges in Data Collection

When discussing data collection it is vital to first distinguish between data that is ethically sensitive and those that is not. Data collected for knowledge creation can be of personal and impersonal nature. Personal data, according to European data protection law either identifies an individual directly (i.e. through a social security number) or it is indirectly indicative of an individual.²

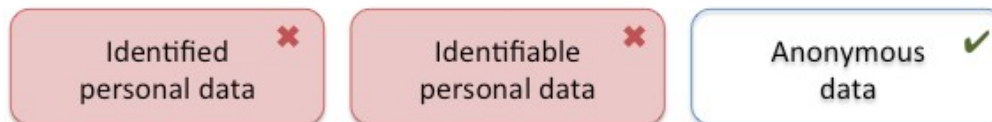
Typically, ethical questions arise only around the collection of *personal* data, and only so when that personal data is put to uses that go beyond the original purposes and reasons for which it was collected and is needed. For example, let's think back of the work scenario. The initial reason for collecting the virtual-reality data is to make the virtual world work and to make it technically interactive and responsive. There is no ethical issue in this. However, if that same data is logged and analyzed for a secondary purpose, which is to monitor employees' moods and behaviors and calculate “cut-throat probabilities”, then we get into an ethically problematic space. In this case a user consent for data use is required (see below). Figure x gives an overview of problematic and unproblematic data collection practices.

² *Identifiable* personal data allows for uniquely re-identifying a person from a larger pool of data. For example, there is probably just one woman that lives at my address and is born at the same date as me.

Data collected and used only for primary service delivery:



Data collected and used for **secondary** purposes beyond service delivery:



✗ ethically and legally problematic; requires user consent

✓ ethically and legally not problematic

Figure x: Distinguishing ethically problematic and non-problematic data collection

When personally identified or identifiable data is collected about us for secondary uses, then the European member states, but also the US and all OECD member states have acknowledged that there could be a potential for it to be used against us (harm us) and they have therefore set up protective regulation or guidelines ([European Parliament and the Council of Europe 1995](#); [FTC 2000](#); [Greenleaf 2011](#); [Organisation for Economic Co-operation and Development \(OECD\) 1980](#)). In Europe this restrictive legal approach to data collection has historical roots. The extent of the Holocaust in Europe was driven in part by the availability of personal data records about Jews, which fostered their systematic prosecution by Nazi officials (see figure x). For this historical reason as well as a legal case history of privacy breaches in the US (([Solove 2001](#); [Solove 2002](#); [Solove 2006](#))) personal data collection is regulated to some extent in most countries ([Greenleaf 2011](#)). Typically data collectors are required to minimize the personal data they collect about us and they are only allowed to collect it for a specified and legitimate purpose (see i.e. Art 6 of EU Directive). Some data categories are not allowed to be collected, except for exemptions or if people give their explicit informed consent to do so. These include personal data revealing racial or ethnical origin, political opinions, religious or philosophical beliefs, trade-union membership, and the processing of data concerning health or sex life (Art. 8, ([European Parliament and the Council of Europe 1995](#))).



Figure x: Records of Jews birthdate, name and town (registered by the Parisian police)

Today's data rich service world makes data scarcity approaches more difficult. In many cases companies and people find it beneficial to collect personal data for more than just service delivery. Think of the mood-barometer that was described in the work scenario where top-management is made aware of the emotional state of the company. Or agent Arthur who gives Sophie advice on the quality of products received from other buyers. Think also about the fitness feedback that Roger can receive from other wearers of the Talos suit. All these service examples have in common that they use (at least initially) personal data and aggregated and analyze this data to build valuable *secondary* information services on top of it.

Does this mean that innovative future services rely on personal data? No. Most of the rich data services I have described in chapter 3 can be build with the help of people's anonymized data sets. It is not necessary to maintain the "personal" nature of data sets in order to foster innovation around data. Let's take the Talos suit example. Roger's personal fitness data could be collected by his own personal agent and then passed on to the Talos service platform in an anonymized form. On that platform all anonymous Talos customers would then pool their fitness data for comparison and benchmarks without revealing their identities to Talos Corp.. They would hence technically exclude the risk that their Talos data could ever be sold or shared. Companies on the other side don't run into any ethical or legal problem while being able to deliver intelligent data-driven services.

Summing up, the propositional ethical question around data collection is whether we should collect *personal* data. Ethical system design as I will introduce it later (chapte x) would start with this question by challenging *whether* the collection of *personal* data is really necessary for the provision of a digital service or whether it is not possible to only use anonymous user data in the first place so that no ethical conflict or legislative issue can arise. Box X discusses anonymization of personal data in more detail.



Box 1:
Techniques for Pseudonymization and Anonymization of Personal Data¹⁾

Data can have different degrees of identifiability. Pseudonymous or anonymous use of data for service delivery protects individuals' privacy and makes data collection ethically or legally unproblematic.

Pseudonymous data means any personal data that has been collected, altered or otherwise processed so that of itself it cannot be attributed to a data subject without the use of additional data. This additional data should be subject to separate and distinct technical and organisational controls. Any re-attribution should require a disproportionate amount of time, expense and effort according to timely technical standards.

Technically the creation of pseudonyms could imply that separate databases for profile and contact information are created in such a way that common attributes are avoided. Steps should also be taken to prevent future databases from introducing common identifiers. Identifiers should therefore be generated at random. Any information that is highly specific to an individual (e.g., birth dates or contact data) should be avoided whenever possible. The general guideline for pseudonymous data is to minimize the granularity of long-term personal characteristics collected about an individual.

Even so, it may still be possible to individually identify a person based on transaction patterns. Pattern matching exploits the notion that users can be re-identified based on highly similar behavior or on specific items they carry over time and across settings. For example, mobile operators may be able to re-identify a customer by extracting the pattern of location movements over a certain time span and extracting the endpoints of the highly probable home and work locations. Typically, only one individual will share one home and work location.

Pattern matching does not always result in the identification of a unique individual. Often, a pattern may match multiple individuals. k-Anonymity is a concept that describes the level of difficulty associated with uniquely identifying an individual.²⁾ The value k refers to the number of individuals to whom a pattern of data, referred to as quasi-identifiers, may be attributed. If a pattern is so unique that k equals one person ($k = 1$), then the system is able to uniquely identify an individual. Detailed data tends to lower the value of k (for example, a precise birth date including day, month, and year will match fewer people than a birthday recorded without year of birth). Long-term storage of profiles involving frequent transactions or observations also tends to lower the value of k because unique patterns will emerge based on activities that may reoccur at various intervals. The values of k associated with a system can be increased by storing less detailed data and by purging stored data frequently.

In some cases, large values of k may be insufficient to protect privacy because records with the same quasi-identifiers do not have a diverse set of values for their sensitive elements. For example, a table of medical records may use truncated zip code and age range as quasi-identifiers, and may be k-anonymized such that there are at least k records for every combination of quasi identifiers. However, if for some sets of quasi-identifiers, all patients have the same diagnosis or a small number of diagnoses, privacy may still be compromised. The l-diversity principle can be used to improve privacy protections by adding the requirement that there be at least l values for sensitive elements that share the same quasi-identifiers.³⁾

Anonymous data' means any personal data that has been collected, altered or otherwise processed in such a way that it can no longer be attributed to a data subject; Anonymity can be provided if no collection of contact information or of long-term personal characteristics occurs. Moreover, profiles collected need to be regularly deleted and anonymized to achieve k-anonymity with large values for k or l-diversity with large values for l.

Data Types	Definition in terms of linkability	Protective System Characteristics
Personally identified data	linked	<ul style="list-style-type: none"> • unique identifiers across databases • contact information stored with profile information
Personally identifiable data	linkable with reasonable & automatable effort	<ul style="list-style-type: none"> • no unique identifies across databases • common attributes across databases • contact information stored separately from profile or transaction information
Pseudonymous data	only linkable with disproportionate amount of time, expense and effort according to timely technical standards.	<ul style="list-style-type: none"> • no unique identifiers across databases • no common attributes across databases • random identifiers • contact information stored separately from profile or transaction information • collection of long term person characteristics on a low level of granularity • technically enforced deletion of profile details at regular intervals
Anonymous data	altered or otherwise processed in such a way that it can no longer be linked to a data subject	<ul style="list-style-type: none"> • no collection of contact information • no collection of long term person characteristics • <i>k</i>-anonymity with large value of <i>k</i> • <i>l</i>-diversity with large values for <i>l</i> • differential privacy

Figure x:

- 1) Spiekermann, Sarah & Cranor Lorrie. „Engineering Privacy“. *IEEE Transactions on Software Engineering*, Vol. 35, 2009, pp. 67-82.
- 2) Sweeney, Latanya, “k-Anonymity. A Model for Protecting Privacy”. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, No. 5, 2002, pp. 557-570
- 3) Machanavajihala, Aswin et al., “L-diversity: Privacy beyond k-anonymity”, *ACM Transactions on Knowledge Discovery*, Vol. 1, No. 1, 2007

Informed consent

Lets assume that personal data needs to be collected and that anonymization or pseudonymization is *not* feasible. Then an ethical question arises around *how* that data is collected. There is widespread legal agreement that personal data should only be collected with the *informed consent* of data subjects.³

³ see i.e. Art. 19 of ([European Parliament and the Council of Europe 1995](#)), US Fair Information Principles including principles of “notice” and “choice” FTC, F.T.C. 2000 "Fair Information Practice Principles," F.T. Commission (ed.), OECD Privacy Guidelines Organisation for Economic Co-operation and Development (OECD) 1980. "OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data."

Consent is historically rooted in the Nuremberg Code⁴. Conceptually the obtaining of consent can be grouped into two distinct activities: One is to inform data subjects about data use intentions. The second is to obtain the consent from the data subjects. Informing about data uses means that consent seekers have to give *accurate* information about the *specific* purposes and reasons of personal data use as well as potential benefits and harms resulting from that use. Friedman et al. (2000) and the European Article 29 Working Party on Data Protection outline that companies need to meaningfully disclose their data usage practices *before* people consent (or decline to consent). “Meaningful disclosure” requires a company to state: (a) what data will be collected, (b) who will have access to the data, (c) how long will the data be archived, (d) what the data will be used for and (e) how the identity of the individual will be protected (p. 2 in (Friedman et al. 2000)). This information should be provided to users in such a way that it can actually be understood by them (principle of comprehension).

The second activity required for consent is to actually obtain it. People need to give their consent freely and voluntarily. They need to be able to exercise a real choice and there should be no risk of deception, intimidation, coercion or any other negative consequences if he or she does not consent. Friedman outlines that opportunities to accept or decline one’s personal data usage should be visible and readily accessible. The European Art. 29 Working Party on Data Protection comments that “consent must leave no doubt as to the data subject’s intention” (p.3 in x). Hence, the menus to accept or decline personal data uses should not be buried under myriad website layers or hidden in obscure locations such that data subjects cannot find them.

Figure x summarizes the requirements for obtaining consent as outlined by US and European scholars and regulators.

⁴ The Nuremberg Code was formulated in response to Nazi doctor’s experimentation with human subjects. The Nuremberg Code outlines how informed consent must be obtained and constituted for medical and health research purposes. The code has been adopted by the U.S. National Institutes of Health. This means that for health research it is required that human subjects consent to the collection and use of their data. This consent has to meet the following requirements: “The voluntary consent of the human subject is absolutely essential. This means that the person involved should have legal capacity to give consent; should be so situated as to be able to exercise free power of choice, without the intervention of any element of force, fraud, deceit, duress, over-reaching, or other ulterior form of constraint or coercion; and should have sufficient knowledge and comprehension of the elements of the subject matter involved as to enable him/her to make an understanding and enlightened decision. This latter element requires that before the acceptance of an affirmative decision by the experimental subject there should be made known to him the nature, duration, and purpose of the experiment; the method and means by which it is to be conducted; all inconveniences and hazards reasonable to be expected; and the effects upon his health or person which may possibly come from his participation in the experiment. The duty and responsibility for ascertaining the quality of the consent rests upon each individual who initiates, directs or engages in the experiment. It is a personal duty and responsibility which may not be delegated to another with impunity.” (retrieved on May 28th 2014 at <http://history.nih.gov/research/downloads/nuremberg.pdf>)

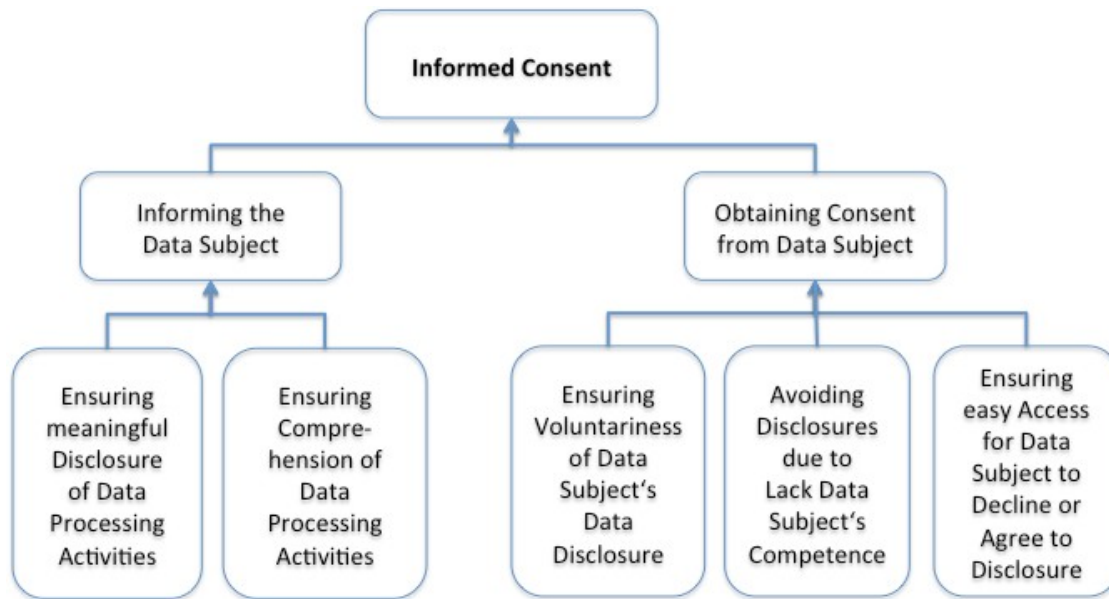


Figure x: Conceptualizing Informed Consent

To meet the requirement of meaningful disclosure outlined above, engineers can embrace protocols such as P3P 1.0 (P3P stands for Platform for Privacy Preference Project) as specified by the W3C in 2002 ([Cranor et al. 2006](#)). P3P describes web browsers that can read standardized machine readable privacy policies published by companies or governments. The openly accessible privacy policies are accumulated according to an XML format encoding a privacy taxonomy that embraces 17 possible data categories collected for 12 possible purposes, 6 possible types of recipients and 5 possible types of retention policies. Company information is pulled off the privacy policy by a user's web browser and can then be translated into a “privacy nutrition label” of the kind displayed in Figure x (figure x). An alternative for how to transmit privacy preferences and find the sources of how data is handled has been proposed by Tim Berners Lee and his group in the form of HTTPA (where the “A” stands for “Accountability”) ([Seneviratne et al. 2014](#)). (see also section x below). An easy way for visualizing data collection and handling practices for customers with the help of an adapted “nutrition label” is shown in figure x.

Bell Group

information we collect

ways we use your information

information sharing

	to provide service and maintain site	marketing	telemarketing	profiling	other companies	public forums
contact information		opt in			opt out	
cookies						
demographic information		opt in			opt out	
financial information						
health information						
preferences						
purchasing information		opt in			opt out	
social security number & gov't ID						
your activity on this site		opt in			opt out	
your location						

Access to your information

This site gives you access to your contact data and some of its other data identified with you

How to resolve privacy-related disputes with this site

Please email our customer service department

bell.com

5000 Forbes Avenue

Pittsburgh, PA 15213 United States

Phone: 800-555-5555

help@bell.com

we will collect and use your information in this way

we will not collect and use your information in this way

opt out

by default, we will collect and use your information in this way unless you tell us not to by opting out

opt in

by default, we will not collect and use your information in this way unless you allow us to by opting in

Figure x: An adapted “nutrition label” to inform users about data handling practices (taken from (Cranor 2012))

User Control in Ubiquitous Computing

Informed consent is a challenge as IT becomes more ubiquitous. “Weaving computing into the fabric of everyday life” is a high vision of the IT world called “Ubiquitous Computing” ([Weiser 1991](#)). This vision implies outspokenly that our natural environments should collect data about us human beings, nature and infrastructure around us at all times. I describe this vision in the mall scenario:

“Going through the mall’s main gate gives the mall implied consent to read out his and his kids’ data and send them tailored advertising and information. “Reading out” involves scanning clothes for RFID tags, recording movements and points of interest. Robots and on-shelf cameras analyze facial expressions and emotions. Video surveillance camera systems that embed security analysis screen their skin type and movement patterns.”

How can informed consent be organized in such a more complex environment where nutrition labels would probably overwhelm users? Timely, proposals foresee that personal software agents serve as mediators between the intelligent infrastructure and us (see i.e. ([Langheinrich 2003](#); [Langheinrich 2005](#); [Spiekermann 2007](#))). Agent Arthur is an example for this kind of mediating software entity. Personal agents can learn and store our privacy preferences and then permit or block requests to collect data about us. Requests for our data as well as data sharing can be logged on the client side ([Danezis et al. 2012](#)), as well as with the requesting data collecting entities. The latter may receive a kind of “sticky policy” with our data ([Casassa Mont et al. 2003](#)). These policies travel as metadata-tags with the information that is collected from us indicating to data controllers and processors whether, to what extent and under what conditions we allow for our data use ([Nguyen et al. 2013](#)). Figure x broadly illustrates the kind of privacy mediation process that could be implemented (taken from ([Langheinrich 2003](#))).

What is crucial in this invisible and ubiquitous data collection process in the long run is that people continue to exercise and perceive control over what is happening. How can this be done? To answer this question it is helpful to first conceptualize the construct of perceived control generally and then apply it to data collection.

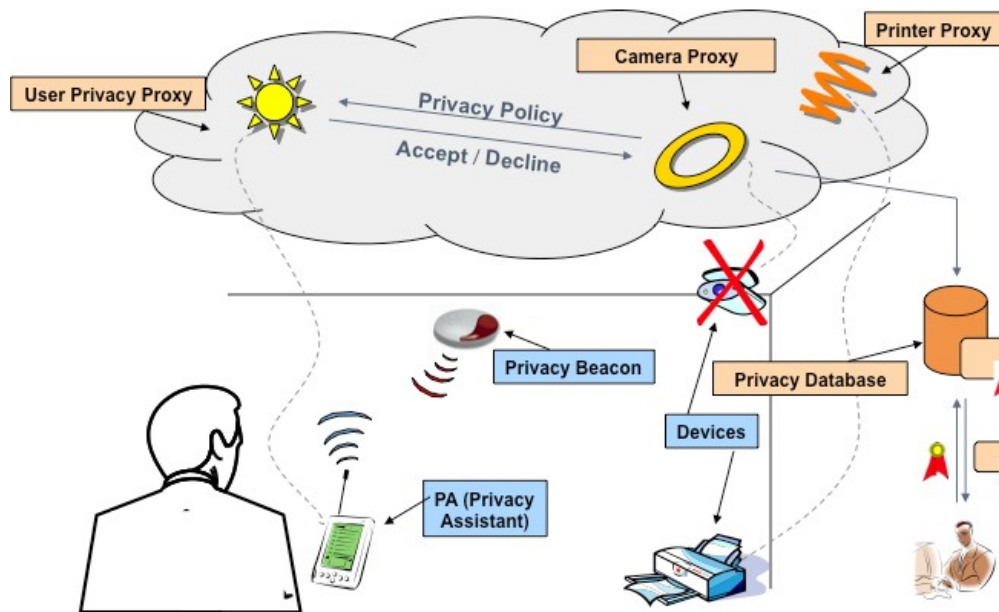


Figure x: Overview of privacy mechanisms for ubiquitous computing as proposed in ([Langheinrich 2003](#))

Perceived control is the conviction that “one can determine the sequence and consequences of a specific event or experience” (p. 385 in ([Zimbardo et al. 1996](#))). According to ([Averill 1973](#)) three types of control can be distinguished: cognitive control, decisional control and behavioral control. *Cognitive control*, which has also been coined “information control” implies that a person has the possibility to understand and interpret a (potentially threatening) event. ‘To know what’s going on’ (cognitive control) is a function of comprehensive and complete information on one side, but it also depends on people’s ability to absorb and understand that information on the other side. *Decisional control* is the opportunity to choose an action among several true choice options. True choice means that the options available must be affordable by the individual. Finally, *behavioral control* is gained when one is able to take an action and thereby directly affect, modify or regulate an event. For example, when one’s WiFi signal is weak and one can just walk to another room where the signal is strong again, then behavioral control is experienced.

Figure x summarizes the three dimensions of perceived control. Depending on the IT service, either all or just some of the controls need to be provided in order for people to feel comfortable. Through system design companies’ influence the levers boxed in bold: It is in their hands to decide what information they provide to customers, what choices they offer and how they implement user feedback.

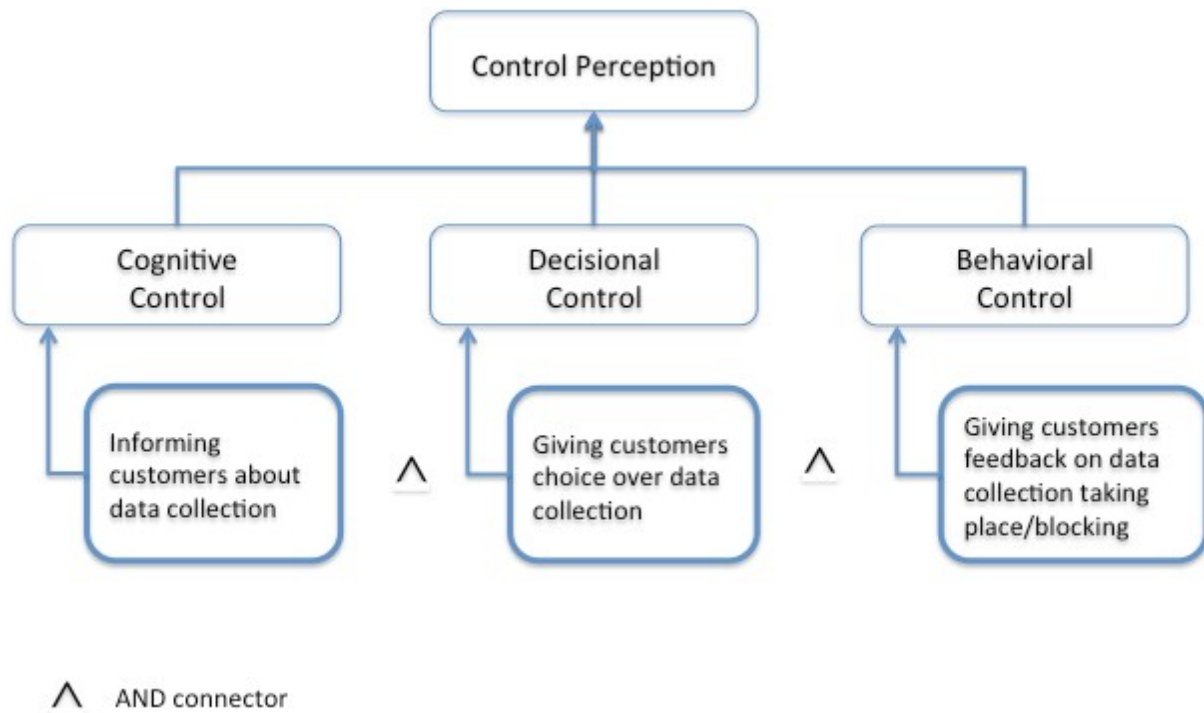


Figure x: Psychological conditions for perceiving control over an environment

Lets now transfer this conceptualization of control to data collection through RFID as described in the retail scenario. Studies on RFID technology have shown that customers are concerned that their personal belongings (i.e. their cloths) could be read out by RFID readers without their knowledge and consent ([Guenther et al. 2005](#)). The challenge of future retail environments therefore is to give people some perception of control over such invisible data collection practices. Information control would imply that a retailer informs customers of read processes taking place. This information should be provided in a form that is easily accessible and easy to read. For example, a notification could be sent to visitors' mobile phone. There could be signs at the mall entry, etc. Such first layer notifications could be linked to more detailed descriptions for all those customers who are really interested to understand the details of data collection and processing. In my stories above Roger is such a customer who wants to know what is going on and acts accordingly. The mall in my story above is transparent as to its data collection- and use practices. Besides informing customers about read processes, the mall gives its visitors a true choice over being read out or not. Customers who want to stay anonymous use a special entrance to the Halloville mall. Alternatively, they can use an agent, such as Arthur, to mediate data collection. The retail story I tell hints to the problem of affordability though. Roger sees himself forced into the data collection process because he cannot afford to forgo the discounts he receives in return for the data collected. If the mall gave price discounts to anyone, regardless of whether data is read or not, then visitors would have a true choice to participate in data collection or not and they would probably feel positively more in control when entering the mall. Fostering behavioral control

perceptions over RFID is more difficult than giving entry choices and information. For example, it is hard to proof for a mall operator that RFID read processes do not take place once a customer has opted out. As a result, RFID technology was shown to produce feelings of helplessness in people that could turn out to be an RFID implementation challenge for retailers in the long term ([Guenther et al. 2005](#)). Recording and visualizing data requests that have been blocked may be of great relief to untrusting customers.

Ethical Challenges in Information Aggregation and Knowledge Creation

Once data has been collected it needs to be aggregated to become information and knowledge. It is in this step of the knowledge management process that the main added value is created for companies and society at large. However, to truly create this value and speak of “knowledge” creation several challenges need to be overcome. One is that the collected data entering the information-processing phase as well as the resulting information product need to be of high quality. The other is that information aggregation and knowledge creation should be as transparent as possible in order to ensure that what we create is actually true and does not distort reality.

Data Quality

*“He who knows nothing is closer to the truth
than he whose mind is filled with falsehoods and errors”⁵
(Thomas Jefferson)*

An important prerequisite for knowledge creation is that the data used is of high quality. Data quality can be characterized as data’s *fitness for use* in a respective application context ([Wang 1998](#)). This fitness is not always given. “Data quality problems plague every department, in every industry, at every level, and for every type of information ... Studies show that knowledge workers waste up to 50% of time hunting for data, identifying and correcting errors, and seeking confirmatory sources for data they do not trust.” (p. 2 et seq. in ([Redman 2013](#))). This observation must be complemented by the fact that even commercial computer programs rarely come without bugs. Statistical surveys suggest that on average two to three mistakes can be observed for every 1000 lines of code even in professionally commercialized software products.⁶ Against this background, using digitally produced “knowledge” must always be regarded with a critical distance, respecting the potentiality for software mistakes, distorted data sources or misinterpretations of data sources.

Incidents of misinterpretation of data sources can be reduced if data is well described with the help of meta-data (see details in box x). Meta-data serve what Wang (1998) calls “representational information quality”: interpretability, ease of understanding, concise and consistent representation (p. 60 in ([Wang 1998](#))). In addition to meta-data it is advisable to

⁵<http://www.brainyquote.com/quotes/quotes/t/thomasjeff157254.html#gz1kxL4otu0cqKwk.99> retrieved on June 3rd 2014

⁶ <http://de.wikipedia.org/wiki/Programmfehler> (last retrieved on June 3rd 2014)

flag information with quality indicators that signal the degree of reliability to its users and provide the contact details of those units which actually produce the data ([Redman 2013](#)).

Figure x summarizes how data can be distorted in various ways ([Scannapieco et al. 2005](#)). The accuracy of data can be compromised due to simple mistakes in the syntax and semantics of data entries as well as duplicates. Often, data is not as fresh as it should be. As soon as data is not stable (like a birth date), but time-variable (such as an address or age) it needs regular updates. Time-related quality dimensions of data (which could be signaled to users) typically include the currency and timeliness of data. Currency measures how promptly data is updated. Timeliness measures how current the data is, relative to a specific task. The latter quality criterion is also relevant for web content (i.e. blogs or news articles), which, unfortunately, often lack the date of their publication.

A common problem in knowledge creation is that the data collected about a phenomenon or about a person is not complete. Lets take the example of an advertising network that has been able to collect the gender of most ad viewers, but for some ad viewers this attribute could not be observed. In this case the knowledge about these respective viewers is not complete. To speak with figure x, the attribute ‘gender’ exists and it is known that it exists, but it is not known to the advertising network. Completeness can be measured by the ratio of known attributes about a phenomenon divided by the total number of attributes.

Finally, data must be consistent. Relational databases often allow for automatically checking the consistency of data sets by looking at their integrity. A distinction is made between intra- and inter relation integrity depending on whether data is part of the same relation (domain) ([Scannapieco et al. 2005](#)). For example, the original year of a film publication must be before the remake year of a film.

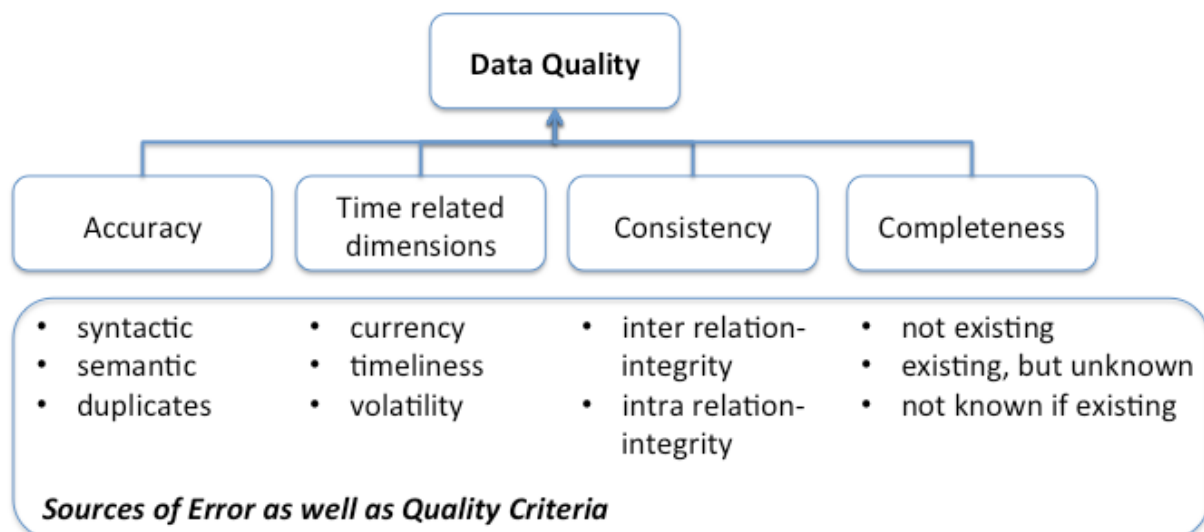


Figure x

To ensure high levels of data quality scholars have proposed Total Data Quality Management Procedures that should be implemented in software firms. They advise to equate today’s “information manufacturing” to traditional product manufacturing ([Wang 1998](#)). Figure x

visualizes this thinking.

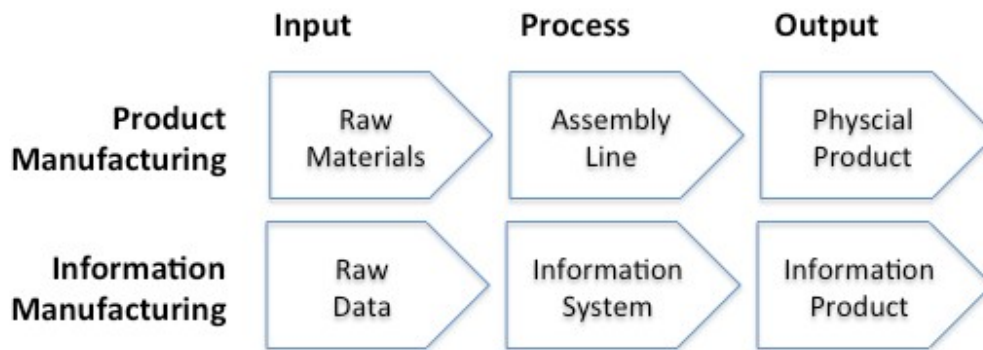


Figure x

Box X:
About the importance of metadata for transparency

For data to be meaningful and comprehensive and also in order to challenge its truthfulness, metadata is highly important. Metadata is “data about data” – for instance, units of measure. Metadata allows data analysts to recap the conditions under which the data has been originally collected and to understand the true meaning of the data collected. Only against the background of this knowledge it is possible for data analysts to further combine and analyze the data they use and draw meaningful conclusions from them.

Harvard Business Review¹⁾ reports on what can happen if metadata is not specified: NASA lost a \$ 125 million Mars Climate Orbiter because one group of engineers used English units for distance (feet and miles) while another unit used metric units for key operations. Another corporate example illustrates how important metadata is to assess a company’s success: Lets take a typical firm in which several words may be used to describe a “customer”. One employee talks about a “customer” when he means that repeated business has been done with a client. Another employee at the same firm may talk about a “customer” as soon as a first contract has been signed. A third colleague may be very bullish by character and will speak about a “customer” already when there is only a sales lead. So what is a “customer” finally for a company? Metadata rules specify what a company will mean by a “customer” and IT systems then reflect this uniform understanding upon which the market and sales operations can be monitored.

The examples illustrate: It is important that meta data rules are defined and it is equally important that they are documented and communicated within a firm.

1) Thomas C. Redman, “Data's Credibility Problem“. *Harvard Business Review*, 2013, pp. 2-6.

Truth

*“Some may want the truth all the time,
and many may want it some of the time,*

I have outlined in the introduction to this knowledge chapter that knowledge requires truth. Spence even argues “information without truth is not strictly speaking information, but either misinformation or disinformation” (([Spence 2011](#)), p. 264). Yet, as the data quality section above has shown, errors and misinterpretation of data create risk that what we believe to be knowledge is really misleading. Hence, when working with machines we always need to be cautious as to the extent we can fully trust their output. We also need to be aware that a lot of machine output that appears as perfect knowledge at first sight is really just a spotlight on probabilities. Take the Google search engine’s result of a person query: The search results page is not a perfect overview of who the person is that is queried. It is just a selection of data on a person with a high probability of relevance. Unlike the common belief that machines are ‘right’ and hence more reliable than people, data scientists and experts will agree: Machine output is rarely perfect.

If we seek truth on the basis of the kind of ubiquitous data collection described in the scenarios above a number of ethical challenges can arise. Let’s go back to the retail scenario where I describe how the mall’s robot choice algorithm determined that female teen-robots would be best suited to accompany some males with pedophilic tendency in their shopping trips.

“Another issue related to the humanoids’ looks. They typically resembled people quite realistically and had all kinds of looks and sizes. Some of them looked like teenage girls and boys, and everybody thought that these younger looking robots were used in the mall as peers for kids. But then some men got teenage female robots to be their shopping companions. And a whistleblower found that this robot choice, recommended by Halloville’s IT system, was related to the system’s knowledge of pedophilic tendencies for some male customers (because they had visited the teen sex porn categories on porn websites on the Internet).”

Several questions arise out of this scenario: Do we want this kind of knowledge to be created at all? Are we sure that we want to produce knowledge on the basis of *all* the data we collect? Who should be allowed to establish such truths about us? And who is liable if truth damages our reputation?

Some scholars have taken a rather critical view on creating truth arguing that it may paralyze people and impact their productivity: “The question whether truth has value or whether knowledge thereof has a destructive effect is old. Who increases knowledge, increases pain; knowledge paralyzes action; consciousness entangles in fear and disturbs the natural course of lively processes; the reach out for knowledge is the fall from grace.” (([Jaspers 1973](#)), p. 142). How would a man feel if he read in the press that the mall assigns teenage robots to pedophiles and he had a teenage girl crossing the mall with him the day before? Or another example: How would we feel if a genetic screen test showed that we have cancer in the next year or so with 80% probability? What does this knowledge do to us? Will it change our lives to the better or to the worse? Has knowledge of our own truth the potential to lead to self-fulfilling prophecies?

As a society we have not found final answers to these questions that will be relevant for us in the years to come. Popper took a very positive perspective in this regard. He argued: “...it is only through knowledge that we are mentally set free – free from enslavement by false ideas,

prejudice and idols. Even though this endeavor of self-education does not exhaust our meaning of life, self-education can decisively contribute to make our life meaningful.” (([Popper 1974](#)), p. 201) If an algorithm found a pedophile tendency in a person would it not be good for that person to find out about herself or himself more clearly? In the future work scenario above I gave an example for enhanced self-awareness: what if algorithms identified bullying behavior in companies and allowed employees to learn about and delve into their behavior retrospectively? Machines are dispassionate about truth. So they are able to hint towards a version of the truth that may be different from the one that we will sometimes want to create about ourselves or remember in a certain way. Machines will force us into a different perspective on ourselves. I presume that one of the biggest challenges of the future will be how we humans will be able to handle this perspective that some would claim to be our ‘objective’ truth.

A very important question in this context is how and by whom we are to learn about our own presumed truth. Should everyone be allowed to establish (a presumable) truth about us and tell or not tell us about it? Shouldn’t we have a say in *who* is allowed to know something about us? In the work scenario I outline how knowledge creation about us could be organized: Companies may collect a lot of data about us to provide services or increase security, but in order to be allowed to and able *to* analyze the data further or use it for secondary knowledge creation purposes they could be required to ask for our permission. Such an obligation to request permission could be organized as outlined above, with the help of sticky policies ([Casassa Mont et al. 2003](#)), dynamic consent mechanisms ([Kaye et al. 2014](#)), personal agents ([Langheinrich 2003](#)) and meta data architectures ([Nguyen et al. 2013](#)). In the work scenario I go even so far as to foresee full data encryption policies that allows personal data only to be used when an individual provides her private key to decrypt her data. Some analysis, such as the company’s emotional mood analysis could still be done on encrypted data ([Gentry 2009](#)).

Another angle to look at truth and the ethics of knowledge creation is to question whether *all* data sources should be equally used in machine calculations? Especially in future times where it may be that almost all of our real-world activities will be recorded by some computer system it is questionable whether we want to process and use all of this data. Some thinkers have proposed an “ethics of ignorance” along the lines of George Pettie (1548–1589) “So long as I know it not, it hurteth me not.”⁷ This approach would imply that we simply decide that some data is not important to consider. We could abstain from collecting this data, we could delete some of it right after collection, give it very little algorithmic weight or forget (delete) it over time. As I have outlined above, European data protection law at the moment effectively integrates a similar approach called “data minimization” (that applies however to all personal data). Also according to US legal case history not using all data is justified on two grounds: one is that some data may simply be too confidential by nature to be collected. The other one is that some data may be too sensitive to be transferred as such a transfer would be equal to blackmailing someone. Lets transfer these arguments to the pedophile scenario above: Collecting, storing, transferring and analyzing porn category choices made on sex-video hubs is technically easy to do. However, we could consider it as simply too confidential to be done and so we – as a society – could decide that knowledge about ‘sexual orientation’ will simply not be created. We may outlaw it even. A search engine company would then simply and automatically delete all search queries that relate to sex or sex categories right after the query has been made. The problem is though that if we go down this road, then - as a global society! - we need to agree on *what* categories of data we don’t want to know about and that we *really* don’t want to know about them. **Our personal and corporate curiosity**

⁷ Presentation at the Oxford Internet Institute entitled „The Ethics of Ignorance“, by Google’s European Director of public policy, Nicklas Lundblad, May 2014

will be our own biggest enemy in making this decision wisely. The emblem depicted in Figure x stemming from Ovid’s *Metamorphoses* depicts the case where three women’s curiosity let them to open a box of knowledge that they later regret to have opened, seen its horrific content.⁸

A way out of this dilemma (to decide what to ignore) could be to *not* ignore, *but* instead create more transparency and awareness around what data is collected and used and by whom. Do sex-video hubs actually store and share category choices? If we had such a primary transparency of corporate practices we would be better prepared to take responsibility for the truth. Potentially we would be able to adjust our behavior according to our counterpart, just as we do it in the offline world.



Figure x: Picture from Ovid’s *Metamorphoses* where Herse, Pandrosos and Aglaulos are too curious and open a box the interior of which horrifies them.

The search for truth is not only a matter of what data is collected, but also a matter of the kinds of analysis we put the data to. Should algorithms be allowed to put our data to *any* analysis possible no matter the ends? Often we may be surprised about the results of data analysis. Even having seen the raw data we would not have expected a specific outcome. Take the example of a German bank that created seven psychological profiles about its customers to better sell insurances and stocks to them. Customers were characterized as preservationists, hedonists, adventurers, wallowers, performers, tolerants or disciplined.⁹ Do we want our identities to be classified like this? And more importantly, do we want to be systematically treated according to them?

The choice over data analysis remains very much with the personal ethics and virtues of the data analysts today as well as the companies they work with. Company policies can provide guidance as to what kind of data analysis should be allowed and what not. In order for social norms and pressures to play out, it would be good to log and openly publish analysis practices that are being done *regularly* on personal data. Julie Cohen writes that we should have

⁸ Herse, Pandrosos and Aglaulos, daughters of Cecrops, were given the task of guarding a box carrying the infant Erichthonius (which was the result of an attempted rape of Athena by Hephaestus. The girls opened the box and were horrified to see a half-man, half-snake, who later grew up to be king of Athens. See Ovid *Metamorphoses*, 2.252ff. taken from <http://www.emblems.arts.gla.ac.uk/french/emblem.php?id=FA075> (last retrieved on June 4th 2014)

⁹ <http://www.spiegel.de/wirtschaft/unternehmen/verkaufshilfe-sparkasse-sortiert-kunden-in-psycho-kategorien-a-727133.html>

“protocols for information collection, storage, processing and exchange.” (([Cohen 2012](#)), p. 1932) Such transparency would be an important lever to control for personal and company curiosity.

Transparency

In the above section on truth in knowledge creation I outline why transparency is vital. True knowledge can be created only through reason, scrutiny and high quality data. But reasoning and scrutiny as well as data quality need regular monitoring, challenge and incremental improvement, especially as technology advances. For these reasons transparency has been embraced as an important political target. The EU Commission embraces transparency as essential in achieving corporate social responsibility, social justice, environmental security, true democracy and well being ([European Commission 2001](#)). In his first memorandum after assuming presidency Barack Obama embraced transparency as a key tool to promote accountability ([The White House 2009](#)). Most multinational companies have embraced the principle of transparency in their codes of conduct ([Kaptein 2004](#)).

However, the kind of “radical transparency” we may need ([Thompson 2007](#)) possesses a number of properties not easily met by companies and governmental institutions. Figure x gives an overview of the information quality criteria necessary to create true transparency (as opposed to opacity...).

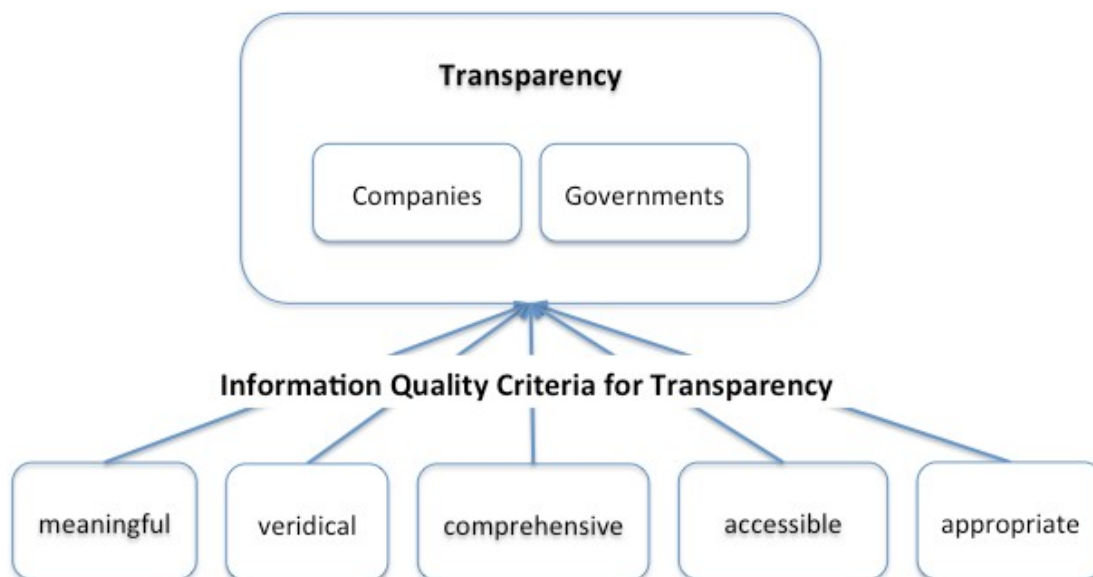


Figure x

Note that the word “transparent” has an almost diabolic ambiguity. In modern times we mostly associate transparency with *visibility*. The borders of a company’s black-box processing are made transparent for us to look inside. However, originally transparency means *invisibility*. We look through something without recognizing it, like through a window glass. This ambiguity in the word reflects some of the reality of many transparency initiatives: Companies and governments can provide us with a lot of information, but the question is

whether that information is the right one to give a true insight into what is happening within corporate boundaries. Take the case of ENRON, once a widely heralded US energy group that went bankrupt when false accounting practices were uncovered. The company was regularly audited. So it provided for an apparent transparency of its practices. But in truth it used opaque instruments to obscure the real bases of profits and bonuses.

“The information disclosed, when implementing information transparency, is supposed to consist of meaningful, veridical, comprehensive, accessible and useful data. This is not a mere litany of properties” write Luciano Floridi and his colleagues (([Turilli et al. 2009](#)), p. 108) Meaningful data means that the information that is given out has gone through some kind of elaboration process and that it conveys a message that has significance for a recipient in a respective context. The significance dimension of transparency is further stressed by the quality criteria of ‘usefulness’ that I rephrase as ‘appropriateness’. “Credibility does not arise from details, but from appropriateness,” writes Armando Menéndez-Viso (2009). He refers to Descartes (1596 – 1650) who once described how a good portrait does not necessarily reflect every detail of a person, but the main lines: “We must observe that in no case does an image have to resemble the object it represents in all respects, for otherwise there would be no distinction between the object and its image. It is enough that the image resemble its objects in a few respects. Indeed the perfection of an image depends on its not resembling its objects as much as it might.” (Descartes, cited in ([Menéndez-Viso 2009](#)), p. 158) As with most portraits, painters are in the dilemma that the one being painted would like to appear in an optimal light. So he or she has to make a decision as to the degree of truthfulness that is reflected by the image. To go back to the corporate world: Arthur Anderson, the auditing company of ENRON, produced meaningful data about its client. Information that was also comprehensive in the sense that it could be easily read. Yet, the audits did not convey the truth of the factual financial status of the company. Arthur Andersen was found guilty of obstructing justice. Once one of the world’s top auditing companies it voluntarily surrendered its licenses to practice as a result of the scandal.

Finally, an important dimension of transparency is accessibility of information. So far, access to company information is only granted with special permission to access proprietary company databases. Little information can also be found on the Web, such as annual reports. This public information is mostly unstructured. Gradually, this situation is changing though. For example, in the US the Gramm-Leach-Bliley Act requires financial institutions since 2009 to publish information about their privacy practices and has given guidance on the structure information should have. The EU and some US States legally oblige companies to publish data breach notifications. Companies in the EU may be required to publish the results of their privacy impact assessments.

Moreover, people in the US and in Europe have the right to access all personal data that a company holds of them. For example, in the US the Privacy Act provides people the right to access their records. So do the Cable Communications Policy Act, the Fair Credit Reporting Act and the Children’s Online Privacy Protection Act. In Europe, the European Data Protection Directive 95/46/EC provides data subjects with a “Right to Access” (Art. 12, ([European Parliament and the Council of Europe 1995](#))).

Standardized and, in particular, *machine-readable* access to one’s data through the Internet change the game of company- and government transparency. While in the past, no one was able to access all the material published and compare it, watchdogs, the press and even private individuals can now increasingly crawl and analyze what has been published or leaked. Such an increased accessibility may incentivize companies to adjust certain behaviors that would

otherwise have remained undetected.

Economically, accessible company information should be beneficial. It helps to reduce information asymmetries in the market between companies and consumers. Customers get a clearer view of who they are dealing with and can make informed choices on who to entrust with their data for what kind of returns. Competition between companies is fostered as a structured analysis of competitive practices becomes feasible. IT companies can compete on the basis of ethical conduct.

Ethical Challenges in the Design of Knowledge Access

One of the most important levers for the flourishing of people and society is our access to the information and knowledge exploding around us. Philosophers consider the access to information as a ‘primary good’ in modern societies ([van den Hoven et al. 2008](#)): “Access to information is relevant to every conceivable plan of life” (([van den Hoven et al. 2008](#)), p. 383). Through the Internet we feel the benefits of efficient access already today. Many of us can download scientific articles online at the click of a button where in former times we had to go to physical libraries that often did not even have what we were looking for. We have free encyclopedias such as Wikipedia at our fingertips. We can instantly and easily search for what we don’t know and want to learn about. Most important: Search has become so easy that we mostly find what we are looking for (which has not necessarily been the case in earlier days). Short: Many of us *have* access to a lot of the world’s information and knowledge and *the way* we are accessing it is hugely more efficient than it used to be in analog times.

As the stories in chapter 3 showed this current situation is just the beginning. In the future we may have technologies in the form of virtual bodies, such as agent Arthur or the Wise Figure, which can explain to us what we want to learn about. They may be so smart that they can adapt the way they communicate with us to our level of knowledge; (just as games today bring us to the next higher level of performance depending on how we succeed at lower level tasks). Accessing knowledge in such a playful way, bears huge potential for the development of cognitive skills in our future societies. Yet, I also mention in the educational story that Roger wonders whether his financial means suffice to afford the holographic private home teacher he wanted to hire for Sophia and himself. I hint to the fact that access to knowledge may not come for free.

In fact, access to services that transmit valuable information and knowledge is rarely for free. Even if the “free”- culture of the Internet suggests this on the surface, access has been and is *conditional* in most cases. Conditional means that – even if we have broadband Internet connectivity – we somehow pay directly or indirectly for service use. We do so either by paying money for information. For example, many news portals now charge for articles. Alternatively, many Internet services condition their service use on the right to monetize our personal data traces. Personal data is a new *currency* with which we indirectly pay for access. Meglena Kuneva, the EU’s former Consumer Commissioner, expressed this economic reality when she said: “Personal data is the new oil of the Internet and the new currency of the digital world”.

Unconditional access to services would, in contrast, imply that services are truly free of cost. Besides the money side, this would imply that we can access the information and knowledge services without paying for them with our personal data and that we can access them publicly (from home or from a public library) without needing to be a member of a specific group; such as a university (educational institution), a company, or an association (e.g. a

standardization organization). One of the most valuable sources of knowledge we have today is our scientific knowledge that is published in academic journals, conference proceedings or standardization documents. These journals, proceedings and standards are, for the most part, not publicly accessible. Independent innovators, small companies, consultants, or interested individuals often fail to meet the condition of access to this knowledge, because they can either not afford the amounts requested or do not have the time to join privileged groups with access.

Figure x summarizes how access to knowledge can be classified today as conditional or unconditional.

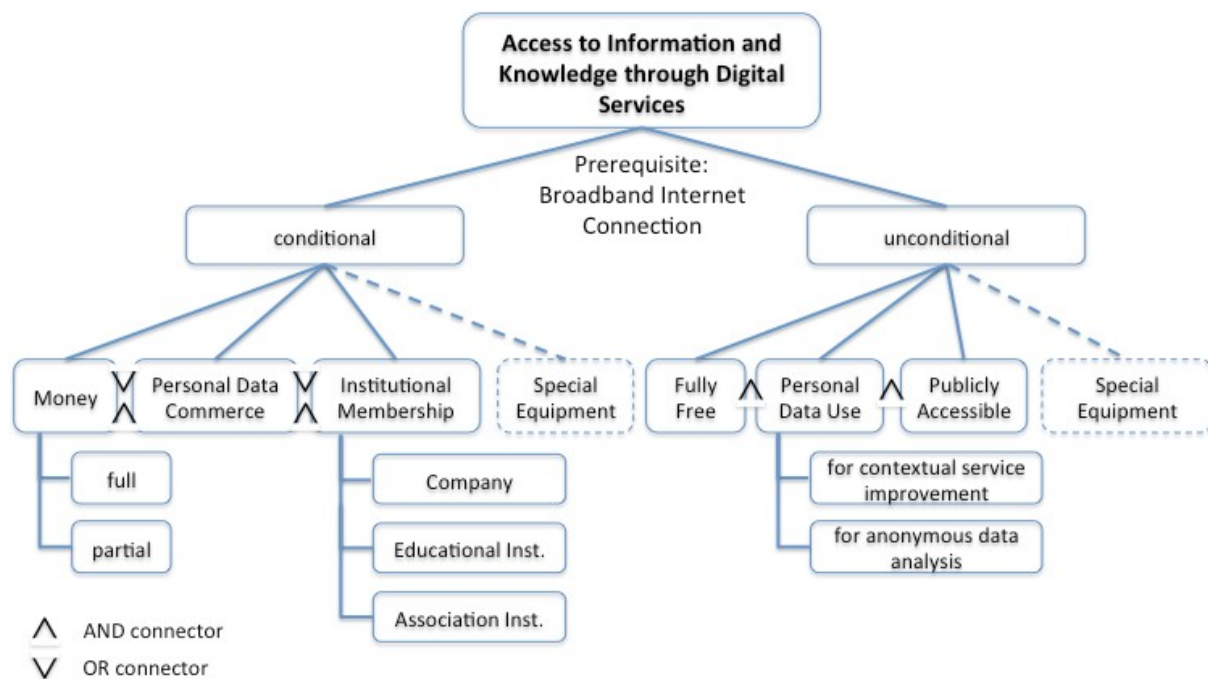


Figure x: Conditional and unconditional access to information and knowledge

Distinguishing conditional versus unconditional access is important for service design, marketing and public policy development, because it untangles the true complexity of access to information and knowledge. Conditional knowledge access has consequences for the digital divide in societies as well as for innovation. Hereafter I first want to concentrate on the effects of conditional access on the digital divide. Below I then reflect on the effects of conditioned access for personal and social development. In chapter x I describe how conditioned access to software, patents and content can hamper creativity innovation.

Access to Knowledge

The Digital Divide is defined as the „stratification in the access and use of the Internet“ ((Ragnedda et al. 2013), p.1). It is of concern to most governments who want to give citizens equal opportunities to participate in the digital service world. Heavy investments have been made to continuously improve the infrastructure coverage of countries in terms of Internet connectivity. As a result, the digital divide in terms of connectivity to the Internet has been reduced significantly in many developed countries. Yet, there still is an important digital

divide in the way people use their access to the Internet. Many people officially own an Internet connection, but don't use it. Many use it, but only so for gaming and entertainment and not for accessing knowledge. The reasons for this "use-divide" are often attributed to individual factors of users, such as their age, sex, ethnicity, employment, etc. Van Dijk critically points out that we tend to "(simply blame) inequality of access on attributes of individuals such as a lack of motivation" ([von Dijk 2013](#)), p. 29). But in truth, the divide is also a matter of people's position in society and whether this position allows them to meet the conditions of access to knowledge. Van Dijk talks about a "relational or network approach" to explain the digital divide. Take the example of employed vs. unemployed or educated vs. uneducated people. One way to argue is to say that people who are unemployed or uneducated are simply lazy and it is therefore not surprising that they don't use the Internet or only use it for chatting and gaming (not effectively participating in the knowledge part of the Web). Another perspective is to note that access to a lot of valuable knowledge on the Web is really conditioned on being a member of a privileged group, such as a university or a company that hold access rights (see above). People who are unemployed often simply don't have access to knowledge through their employer. Even if they wanted to access, they could not. A recent study showed, for example, that of the 114 million English-language scholarly documents on the Web, only 27 million (24%) are available free of charge ([Khabisa et al. 2014](#)).

The stories in chapter 2 describe how only gamers of *Playing The World* have access to the saint figure that can answer them all questions. In addition, only those players who can afford an extra € 5 per month can port their personal agent to the real world who give them access to information (Sophia's agent Arthur). This sounds like small amounts. Note though that the price tags could be much higher.

Another 'deal' to get access to information is described in how the mall functions in the retail scenario. This deal is the one for personal data:

Going through the mall's main gate gives the mall implied consent to read out his and his kids' data and send them tailored advertising and information. "Reading out" involves scanning clothes for RFID tags, recording movements and points of interest. Robots and on-shelf cameras analyze facial expressions and emotions. Video surveillance camera systems that embed security analysis screen their skin type and movement patterns. In return, Roger gets 3% off all his purchases in the mall plus free parking. The only exception is Sophia, who is able to use Arthur to reliably block her personal information exposure and provide her with neutrally tailored product information.

... rich people who are on a truly anonymous scheme can use a separate smaller mall entrance on the east side of Halloville that does not track any data. People's personal agents (a kind of app running on their mobile phones; function-wise similar to Arthur) block RFID read-outs and send their owners' data usage policies to the mall infrastructure, indicating that video and voice data must be deleted. However, when people go through that entrance, they don't receive the 3% discount he gets and have to pay for parking, a luxury that Roger cannot afford. Personal robots are also available only for an extra charge and base their recommendations on the personal agents of those richer folks.

The story outlines how people's access to unbiased information and a protection of their privacy may cost them both money and effort (to use a separate mall entrance). This is problematic from an ethical perspective. As Edward Spence outlines "The epistemology of information ... commits its disseminators to certain ethical principles, values and virtues, such as honesty, sincerity, truthfulness, trustworthiness and reliability, and fairness, including justice, which requires the equal distribution of the informational goods to all citizens" ([Spence 2011](#)), p. 264). Currently such an ethical standard is not the reality though. Instead, a

digital divide could start to widen between those of us who can pay for unbiased information access and consciously chose to protect our personal data versus those of us who need to trade personal data and live in “filter bubbles”.

Objectivity or Filter Bubbles

The term “filter bubble” has been coined by Eli Pariser ([Pariser 2012](#)). It refers to the fact that many information services, such as the social network service Facebook or the search engine Google pro-actively filter the information that is provided to us. More precisely, they *personalize* the information that is displayed to us. Google has been reported to use 57 different variables to decide what search results are shown ([Halpern 2011](#)). This means that every user gets a different answer to the same search query term. Large data brokers and advertising networks have accumulated databases that hold and track our daily behavior online and represent each of us with between 1500 and 2500 personal attributes. This information is then used to select individually relevant ads for us as we surf the Web. According to Eli Pariser, Facebook characterizes its users on multiple dimensions, including political attitudes. When a Facebook user is observed by the platform to hold a conservative political attitude she will from thereon receive mainly the conservative news from her network ([Pariser 2011](#)).

In her article on “Mind Control & the Internet” Sue Halpern (2011) analyzes why this way of serving filtered information can become problematic for political and democratic stability. She describes the example how Republicans and Democrats in the US have gained very different perceptions about climate change between 2001 and 2010. While in 2001 49% of Republicans and 60% of Democrats agreed that the planet was warming, this relatively similar perception of reality fell apart in the following years. Exposed to probably different news feeds only 29% of Republicans believed in global warming by 2010. In contrast, Democrats’ awareness of climate change increased in the same time period. By 2010 70% of Democrats believed in global warming. “When ideology drives the dissemination of information, knowledge is compromised”, she writes ([Halpern 2011](#)). In other words: when people are enveloped in their personal ideologies by personalized digital media, then their perception of reality is narrowed down to what they have always tended to believe anyways. They are not challenged any more. They are captured in a narcissistic spiral of “information cocoons” that continuously re-enforce themselves.

To avoid such developments scholars have pointed to the importance of *objectivity* in the way information and knowledge is provided to us and can be accessed by us ([Tavani 2012](#)). There are various ways to define objectivity. Box x gives a short historical introduction into how perspectives have varied over the course of time on what objectivity actually is and how objectivity relates to art and modern management. For sure, and throughout time, the goal of objectivity has been to capture reality truthfully and free from bias and emotions. In its definition of “objective evidence” Anglo-American law gives a hint to two characteristics of objectivity: that is information can be used in court as objective evidence when it is relevant and genuine. Furthermore we can distinguish between (1) objectivity in the way information is selected and (2) objectivity in the way information is distributed to multiple users. Figure x summarizes this conceptualization of objectivity.

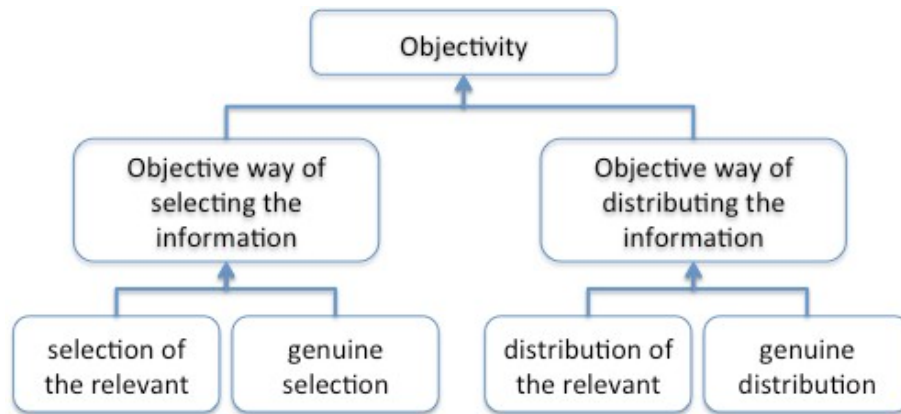


Figure x

We can apply this conceptualization of objectivity to online information and knowledge access. Lets take search engines as an example. ([Tavani 2012](#)) points out that two traditional criteria to rank search results are relatively objective. According to these criteria a site is more *relevant* the more other pages link to it and the more visits it has. *Genuineness* of the selection is provided when search companies make an effort to include (index) a maximum number of available sites regardless of their content. The distribution of information is objective if we as users can all see the same sites scrolling down the long list of search results (genuinely) and see them in the same order of relevance. The search results in this case would of course only be minimally ordered alongside the objective selection criteria given above. To stay with the terminology of objectivity described in box x, “trained judgment” would be required to make sense of the search results returned. As in the old public library days we would be forced to scroll through long alleys full of search results to find what is relevant for us. In doing so, we would find a lot of stuff we need and a lot of stuff we don’t need and thereby explore and learn.

Box x:
On Objectivity and its Relation to Modern Data Driven Business

In their book “Objectivity” (2007), Lorraine Daston and Peter Galison distinguish between three kinds of representation in science that have dominated over the course of history: Truth-to-Nature, Mechanical Objectivity, and Trained Judgment ([Daston et al. 2007](#)).¹⁾ These different ways to “objectively” describe reality can be best understood when looking at three different pictures Daston and Galison have used to illustrate their argument (see also Cohen 2008²⁾).



Figure x a
Campanula folis hastatis dentatis,
Carolus Linnaeus, *Hortus*
Cliffortianus, drawn by
Georg Dionysius Ehret and
engraved by Jan Wandelaar
in 1737

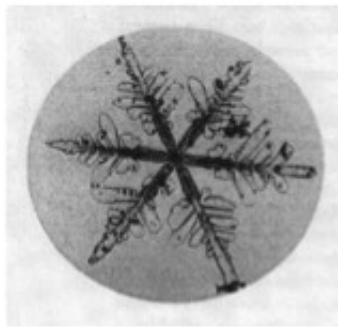


Figure x b
From Gustav Hellman, with
microphotographs by Richard
Neuhauss, 1893

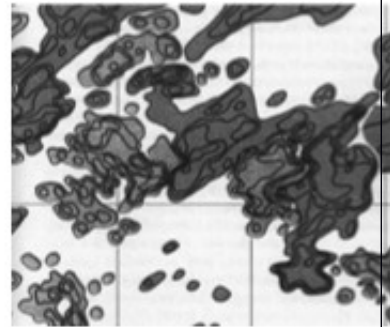


Figure x c
By Robert Howard, Vaclav
Bumba, and Sara Smith, *Atlas of*
Solar Magnetic Fields, 1959,
derived from the Observatories
of the Carnegie Institute of
Washington, DC.

Originally scientists used illustrations to describe nature. They aimed to capture the underlying ideal truth in nature (Figure x a). This meant that they did not work out all the specificities that each specimen could have in their diversity, but they were more interested in the fundamental recurring and essential characteristics of nature's reality. This scientific period was followed by a time of thinking in terms of „Mechanical Objectivity“ (Figure x b). Here scientists were trying to not let their own imagination influence their perception of reality, but to stay as closely as possible to what they could see. At the same time, of course both their selection and perspective did influence objectivity. Mechanical objectivity found its zenith in the late 19th century, supported also through the advent of photography. Figure x b shows how all the salient details, formations and peculiarities of the snow flake are captured.

A radical break with this kind of understanding of objectivity can be observed in Figure x c that is an example of what Daston and Galison call the “trained judgment period” of objectivity. Figure x c is an “image of the magnetic field of the sun [mixing] the output of sophisticated equipment with a ‘subjective’ smoothing of data—the authors deemed the intervention necessary to remove instrumental artifacts...” (p. 21). This 20th century handling of objectivity places much more responsibility on both the scientist who creates the image and the audience that receives it. The scientist “smoothens” the data. He “cleans” it, as modern data scientists would say. And then he leaves the audience with the challenge of making sense of this reality.

This 20th century invitation of scientists to *construct* objectivity is also reflected in the modern art of this time and contemporary management. The artistic works of the Cubistic period for example (see Picasso's “Poet” in Figure x) expect a considerable maturity and thinking ability from their spectators. Any business analyst or modern manager that is confronted with piles of data, graphics and statistics on his business will be able to recognize that looking at this input produces a similar reaction (of helplessness?) than looking at a cubist painting.

What modern art has done and what contemporary business intelligence output is doing is to constantly try to understand reality through its decomposition and reassembly. In art one of the most radical statements of decomposition is Malewitsch's Black Square (figure x b). In modern management an individual sensor data point would reflect this. Objectivity is both enhanced and relativized. Enhanced, because the composed individual data point is there for a fact. It is not gut feeling or (highly plaid) personal opinion. On the other hand, we relativize objectivity through subjective reassembly. The way we clean and aggregate and interpret our sensor data points really bears a lot of subjectivity that again relativizes objectivity. Yet, this objectivity is nevertheless accepted as truthful by management, analysts and society today.

This 20th century way of creating both more and less objectivity is an achievement, because we are

getting closer to truth through our technologies while at the same time being challenged to understand it. This process is allowing us to become what Kant would call "mündig". We are challenged to develop "trained judgment". And through this effort we grow.



Figure x a
"The Poet" (1911)
Pablo Picasso

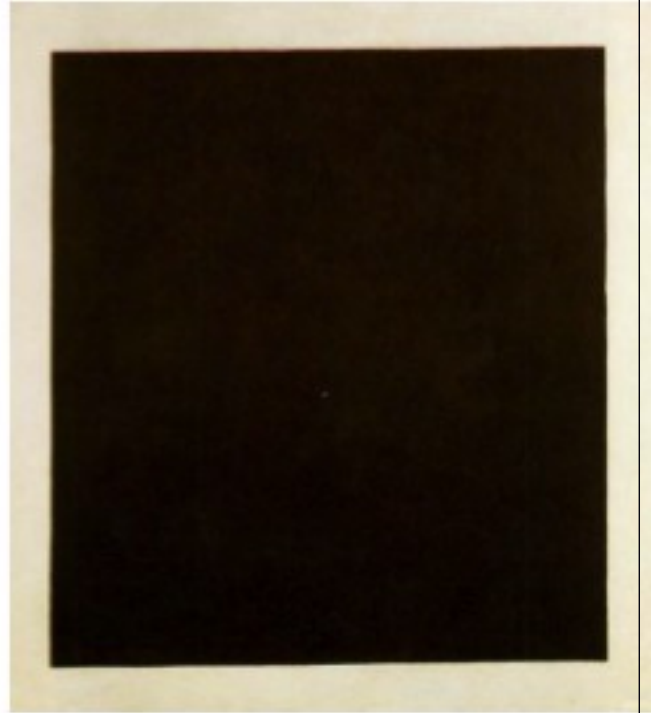


Figure x b
"Black Square" (1915)
Kasimir Sewerinowitsch Malewitsch

- 1) DASTON, L. & GALISON, P. 2007. *Objectivity*, New York, Zone Books.
- 2) COHEN, B. 2008. Objectivity: True-to-Nature, Mechanical, and through Trained Judgment. *Science Blog* [Online]

Teaching us "trained judgment" is not the way search engines took. Efficiency and time cost have been the primary values that search engine companies have embraced. Since governments have not outlawed the manipulation of search results, search engine companies have found themselves forced to replace the objectivity in their access criteria with obscurity (to avoid the manipulation of their search results). There is little transparency as to whether search engines like Google are objective in the way they select the sites they rank. Introna and Nissenbaum have reported that certain sites are systematically excluded ([Introna et al. 2000](#)). What we can observe is that they are not objective in the way they distribute information to us, because they anticipate what is relevant to us and then personalize the information they display. To speak with the terminology of objectivity, today's search engines (just as all other companies who filter content for us without consulting us) throw us back into 19th century "mechanical objectivity".

What could an ethical design for search engines or social networks look like when we reason along the concept of objectivity and embrace the Aristotelian virtue of *philotimia* (healthy ambition; controlling the urge to be superior and decide for others what is good for them)? The "trained judgment" perspective (box x) would embrace that data companies select the information for us and we need to trust in their selection (just as the scientists in figure x c smoothed the magnetic field image so that it is at all readable). Objective access design could set in when it comes to the distribution of content and here ethical companies would need to allow us as users to determine relevance ourselves. Assuming that content is initially listed genuinely, transparent end-user tools for prioritization of content should then be put at our disposition for us to play around with the content, determining for ourselves what and who is relevant for us and how far back in time we want to search.

The reason why we should use such tools to determine for ourselves what's important and what isn't has a lot to do with flourishing. Scholars like Aristotle, Wiener and Terrell Bynum all agree: the purpose of human life has a lot to do with information processing. People need to engage in a diversity of information processing, organizing, remembering, inferring, deciding, planning and acting activities. Even if our machines were able to do all of this for us, the question is to what extent they should do it for us completely. The right balance must be found instead so that people can still do the final selections and also then carry responsibility for the information use.

Ethical Uses of Information and Knowledge

A large part of the information and knowledge we gain in the future through our IT systems will somehow be based on data about people. Video cameras, drones, enhanced video and AR glasses, etc. monitor individual and social interactions in public space. Sensors on our body or integrated into our homes as well as the public infrastructure will potentially measure our movements, electricity consumption, communications, noise levels, etc. As soon as our IT systems directly or indirectly monitor us in this way and subsequently use the data for more than the initial collection purpose, then privacy issues emerge.

In 2014 Microsoft published a study together with the World Economic Forum entitled "Rethinking Personal Data: Trust and Context in User-Centred Data Ecosystems" ([World Economic Forum 2014](#)). In this study the company investigated from the individual's perspective, what constitutes the acceptable use of personal data. Almost 10.000 people were interviewed in eight countries from Europe, Asia and the Americas. They indicate that the way data is collected from people and the way data is used influence acceptability. The

collection method investigated in the study referred to various levels of control people could have over the collection process. This control turned out to be the most important factor for respondents' service acceptance (31-34% of overall data use acceptability is driven by this construct). The finding re-emphasizes the importance of the chapter above on ethical data collection practices, the necessity to give people choices, ask them for consent and build perceptions of control.

The second most important driver of acceptability is how the data is then used. "When data is actively collected, users prefer scenarios where the use of the data is consistent with what they originally agreed (p. 8)." Purposes of data use can be explicitly agreed upon when data is initially collected from users or by giving them preference options that can be revised dynamically at any time (see i.e. today's Facebook privacy settings). However, a lot of the information we exchange in our everyday communication is apt to implicit agreements on data sharing. We typically expect that the information we share in a specific role and context is treated according to the ethical information norms of the respective communication context. For example, in the role of being a patient we share details of our health with a doctor. Implicitly we expect the doctor to respect the information norm to keep our health information confidential. Helen Nissenbaum has called this kind of implicit agreement a respect for the "contextual integrity" of information use ([Nissenbaum 2004](#)).

Contextual Integrity

Contextual integrity recognizes that societies have developed norms of information flow and use. Two types of norms are particularly important: The first set of norms regulates the "appropriateness" of information flows. Appropriateness means that within a given context the type and nature of information exchanged may be allowed, expected or even demanded. It is allowed, expected and demanded to share medical information with one's doctor for example. The second set of norms relates to the distribution of data, the "movement, or transfer of information from one party to another or others" (Nissenbaum, 2004, p.122). For example, when a medical professor working at a university clinic collects information about the personal health status of his patients, then we probably accept that this professor (who is not a regular doctor) uses the information not only to have a history of our health development and to potentially compare our case to other cases. These are expected and demanded uses. We are likely to also accept that as a university professor he wants to integrate patient cases into his research. The university context of the professor allows for this kind of data use. What we would not find appropriate is that patient data is also distributed outside of the university context; i.e. that the clinic sells the health records to insurance companies or international data brokers. Such a use of data would probably be judged as inappropriate and as breaching norms of distribution. Figure x visualizes contextual integrity and its potential breaches.

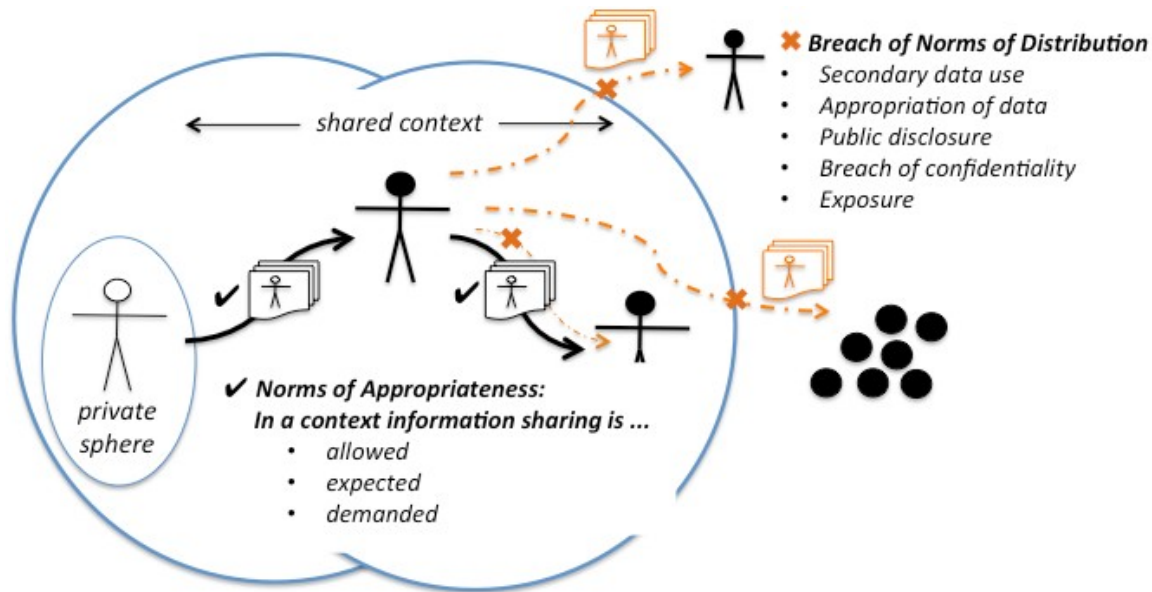


Figure x: Contextual Integrity and Data Sharing Norms

Privacy Harms

The respect of contextual integrity in data and information use has found widespread acceptance by governments and industry. The World Economic Forum (2012) has pointed to the need for context-aware usage of data as a key outcome of global dialogs it had conducted as part of its “Rethinking Personal Data Initiative” ([Nguyen et al. 2013](#); [World Economic Forum 2012](#)). Equally, the US FTC and the Whitehouse have embraced the concept ([The White House 2013](#)). In his development of a “Taxonomy of Privacy” Daniel Solove discusses privacy harms resulting from breaches of contextual integrity ([Solove 2006](#)). These are unforeseen and unwanted secondary uses of data, appropriation of personal data, public disclosure, breach of confidentiality and exposure.

Secondary data use involves using our information in ways that we do not expect *and* might not find desirable. Sex sites on the Web sharing category preference data with third parties is an example for such an unwanted secondary data use that is at the same time a breach of privacy. When a transmission of personal data furthermore involves monetary transactions that we ourselves do not benefit from, then we can even speak of appropriation. *Appropriation* means that our data is used to serve the aims and (potentially monetary) interests of others. Solove calls this “exploitation” (([Solove 2006](#)), p. 491). In the stories above I give the example of United Games selling their VR customers’ gaming data to third parties. The data leaves the gaming context in which it was collected and United Games benefits from the sale without its customers sharing in the profit. From a privacy perspective the data sale may not necessarily be harmful. Yet, contextual integrity is breached when the data is used.

Public disclosure of private matters means that data is made public, which is not of legitimate concern to the public and which is at the same time offensive to a reasonable data subject.

“Jeremy had filmed one of his teachers with his new AR glasses when she made a mistake in front of the class, and he had then published this mistake on YouTube.”

Public disclosure focuses on the content of a message being disclosed; in this case on the teacher making a mistake. The harm caused by public disclosure typically involves damage to the reputation of an individual. This damage is caused by the dissemination beyond context boundaries and to a larger group. In the case of Jeremy, the teacher’s mistake should have stayed as an information within the boundaries of the class. Solove (2006) notes that public disclosures bear risks for people’s long-term reputation. Employers may base decisions on information they learn from some other context. With easily accessible digital information people can become “prisoners of (their) recorded past” (p. 531).

Unlike the tort of public disclosure, the tort of *breach of confidentiality* does not require that a disclosure be “highly offensive”. A breach of confidentiality occurs between friends and acquaintances, but also in fiduciary relationships, such as with doctors, employers, bankers, or other professionals with whom we engage. In the future work scenario in chapter 3 employees hold private keys to their transaction data. Still the HR department has used a “cut-ties” algorithm to suspect Carly of wanting to leave the company. Obviously there must have been a breach of confidentiality by the employer, seen that HR could apply the algorithm to her data without her consent. Breaches of confidentiality violate the trust in a relationship, because information is passed on that should be kept between parties.

Finally, *exposure* is one of the strongest forms of breach of contextual integrity, because there is certain information about us that we want to keep in our immediate private sphere, such as certain physical and emotional attributes (denoted in figure x by the circle around the data subject). These are attributes that we view as deeply primordial, and their exposure would create embarrassment and humiliation. “Grief, suffering, trauma, injury, nudity, sex, urination, and defecation all involve primal aspects of our lives—ones that are physical, instinctual, and necessary. We have been socialized into concealing these activities.” ((Solove 2006), p. 533). If such information of us is exposed to others, then it rarely reveals any significant new information that can be used in an assessment of our character or personality. Yet, the exposure creates injury because we have developed social practices to conceal aspects of life that we find animal-like or disgusting. „Exposure strip-ping people of their dignity”, writes Solove (p. 535). Reports in the press of teenage girls stripping on Skype for their boy friends, who then publish these private videos via e-mail or on Facebook is an example for the tort of exposure.

Computer Bias and Fairness

I have outlined above how objectivity is important for us in accessing knowledge. A special form of lack of objectivity that undermines our dignity, honor and potentially self-esteem is when machines treat us with bias. In line with Batya Friedman and Helen Nissenbaum I use the term bias to refer to “computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others” (p. 332, (Friedman et al. 1996)). Discrimination means that due to a machine judgment we are denied certain opportunities or goods or confront some undesirable outcome. Information about us is used against us. For example, when a machine decides that we are not credit worthy and as a result we are denied a loan or confront a prohibitively high interest rate. Alternatively, the machine may decide that we are rich enough to pay for higher priced flight tickets. Certainly such forms of discrimination are undesirable by the people who are impacted. Figure x summarizes the concept of machine bias.

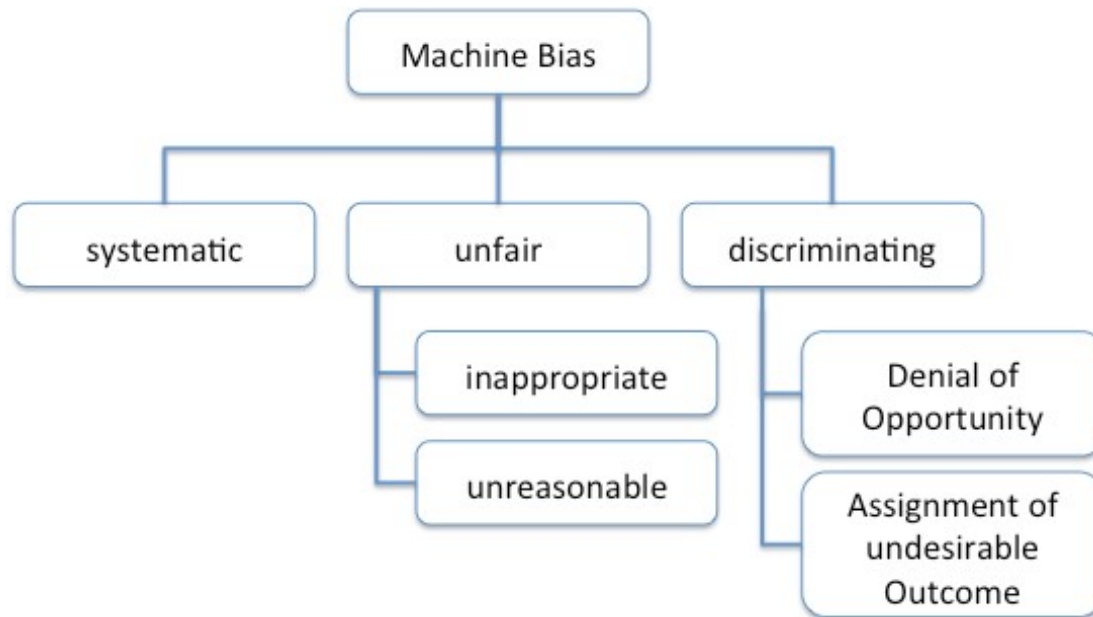


Figure x: Machine Bias according to ([Friedman et al. 1996](#))

Yet, let's think about the flight ticket example again. Couldn't it be considered fair that rich people pay higher flight ticket prices than poor people? Don't we have some forms of price discrimination all along that distinguish between rich and poor and that are considered fair? Senior tariffs for public transport or free entry for kids are examples. Unfairness perceptions arise according to Friedman and Nissenbaum when the behavior of machines towards us is “inappropriate” or “unreasonable”. They specify the loan example described above: If a person is denied a loan, because she has continuously failed to pay her bills in the past then it seems appropriate and reasonable to deny her the loan. There is no bias. Note that bias is only created when the access is “systematically” denied. If a loan is not granted once, because a person has recently not paid her bills then this is not a bias. Only if from now on the person systematically does not get a loan any more (even though she does start paying her bills again) then this is a bias.

What is fair information use? Synthesizing a number of influential fairness theories I broadly want to distinguish between opportunity-based fairness perceptions and equality-related fairness perceptions. By “opportunity” I mean a person's possibility to influence the outcome of events. By equality-related fairness perception I embrace the fact that people compare themselves to others and build fairness perceptions when they are treated similar to those to whom they compare themselves.

Opportunity based fairness can be built by companies through procedural justices and desert based distributive justice. Procedural justice is created through (1) transparency of the processes (procedure) by which information and knowledge about us is used and (2) an opportunity to influence this process ([Lind et al. 1988](#)). Take the example of Jeremy who suspects that he does not get into university because of his outdoor times. Denying access to a student on such intransparent ground would not be perceived as fair or procedurally just. In contrast, if it was publicized in due time that only those students get access to a certain university if they have spent sufficient time outdoor, then applicants can adjust their behavior in advance and have the opportunity to make it. That said, a challenge for future societies will be that Big data analysis will suggest all kinds of beneficial and detrimental behaviors that we humans should try to live up to or avoid. Meeting all the requirements that follow out of such

analyses may be procedurally just and fair (such as a university requiring pupils to have been outdoors). Yet, the number of requirements machines will figure out to be useful may overwhelm us humans. Procedural justice may have the high price of people being enslaved by their "information CVs".

Another set of theories related to fairness perception is dealing with distributive justice. Distributive justice is a philosophical concept that has informed political thinking of how economic benefits and burdens should be distributed in a society. Strict egalitarians call for the equal allocation of material goods and services to all members of society. On such philosophical grounds, machines could have a lot of information and knowledge about people, but the use of this knowledge would not be allowed to lead to any differential treatment. Lets say a bank would know about the different degrees of credit worthiness of people. According to egalitarian distributive justice the bank would not be allowed to adjust loan terms accordingly. John Rawl (1921 – 2002) in his Theory of Justice (1971) introduced a slightly different theory of distributive justice called the 'difference principle'. The difference principle corresponds more to what I call equality-related fairness perception ([Rawl 1971](#)). Rawl would argue that differential unequal treatment can be fair, but only if it leads to the benefit of the least advantaged in a society. So, lets say, a bank knew that some of its customers are rich and others are poor and would then use this knowledge to make the rich pay more for a loan than those who are poor. Finally, distributive judgment has also been based on merit ([Lamont 1994](#)), which is more of an opportunity-based fairness approach in the sense that people earn access to resources; they 'deserve' them.

While philosophers' reflection on fairness is rooted in political thinking about how to best distribute goods in societies, psychologists have studied the concrete construction of fairness perceptions in people as a response to the environment. Procedural justice described above was shown to psychologically influence fairness perceptions in various contexts (see i.e. ([Cox 2001](#)) ([Campbell 1999](#))). Equally, equity theory ([Adams 1963](#)) outlines how people build up fairness perceptions. Equity theory posits that individuals who are similar to one another gauge fairness (or equity) of an exchange by comparing the ratios of their contributions and returns to that of peers in their direct reference group. Lets take again Jeremy's denial of access to university. Equity theory posits that Jeremy will compare himself to other pupils in his school who he feels similar to. If they have worked similarly hard in school and are of similar intelligence then Jeremy would find it unfair that they get access to university and he does not. Inequity leads to 'social tension' that people strive to reduce. Such a reduction of tension can take place already by changing one's reference group. Instead of comparing himself to other students of similar intelligence Stanford Online could inform Jeremy that the reason for non-admittance was his behavior in school and his track record of breaching his teacher's privacy. They could let him know that all pupils with similar behaviors are not admitted to Stanford Online. Jeremy could then compare himself to those pupils who committed similar tort. Seen this comparison, Jeremy would then not see his dismissal as unfair any more. The example shows that following equity theory online companies are well advised to communicate to customers their reference group. Potentially they could give them also the chance to intervene and change the reference group that the company chose for them initially.

Philosophers have added the Principle of Equality which entails that for any two people, A and B, if it is ethically permitted for A to treat B in a certain way, then, in ethically similar circumstances, it is permitted for B to treat A in a similar way.

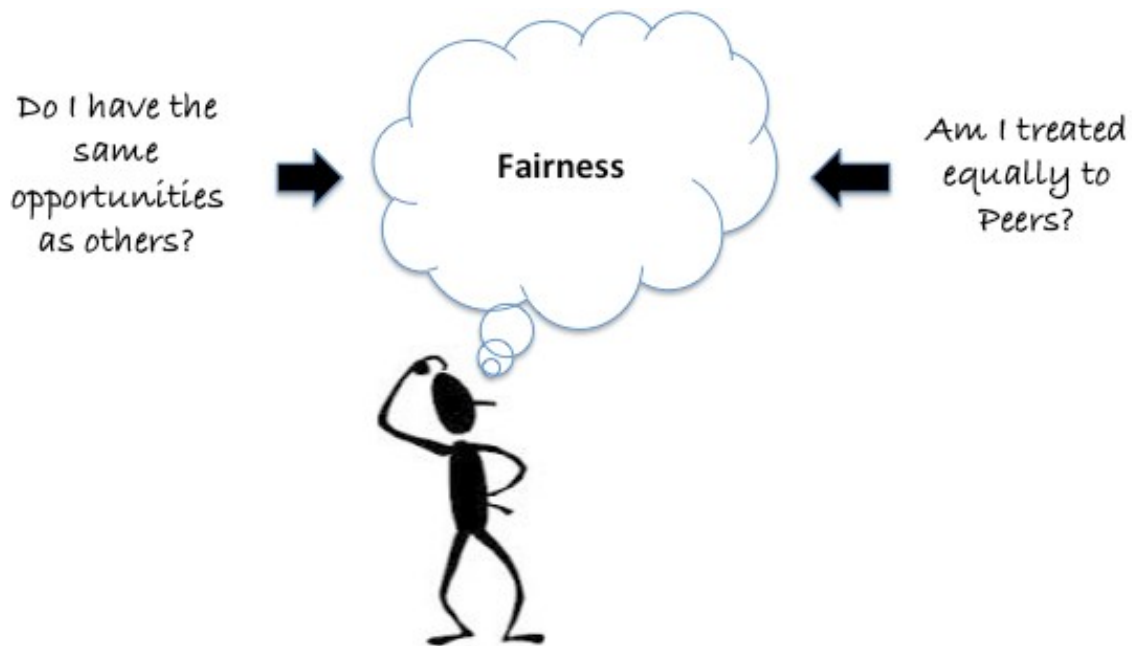


Figure x: Origins of fairness perceptions

Summing Up: Ethical Knowledge Management

*"Cyberspace is the land of knowledge,
and the exploration of that land
can be a civilization's truest highest calling."*
(Ethster Dyson, 1994)

Knowledge is an intrinsic value, a recognized primary good to societies across the world. For entitling this chapter I have used the term "ethical" knowledge. Ethical knowledge is what we create if we embrace and respect the intrinsic value characteristics outlined in this chapter.

At the basis of knowledge is the data and observations about the world and about ourselves. Collecting this data without the control and informed consent of the people is not advisable. Not only is consenting a legal necessity. Consent and control also ensure the long-term availability of data. If people lose trust in the data collection infrastructure and feel excluded and out of control it is likely that they quit. Psychologists know that humans avoid environments in which they are out of control ([Mehrabian et al. 1974](#)). If they are forced to stay, they physically suffer ([Langer 1983](#)). Both of these reactions are certainly not an aim of the knowledge society we want to build.

Information aggregation and knowledge creation must withstand the expectations of truth. Without truth there is no knowledge. And truth at all times needed discourse and challenge. If we don't allow for transparency in our knowledge creation processes and have our information aggregations and conclusions regularly challenged, then we are likely to hamper progress. We will create suboptimal half-truth on the basis of which neither science nor management nor economics can operate efficiently in the long-term. Current data quality problems in companies are a warning pre-cursor of such developments.

When it comes to knowledge accessibility an ethical vision of society, as a free –minded and mature community of individuals, implies that as much people as possible should be granted the possibility to learn the information and knowledge created. Learning about the world should however not be pre-determined by simplistic filter bubbles. Human beings have a high capacity for filtering reality themselves and according to their own needs. Machines should respect this human capacity and give people multiple filter technologies that allow them to find the “objective” truth for themselves.

Finally, knowledge is a treasure and an asset if it is produced ethically. Societies should treat it as such as they use it. Using knowledge against the people, breaching their privacy, undermining fairness and establishing ubiquitous machine bias will undoubtedly be a short-term strategy on which a knowledge society cannot flourish.

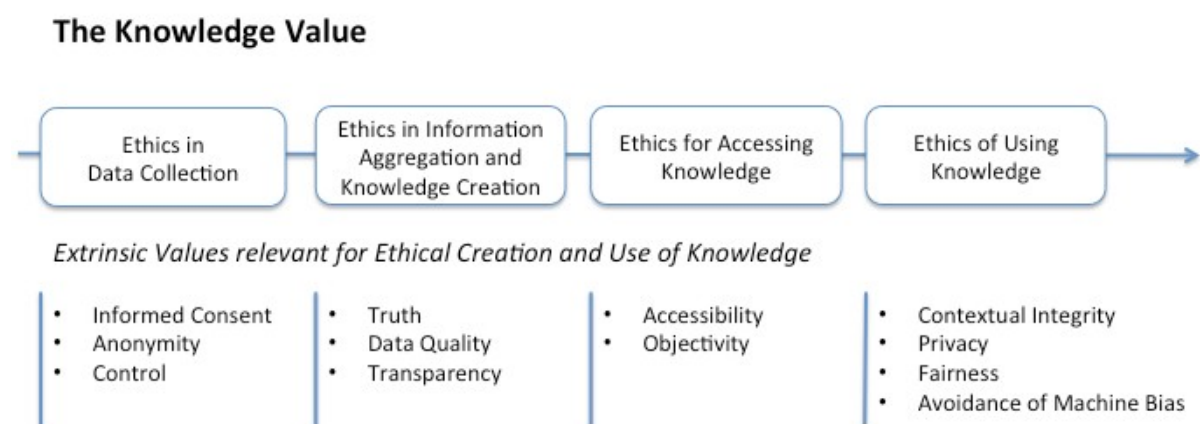


Figure x: Summary of ethical knowledge creation and values supporting it

Figure x summarizes all those extrinsic values that are relevant along the process of managing data, information and knowledge.

A short disclaimer is necessary at this point: In this chapter I have only talked about knowledge that we gain through the use of machines. Thereby I have understated that a great extent of knowledge is actually gained without the help of machines and should so in the future. Maslow once wrote: “Science is only one means of access to knowledge of natural, social and psychological reality. The artist, the philosopher, the literary humanist, or for that matter, the ditch digger, can also be the discoverer of truth, and should be encouraged as much as the scientist.” (p. 8, ([Maslow 1970](#))).

Exercises:

- Reflect on the scenarios in chapter 3 and retrieve those scenes where the ethical use of knowledge is at stake. Align these incidences with figure x.
- Discuss the different uses and timely definitions of the term “knowledge” and argue from a utilitarian perspective whether it is beneficial for society to adapt an

information processing perspective on knowledge

- Is it ethically correct to work on illegitimately collected personal data?
- An ethical question that arises is whether software protection from data collection through agents could not create a digital divide between those people who can afford privacy protection and those who cannot. In the scenarios I describe how "rich" people can forgo economic incentives and stay anonymous while others who cannot afford this, need to reveal their personal data. Economic benefits are traded for personal data. Debate in class whether it is ethically ok to trade privacy for economic benefits.
- Do you think that there should be a right to be forgotten? Debate in Class...
- Think of a contemporary IT application and outline whether, how and to what extent it respects the principle of contextual of integrity for the use of its customer data.

Exercises Later Chapters

- Having a say in who creates knowledge about us seems a vital ethical claim. Please model a UML sequence diagram that outlines how such a 'human-in-the-loop' process could be organized, specifying technical entities required to make it work. What will be the challenges? What will be legal and technical requirements?

Freedom and Liberty in the Machine Age

Freedom is one of the most cherished values in today's democratic societies. It is embedded in most constitutions and international conventions. Human possibilities increase when machines give us greater autonomy in the way we live. For example, we can now work remotely from almost any geographic location and delegate many tedious or time-consuming activities to machines. As a result, we can spend more time on activities we enjoy doing. However, the control we gain is offset by the control we delegate to machines ([Brey 2004](#); [Spiekermann 2008](#)). When machines exercise control on our behalf, they can also infringe on our freedoms. For example, machines may force us into behaviors, deny us access to locations, tell us how and where to drive, etc. We must therefore carefully design machines with the goal to strike a fine balance between delegated tasks and tasks kept by humans.

Before delving into how to achieve this balance, I first want to clarify what freedom liberty and autonomy actually are as constructs. Do we need to distinguish between the terms 'freedom' and 'liberty' for instance? Pitkin (1988) makes a distinction between freedom and liberty due to the different etymological heritages of the two words. "Free" comes from the Indo-European adjective 'priyos,' which means something like 'one's own,' 'dear' or 'the personal,' with a connotation of affection or closeness. So freedom has been associated more with an inner state (similar to 'positive liberty' introduced below). An abuse of freedom means a threat "to engulf the self, to release to uncontrollable and dangerous forces" (p. 543 Pitkin). In contrast, "liberty" stems from the Indo-European verbal root 'leudh,' which means 'to grow' or 'to develop' in the face of control. It has been more associated with external states, like a political system in the face of which someone is permitted to grow (similar to concept of 'negative liberty' explained below). Despite these different linguistic roots most languages (surprisingly) don't separate the two terms. For example, Germans use the word "Freiheit" (similar to freedom) while French use the word "liberté" (similar to liberty) to mean the same thing. Most political and social philosophers have also used the two terms interchangeably ([Carter 2012](#)). Hereafter, I will therefore not discern them.

David Hume (1711 – 1776) defined liberty (or freedom) as the "power of acting or of not acting, according to the determination of the will; this is, if we choose to remain at rest, we may; if we choose to move, we also may ([Hume 1748](#))." But Hume recognized that choice alone does not lead to concrete action. After we make a choice, we need to be able to carry out what we choose. Our inner and outer environment sometimes impedes the execution of our choice. Therefore, we must distinguish two constructs: Freedom of will and freedom of action.

Closely related to this distinction between freedom of action and freedom of will are the concepts of negative and positive liberty. Negative liberty looks at external obstacles and is therefore close to the concept of freedom of action. In contrast, positive libertarians recognize that not all constraints on freedom come from external sources. Instead, positive libertarians emphasize internal constraints, such as irrational desires, addictions, fear, ignorance, and so on. They argue that to be free means to be independent from too much external influence. Since external barriers to action are the most prominent form of infringement I will give an introduction to negative liberty first.

Negative Liberty and Machines

Negative liberty is defined as "the absence of obstacles, barriers or constraints...(meaning) to

be unprevented from doing whatever one might desire to do” (Carter 2012). Some authors have called this perspective on liberty the “republican” view” (Pettit 1979). According to this view, liberties include freedom of movement, freedom of religion or freedom of speech. (Berlin 1969) described negative liberty as a kind of free space in which people are sovereign: “What is the area within which the subject – a person or a group of persons – is or should be left to do or be what he is able to do or be, without interference by other persons?” (p.2).

Technology may be designed as an external obstacle to our freedom of action. In chapter x, I described several ways in the scenarios in which this can be done. A very subtle form of interference with our freedom is when objects act autonomously and make decisions for us without us (their owners) being in the loop:

“As he turns around in his bed, he knows that his bracelet has now signaled to the coffee machine to prepare his morning café latte – a friendly nudge to get up and get going. But Stern doesn’t feel like it at all today...Stern slowly walks up to the kitchen. His café latte is not as hot as he likes it, and the machine has not put as much caffeine as usual into his cup due to his raised emotional arousal. But never mind.”

This example is one for *direct negative liberty infringement* or “technology paternalism” (Spiekermann et al. 2005). *Technology paternalism* involves autonomous actions of machines that interfere with users’ liberty and cannot be overruled by users. We already confront technology paternalism regularly, such as when we start driving a car without wearing a seat belt. Most cars autonomously start to beep and force us to wear the seat belt. An even stronger form of paternalism occurs when whole infrastructures and processes make us incur extra time cost and money. An example is the Halloville mall in the retail scenario:

“...people who are on a truly anonymous scheme can use a separate smaller mall entrance on the east side of Halloville that does not track any data....However, when people go through that entrance, they don’t receive the 3% discount he gets and have to pay for parking, a luxury that Roger cannot afford.”

Another form of negative liberty infringement is of an *indirect* nature. The infringement is indirect because the controlling machine entity or entities interfere with our activities without revealing their identity or the identities of their operators. Take the education scenario, where big data analysis is done by an unspecified entity, which then determines that Jeremy is not allowed to join Stanford’s online university program:

“I guess they think I can’t do any better because of my outdoor times.”...“What do you mean by outdoor times?” Roger asks...Jeremy doesn’t know whether it’s true, but a whistleblower software agent told him that big data analytics found that individuals’ intelligence were highly correlated with their average time outdoors over the past ten years. Since Jeremy had stayed indoors a lot when he played the Star Game VR, his average outdoor 10-year rating was probably pretty low. And he now suspected that this data was being used to predict applicant performance.”

We sometimes confront a similar situation today, as when credit scores are used against us and impede us from getting a loan or an attractive interest rate. Some people argue that it is appropriate to use technology in this way, and therefore the technology does not infringe on freedom. They would probably say that it was Jeremy’s choice to not exercise and spend more time outdoors. They would point out that if he had known about the outdoor expectations of Stanford Online and if those expectations were stable over time, he could have complied with them. They would also point to the fact that Jeremy is free to go to another school. Perhaps his desire to go to Stanford is simply too ambitious? Philosopher Ian Carter replies to this line

of argument as follows: “If being free meant being unprevented from realizing one’s desires, then one could, again paradoxically, reduce one’s unfreedom by coming to desire fewer things one is unfree to do. One could become free simply by contenting oneself with one’s situation. A perfectly contented slave is perfectly free to realize all of her desires” ([Carter 2012](#)).

Positive Liberty and Machines

“Positive liberty is the possibility of acting...in such a way as to take control of one’s life and realize one’s fundamental purposes” ([Carter 2012](#)). Positive freedom does not focus on the content of desires. Instead, positive freedom focuses on *the ways* in which desires are formed and whether they are the result of an individual’s reflection and choice or the result of pressure, manipulation or ignorance ([Christman 1991](#)).

There are four kinds of challenges to positive liberty in the machine age that I will describe: manipulation, addiction, denial of autonomy, and allocation of attention. I start with the danger of manipulation. Remember the retail scenario, where I describe the user dynamics of the Talos suit:

“The suit tracks all body functions and analyzes his moves and progress. Unlike some of his neighbors, Roger does not think that he will get paranoid about the suit. Many of his friends have gone crazy. The textiles transmit everyone’s activity data to a regional fitness database which displays everyone’s performance. So, many of his peers became preoccupied about their physical condition when seeing how they perform in comparison to their peers. They feel like they have to meet at least the average performance standard in the region, which is pretty high. One of his friends was so thrilled by the Talos force that he exhausted himself in a 12 hour run in the woods. He later had to be hospitalized for his exhaustion.”

Machine feedback – as the Talos case shows - may manipulate our “will” in good and bad ways. In the example, Roger’s neighbors become pressured to improve their fitness to a level that is not necessarily healthy. But they are also motivated to exercise more. No matter the positive or negative effects, we must note that machines (like the Talos) *do* influence our will. Some philosophers have claimed that the will is always free. Descartes famously wrote, “the will is by its nature so free that it can never be constrained” (Descartes, 1650, I, art. 41). But the majority of scholars agree that there are many situations where the will is not free; which is true for the digital world just as much as for the physical one. Physical, biological and social factors influence how we think and act. Machines do, however, amplify this influence. They amplify, because they constantly access our consciousness. They make comparative factors (such as peers’ performance) or behavioral rules more visible and nudge us to behave in a certain way.

Positive liberty is not only about the way in which we form our desires, but also about our active commitment to them. Harry Frankfurt (1982) distinguished between two kinds of desires: “First-order desires” are those we share with animals. We feel that we want something, for example, a cake. In contrast, “second-order desires” reflect on first-order desires. And here we may decide to not follow our first-order desires. For example, we may decide not to eat the cake because we don’t want to get fat. Frankfurt argues that we act freely when we are able to act on our second-order desires. These desires are the ones that we actually identify with and that actually satisfy us. The second-order desires reflect the true self ([Frankfurt 1971](#)).

Frankfurt’s distinction allows us to further distinguish free people from addicts. Addicts can only follow their first-order desires. They are not free because their ability to follow second-order desires is impaired. The scenarios mention the potential for addiction in machine environments:

On average, players spend 3 hours in the game per day, with 10% at 6 hours a day. And – good for Stern – the hours are growing. The game is really addictive, or, as Stern would put it, “*compelling*.”

In the chapter on health below, I will return to the problem of addiction to machines.

One challenge relating to positive liberty is how we can guard our *autonomy* vis-à-vis our intelligent machines. Human autonomy is regarded as one of the cornerstones of social enlightenment. Kant’s classical definition of autonomy is that it is “the property that the will has of being a law to itself” ([Kant 1795](#)). In Kant’s perception, autonomy means that we are sovereigns of our own will. Autonomy is hence a form of positive liberty. In the future, it may be difficult to maintain this sovereignty, which I hint at when I describe the intimate relationship between Sophia and her agent Arthur:

“Sophia chats with her 3D software dragon Arthur, who gives her advice on what products and shops to avoid for bad quality and where to find stuff she likes and needs. Sophia almost can’t live without Arthur’s judgment anymore. She really loves him even though he recently started to criticize her sometimes; for example, when she was lazy or unfair to a friend.”

Arthur is described as a highly intelligent machine being that influences Sophia’s thinking and behavior to an extent where she “almost can’t live without Arthur’s judgment any more.” So, agent Arthur is crucial to Sophia’s autonomy and positive liberty. Note Berlin’s definition of positive liberty: “What, or who, is the source of control or interference that can determine someone to do, or be, this rather than that?” ([Berlin 1969](#)), p.2). Berlin and Kant’s reflections highlight the special care we must take as to the sources of our future decision-making.

Finally, a subject that has not been associated with positive liberty infringement yet is the effect machines have on our attention allocation. Already, our IT systems channel a lot of our attention, which in turn controls what we see and do at a given time. The systems force us to attend to things other than what we have actually chosen to do at that moment. For example, machines capture our attention when we receive a phone call or message while in deep conversation with a friend. As a result, we are often not in control over what to look at and attend to. Of course, some people would argue that we don’t have to pick up a ringing phone, and we don’t need to look at the ads we receive. This argument is, however, true only to the extent that the incoming signals can be ignored. Depending on the design of interruptions, we sometimes can ignore them. We can set for instance our handsets to be silent and we can determine that they do not ring at certain hours. Often however we cannot ignore the machines surrounding us. This is the case for instance when incoming messages are highly salient, moving or pop up in front of us hindering ongoing work. I will detail below how systems can be better designed to be less intrusive so that they don’t infringe on our attention priorities and hence our positive liberty.

Figure x summarizes the various machine characteristics that this chapter covers as issues in our struggle to maintain positive and negative liberty in the machine age.

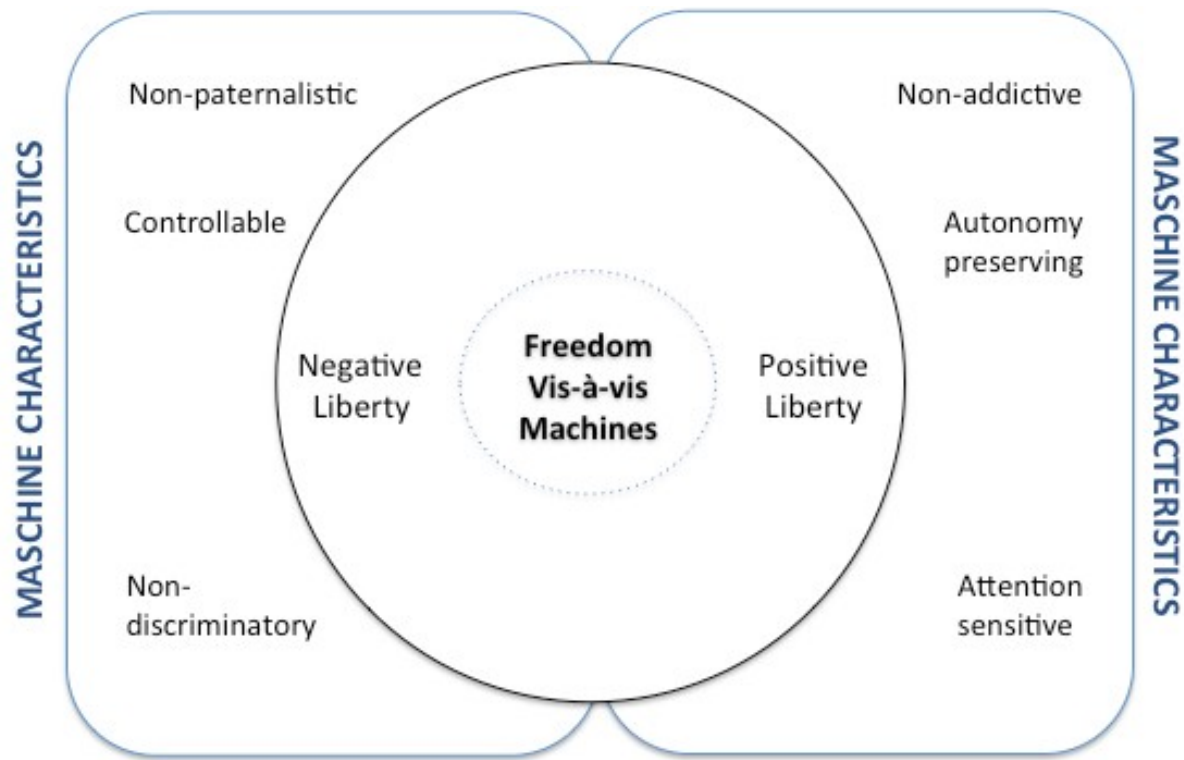


Figure x: Machine characteristics protecting our freedom

Technology Paternalism and Controllability

In his famous article on the computer of the 21st century, Marc Weiser wrote, “The [social] problem [associated with Ubiquitous Computing], while often couched in terms of privacy, is really one of control” ([Weiser 1991](#)). Weiser was working to embed computer power into objects. He thought that we will weave digital functionality into most of our ordinary objects and that clumsy computers would be replaced by machines that exercise power invisibly and through our objects. Some people have started to call this vision “The Internet of Things”, Stern’s bracelet is an example of the new machine. Because it knows when Stern usually wakes up and recognizes that Stern is moving in his bed, it prepares Stern’s morning coffee. It also knows, based on Stern’s pulse and skin conductivity, that he is relatively stressed. As a result, the bracelet signals the coffee machine to not only prepare regular coffee, but to reduce the caffeine level so that nutrition is optimized for Stern’s body state. Much of this scenario addresses paternalism.

“[Paternalism is] a system under which an authority... regulates the conduct of those under its control in matters affecting them as individuals as well as in their relations to the authority and to each other” (Merriam-Webster’s Collegiate Dictionary, 2003). The goal is “protecting people and satisfying their needs, but without allowing them any freedom or responsibility” (The Longmans Dictionary of Contemporary English 1987). Together with Frank Pallas, I extensively discussed and developed the concept of *Technology Paternalism* in an earlier publication ([Spiekermann et al. 2005](#)). Against the background of RFID and sensor technology, and with the help of several focus groups, we identified the main conditions under which we can talk about paternalist machines and what should be done to avoid them.

The first trait of a paternalist machine is that it starts acting autonomously. It is independent

and out of the control of the machine owner. Because the machine is not controlled, it cannot be overruled. Stern cannot stop his bracelet from ordering coffee even though he knows that morning that he needs more time in bed. The result of the action cannot be disregarded: the coffee is there. And this activity might limit or infringe on freedom. Stern would have liked to have his coffee later and with the usual caffeine level, but the machine does not react to that desire or even give him an option. Instead, the lower caffeine level is legitimized by the argument that Stern’s body is better off with less caffeine. Figure x summarizes the traits of paternalistic systems.

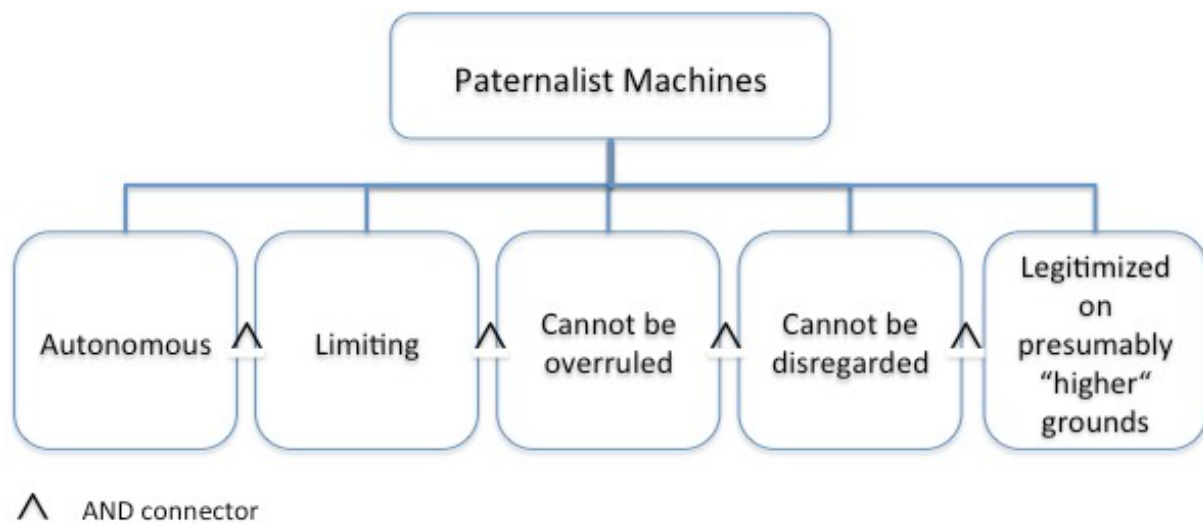


Figure x: Traits of paternalistic machines

How can we avoid such paternalist machines? A major result from our empirical research showed that owners should be able to overrule machines. “Decisions made by technology and any exceptions from this should be considered very carefully. People should always have the last word!” said one of the study participants ([Spiekermann et al. 2005](#)), p. 9). In Stern’s scenario, Stern could have the last word if his bracelet only signaled that it was ready to order coffee. Stern could confirm the order before it is placed. In a more sophisticated version of the bracelet, Stern may possess an agent Arthur as well. Arthur could ask him whether he wants his coffee prepared and whether he wants it with less caffeine. This version of the future sounds much more promising. The example shows that simple system design elements that determine how control is allocated between men and machines can alter the whole relationship. A freedom-depriving scenario, where coffee is prepared without request in an undesired way, is turned into a freedom-enhancing one where coffee is prepared for us by a machine exactly when and how we want it.

Optimal control allocation (often called “function allocation”) is at the core of a scientific field investigating “human-centered automation” ([Billings 1991](#)). In this field, a traditional approach to understanding function allocation between humans and machines was Fitt’s MABA-MABA list ([Fitts 1951](#)). Fitt suggested allocating tasks to humans and machines in accordance with their relative strengths and weaknesses.

Fitt’s list built on humans being better than machines in:

- detecting small amounts of visual, auditory or chemical energy,
- perceiving patterns of light or sound,
- improvising and using flexible procedures,
- reliably storing information for very long periods of time and recalling appropriate parts,
- reasoning inductively and at exercising judgment.

In contrast, machines were found to be better at:

- responding quickly to control signals,
- applying great force smoothly and precisely,
- storing information for some time and erasing it completely,
- reasoning deductively.

Even though these relative strengths of men and machines were formulated over 50 years ago, they still hold some truth. Machines today can detect and perceive quite a lot of obvious, bold and repeated patterns. Truly 'understanding' a situation though beyond what's encodeable is an exclusive human skill. While reliable storage of information is often identified as a clear machine advantage, digital storage is apt to degradation (depending on the storage medium). An this decay is unfortunately independent of information relevance. Human beings can often remember at least the most important parts of history. When they are trained they can well remember details.

That said, the relative skills of men and machines are rapidly evolving. Machine capabilities progress extremely rapidly, while human capability needs long training and practice. Consequently, no clear and long-term guidelines on how functions should generally be shared between humans and machines are available. In contrast, Sheridan expressed any attempt to develop such guidelines as "alchemy" ([Sheridan 2000](#)).

What is clear, however, is that we must carefully consider our options for allocating function control. Sheridan himself identified eight levels of relative control between humans and machines ([Sheridan 2002](#); [Sheridan 1988](#)). These levels range from one extreme, where a computer does everything and people have no control, to the opposite extreme, where individuals do not use machines (full human control). Table x summarizes this control manipulation scale and demonstrates how it could be applied to the scenario of Stern's bracelet interacting with the coffee machine.

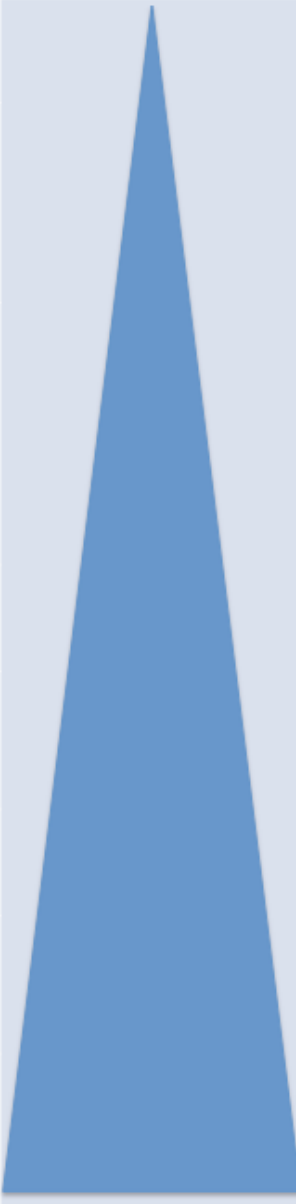
8 -10 Stages	Automation and Control Allocation between Machine and Human	Example: Smart Bracelet (SB) and Coffee Machine (CM)	Degree of Paternalism
1	M does not offer assistance; H must do the task completely herself.	Stern gets up and brews coffee himself. SB and CM do nothing (status quo).	
2	Upon request, M shows all alternative options to do a task. H executes.	Stern pulls up CM menu (i.e. on his mobile device or bracelet display) or starts Agent. CM menu or agent state <i>all</i> order options (time, caffeine level, etc.). Stern manually choses one <i>and</i> presses order button.	
3	M recommends specific way to do the task. H has to execute or not execute recommendation.	Stern pulls up CM menu (i.e. on his mobile device or bracelet display)/starts Agent. CM menu or agent state all order options (time, caffeine level, etc.) and <i>recommends one</i> in particular. Stern chooses. His choice is <i>automatically interpreted as an order</i> and is executed by CM.	
4	M recommends a specific way to do a task and it executes upon H's approval.	SB or Agent signals an option to order coffee. When Stern approves, CM executes.	
5	M recommends a specific way. Allows H a restricted time to veto before automatic execution.	SB or Agent signals an option to order coffee and recommend less caffeine. If Stern does not veto within a certain time frame CM brews coffee automatically.	
6	M executes automatically and informs H about action taken.	CM brews coffee automatically when receiving signal from SB. SB/Agent informs Stern that coffee is ready.	
7	M executes automatically and informs H only if asked to.	CM brews coffee automatically when receiving signal from SB. Stern can consult SB/Agent whether coffee is ready.	
8	M selects the method and executes task. H is out of the loop.	CM brews coffee automatically when receiving signal from SB. Stern has no information. Goes to the kitchen to get the coffee that was chosen for him. (scenario status)	

Figure x: Levels of output automation as distinguished by ([Sheridan 2002](#); [Sheridan 1988](#))

Although engineers benefit from knowing the different options for fine-grained control, they must base their decision on the ultimate goal of a machine service. Is it efficiency and productivity? Or are human freedom, dignity, growth and emotional well-being more important? In classical automation environments, where productivity, efficiency and safety have been the main design priorities, control has been delegated to machines. For some reason, the notion that more automation is always better persists. But in the coming machine age, where man and machine will interact ubiquitously in daily life, efficiency and productivity may not necessarily be the best strategy. The idea of full automation – some authors talk about fully “autonomous agents” – will probably need to make

room for a more balanced man-machine vision. This is because, as Norbert Wiener once said, “communication and control belong to the essence of man’s inner life” (([Wiener 1954](#)), p. 18). Humans’ emotions and behavior are strongly determined by the degree of control they have over their environments. In the 1970s, ([Mehrabian et al. 1974](#)) found that perceived control (“dominance”) over an environment lead people to approach that environment. In contrast, when people are deprived of control, they avoid environments, show reactance ([Brehm 1966](#)), feel helpless ([Abramson et al. 1978](#); [Seligman 1975](#)), are unhappy ([Thompson et al. 1991](#)) and even die earlier ([Langer et al. 1976](#)). It is therefore not surprising that early studies on control perceptions and Internet use found that people are more motivated to use e-commerce sites over which they have control ([Novak et al. 2000](#)). Control allocation in favor of the machine is therefore a less obvious decision in consumer markets than it has been in industry environments. A simple illustration of how ‘the pleasure of control’ can play out in tech-driven consumer markets is the continued use of stick-shift cars in Europe, where over 70% of cars still have manual transmission.

When engineers decide for more automation, they must consider how to provide user feedback. Human Computer Interaction (HCI) scholars outline how systems’ *feedback* is used to foster perceptions of control (see appendix x). In industrial environments that are highly automated, people have been observed to often don’t know what is going on. For instance, pilots in cockpits most frequently ask questions such as “what is it doing?”, “why is it doing that?”, “what will it do next?” and “how did it ever get into that mode?” ([Woods 1996](#)). Similar problems now arise in everyday life. People regularly “feel stress due to subjectively unpredictable behavior of technical systems” (([Hilty et al. 2004](#)), p. 863). For example, modern cars sometimes brake autonomously, even when on a motorway. People become stressed in such cases due to a lack of *situational awareness* ([Endsley 1996](#)). They do not know the mode a machine is in. Often they don’t know or forget about underlying machine mechanisms at work ([Endsley 1996](#)). Poor system design contributes to a lack of situation awareness. Too often a discrepancy can be observed between an engineer’s “conceptual model,” which determines how a machine acts, and the “user’s mental model,” which determines how humans understand this action ([Norman 1988](#); [Scerbo 1996](#)). Work by Donald Norman and others on control affordances in “The Design of Future Things” is therefore vital ([Norman 2007](#)). Intelligent machines should meaningfully interact with users, provide reasons for suggestions, allow users to easily ‘pause’ and ‘resume’ activity, only gradually advance to take over decisions for humans and respect that people have very different predispositions for how much control they want to delegate ([Maes et al. 1997](#)).

Coming back to Stern’s coffee example: If we used stage 4 of automation as shown in table x, we would see that Stern’s bracelet gives him the option to order coffee. Only when Stern approves this order does the coffee machine actually start brewing. If we implement the machine in this way, we don’t have a problem of liberty infringement or technology paternalism, and Stern is aware of what is going on. But let’s think of a case that is more ethically ambiguous. Imagine that Stern had a heart

attack in the past. He loves coffee, but it is not good for him. He always fails to comply with his own wish to drink less caffeine. His second-order desire is hence constantly undermined. Also, his health insurance company does not want him to drink coffee because drinking coffee increases his risk of another heart attack. The coffee machine is therefore set to not put caffeine in the brewer by default, a rule that Stern cannot override. To ethically judge this machine design, we now need to consider another construct related to positive liberty; that is how humans’ *autonomy* can be guarded (or lost) in the face of machines.

Autonomy vis-à-vis Machines

“Autonomy ... refers to the capacity to be one’s own person, to live one’s life according to reasons and motives that are taken as one’s own and not the product of manipulative or distorting external forces. Autonomy might be defined as the freedom to make self-regarding choices, in which a person expresses his/her authentic self” (([Koopmans et al. 2011](#)), p. 177). Taking this definition of autonomy and seeing the concept’s close link to positive libertarian thinking, what is key to Stern’s case is to consider *who* decided to put less caffeine in the coffee. To speak with Berlin’s words (1956, p. x): “What, or who, is the source of control or interference...?” Consider these potential sources of control:

- 1) Coffee machines in the future will generally not put caffeine in coffee at all. Some political entity has decided that the health risks are too high for everyone, and so by default all coffee machines comply with the zero-caffeine rule.
- 2) The manufacturer of the coffee machine has decided that the company wants to compete based on healthy coffee and therefore markets most of its machines with the zero-caffeine default.
- 3) The health insurance company has asked Stern to allow it to monitor his caffeine consumption by obtaining the coffee machine’s usage data and will deny him insurance for another heart attack if he drinks any caffeine.
- 4) The machine is flexible, and Stern can overrule it and brew his coffee however he wants it.

With options 1 and 3, Stern effectively loses his autonomy: he is not in control. Instead, the regulator (option 1) or the insurance company (option 3) has taken over and infringed on his liberty. To quote Kant, Stern is “constrained by another’s choice” (([Kant 1795](#)), p.x).

With option 2, Stern has a bit more autonomy as long as there is competition in the market for coffee machines. He can still purchase from a vendor who gives him more freedom. Finally, option 4 respects Stern’s freedom. He can choose each day how much caffeine he wants to drink. Here, theoretically, Stern is autonomous. Yet, if he has this freedom, then he is tempted each day to have a little real coffee and so to forgo his

second-order desire to remain healthy and avoid another heart attack. So we can consider a fifth design option, one that optimizes both his freedom in terms of autonomy and his ability to realize his second order desire:

- 5) The vendor sets a zero-caffeine default in the coffee machine. Stern can easily override this default. But to do so, he incurs a small transaction cost, such as confirming his choice. The machine asks him: "Given your health status, do you really want so much caffeine?"

This option - setting machine defaults to protect individuals while leaving them meaningful room to make a different decision - has been called "nudging" ([Thaler et al. 2009](#)). Nudging has recently found strong resonance in political practices. Of course, "by choosing their actions, one by one, humans continually create and adjust their own ethical characters - and their own lives and personal identities as well" (([Bynum 2006](#)), p. 160). Nudging people interferes with their continued identity construction. Yet, from psychology and behavioral economics, we know that humans often have trouble making rational decisions ([Kahneman et al. 2000](#)). Many of us can't effectively judge short-term and long-term risks. We are often tempted by immediate gratification ([O'Donoghue et al. 2000](#)). We tend to hyperbolically discount long-term risks ([Laibson 1996](#); O'Donoghue et al. 2001). The last option, where the machine nudges a person to his true advantage, is an example of a machine helping us to follow our second-order desires while preserving our liberty to choose. This, of course, presumes that a machine or rather a machine's designer knows about our second-order desires. If such nudging is then not done too intrusively, but is transparent and can easily be countermanded, machines can support our positive liberty. Continuing in the line of positive libertarian thinking, it is crucial for a nudging machine to reveal the source of its defaults though: why was the default set, who set it and how can it be changed? This information should be an easily accessible information. At the very least it should be part of the machine's manual.

In the case of the coffee machine, giving this kind of information seems doable. But think of the autonomous Alpha1 robots or agent Arthur. These extremely advanced machines may be called "agents" because they display three key characteristics: interactivity, autonomy and adaptability. Box x details these machine characteristics. . It may be much harder to remain autonomous in the face of these systems.

Box x

Autonomous Agents: A Characterization

Some scholars have proposed that machines qualify as "agents" if they are interactive, autonomous and adaptive ([Allen et al. 2000](#); [Floridi et al. 2004](#)). These scholars define these three traits as follows:

Interactivity means that machine agents react to input from their environment. For example, they know where they are based on geo-coordinates. They can sense their environments, receive and interpret video streams, and use this data input to react. Interactivity may be realized with a simple "if <some external state> = x, then do a"

Machine *autonomy* means that the system can trigger an action "on its own." It does not necessarily need an external stimulus or command. It can perform internal transitions to change its state. For example, a machine can contain an internal clock that measures its life-time. After two years, the internal clock tells the machine to stop functioning. Autonomy may be realized with a simple "if <some internal state> = a, then do b"

Adaptability means that a machine seems to learn. Through its interactions, it can change the rules by which it changes state. Thus the machine takes decisions based on factors the combination and sequence of which cannot be perfectly predicted even by a machine's designer. The life-time algorithm of it depends on too many internal and external states. Take the example of a house cooling system, which may start with an initial rule to balance temperature at 18° C (64° F) based on inhabitants' sensed body heat. When three out of four people in the house fall below the recommended body temperature, the system increases the house temperature to 20° C (68° F). Later, the machine might detect that body temperature is a little too high for some of the house's inhabitants. Instead of decreasing the temperature to 19° C (66° F), it adds perspiration as an additional indicator for bodily health to its 'optimal-temperature-setting' algorithm; it hence effectively changes (extends) its initial rule. The system might later determine that maintaining a temperature of 20° C (68° F) is best because this temperature guarantees a good balance between body heat and transpiration. Alternatively, by retrieving the newest research from the Web, the house temperature system may learn that for people beyond 80 years of age, the optimal room temperature is 23° C. It recognizes that one lady in the house has just turned 80 and it therefore sets the room that she is in to 23°.

A key challenge for *human autonomy* in the face of agents will be to have the agents correctly model users and their decision environments in real-time. Vendors will not always be able to set simple defaults correctly. In scenarios where humans interact with agents like Arthur, adaptive machines (see box x) regularly make decisions based on some internal or external stimuli and states and more or less comprehensive reasoning. The question is how these future machines will learn from us, "adapting" to their users and owners. How will we communicate our desires to our machine agents, and how strongly will our agents' algorithms then respect these desires when they make decisions for us? As Batya Friedman and Helen Nissenbaum outline in their paper on "Software Agents and User Autonomy," "A lack of technical capability on the part of the software agent - to be able to accurately represent the user's intentions - can lead to a loss of autonomy for the user" (([Friedman et al. 1997](#)), p. 467). The same is certainly true when it comes to the capturing of second-order desires that precede intentions.

When machine agents don't possess the right level of capability and make the wrong decisions for us, they undermine our autonomy and may become a nuisance for us as users (as some machines are already today). Take the cooling system example: What if the old lady in the house has always loved to sleep in a cold room below 17° C (62°F)? A way to support machines' capability is to design them in a fluid way. *Machine fluidity* → means that agents can adjust to changes in users' goals, even smaller subgoals ([Friedman et al. 1997](#)). Returning to the coffee machine example: if Stern's health improves, he may be able to have real coffee from time to time without negative health effects. The coffee machine needs to be able to accommodate this change of Stern's goals or spontaneous deviation from his default. Machines are more fluid when users can influence a machine's inner workings. The fluidity implies that users should be able to access machines' settings and rules. They should be able to alter them and have insight into and choice over the behavioral options available. This requirement seems like an irrational call from the past. As Floridi and Sanders note, “The user of contemporary software is explicitly barred from interrogating the code in nearly all cases” ([Floridi et al. 2004](#)). Does this need to be the case? And is this even desirable? Some people would argue that people are lazy and that they are not interested in manipulating anything nor are they qualified. They can do more harm than good. On the other hand, the history of technology is paved with evidence that people do manipulate their machines. Take the example of the automobile. People not only became innovative around machines thinking about the most eccentric uses (figure xa), but many spent parts of their lives understanding them, repairing them and ramping them up (figure xb). A whole youth culture has evolved around the manipulation of computers, testing out their limits and capabilities.

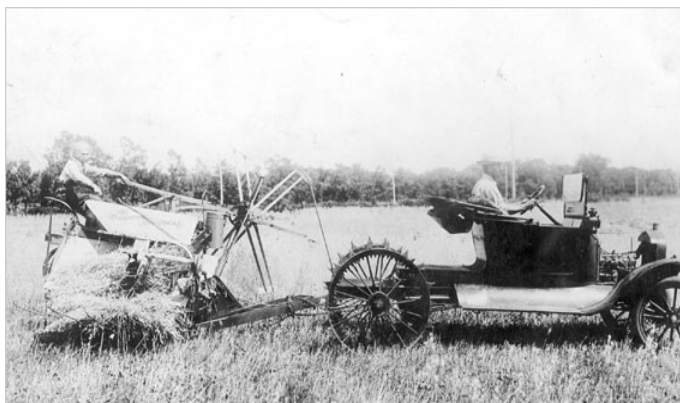


Figure x a: People putting their machines to new uses



Figure x b: People manipulating their machines

This last point of *machine accessibility* → is one that I will take up again in more detail in section x, where I discuss the benefits of free and open source software. For people to not lose freedom and make decisions in co-operation with their machines, they need to be able to control the code of their machines. Some scholars will firmly contest this claim. We don't know all of the details about how a car works, but we can still use it and feel

very free as we speed up and down the highway. Furthermore, many machines now come with "software as a service" architectures. Economic reality is that the code is black-boxed on some remote server that is not accessible to users. An open question for machine's future design is therefore *how much* insight we really need *to feel* in control over the machines we use. And at what level should we then effectively be allowed to manipulate them? Only at the application layer? Or also at lower levels of a machine's design?

From a positive liberty perspective, it seems to be important for a machine owner to determine and change "the source" of recommendations coming from the machine. For example, Sophia should be able to determine whether Arthur's recommendations are based on the mall's advertisement system, the GX1 loyalty card system or some NGO.

"Sophia receives less information, but the information she receives is of higher quality and more tailored to her preferences. She can let Arthur know from what sources he should retrieve recommendations. She trusts Playing The World and believes that Arthur respects her orders, looks after her privacy and recommends what is best for her."

Even if the influence of a machine's sources and defaults remains unknown for most users, who are neither willing nor capable of digging into machine details, the potential freedom to access and change positively influences liberty at the societal level. As I will argue below – and many scholars have argued before me – open and free code is vital for freedom ↯.¹⁰

If individual users don't manipulate their machines at a deep level but want to access and control their agents to some degree, they should be able to do so on the application layer at least ↯. This means of course more work for desires who have to think about who to make their creations comprehensible. Users must be able to easily understand the application layer interface. The application layer interface should fully represent how machine states are determined and how they can be altered. Most importantly, users should be able to make changes easily and at minimum transaction cost. Some authors have noted critically: "In some instances, software agents may supply users with the necessary capability to realize their goals, but such realization in effect become impossible because of complexity. That is, the path exists to the state the user desires to reach, but negotiating that path is too difficult for the user" (([Friedman et al. 1997](#)), p. 467).

Against this background, it would make sense for system designers working on highly autonomous machines or agent systems to embrace progressions in the field of end-user programming and development ([Ko et](#)

10 Machine manufacturers will argue, of course, that deep tampering with machines is dangerous and creates liability issues if a machine fails. I would argue that tampering below the application layer could shift liability to the user. This liability is no different from the liability of people who customize or tinker with their cars: When those machines break, the driver is often responsible. Digitally enforced logging of code changes in future machines could ensure that liability goes where it belongs.

[al. 2011](#)). "People who are not professional developers can use EUD (End User Development) tools to create or modify software artifacts (descriptions of automated behavior) and complex data objects without significant knowledge of a programming language"¹¹ I have pointed to this kind of user control over future machines in the scenario when I describe how the 8-year-old child Sophia is able to access and manipulate her agent Arthur:

"One cool thing about the game is the fostering of players' creativity. Anyone can design his or her own game characters with an easy-to-use programming tool and have these characters accompany them or engage with other people and their characters. The pink dragon displayed on the meeting room's screen is actually the toy of an eight-year-old girl named Sophia, who has chosen to be accompanied by this dragon in the game. She calls it Arthur...[Sophia] has configured Arthur to run on top of her personal data vault. She could do so because of an agreement between the game company Playing The World and her personal data vault provider..."

In my sci-fi cases, Sophia is smart enough to manipulate her agent Arthur. If people are anxious about changing how their agent works, then at the very least machines should explain why they act as they do. Petty Maes, a software agent pioneer, wrote that "...the particular learning approach adopted allows the agent to give 'explanations' for its reasoning and behavior in a language the user is familiar with, namely in terms of past examples similar to the current situation. For example, 'I thought you might want to take this action because this situation is similar to this other situation we have experienced before, in which you also took this action' or 'because assistant Y to person Z also performs tasks that way, and you and Z seem to share work habits...'" (([Maes 1994](#)), pp. 32-33). When designers account for accessibility, they must match users' abilities to what the machine assumes the user is capable of ([Friedman et al. 1997](#)).

Finally, machine agents need to be *reliable* –I. If machine agents use inaccurate or false information, people can't trust them. And if people then still have to rely on the agents, they will naturally feel out of control. At first sight, this requirement sounds easy to meet – of course machine agents need to work with correct information! However, much machine judgment and feedback today is based on probabilities. The responses we receive from machines are based on what the machines think is correct. To date, machines don't reveal that their feedback is just a statistical probability and cannot be taken for granted. For example, timely ad networks often make projections of people's likely traits and interests. The networks do so based on observed and probabilistic behavioral patterns and demographics. In many cases, the resulting judgments are outdated or false as I also described above in the context of machine fluidity and an old lady preferring a cold room. Future machines cannot be based on such suboptimal user knowledge (see the section on truth). If we want to build autonomous machines that gain our trust and don't undermine our

¹¹ http://en.wikipedia.org/wiki/End-user_development (URL last visited on July 20th 2014)

autonomy, then users and agents must co-operate to ensure that agent behaviors are based on timely and true user data₁, that is available in real-time₂. We need excellent and reliable *user models*₃ (Kobsa 2007). Again, this data collection needs to happen with the consent of users, as was outlined in section x.

Figure x summarizes the factors that influence how we will perceive our autonomy vis-à-vis machines in interacting with them and the technical factors this perception depends on. The empty boxes in the figure indicate that other factors not covered here might influence our perception of autonomy.

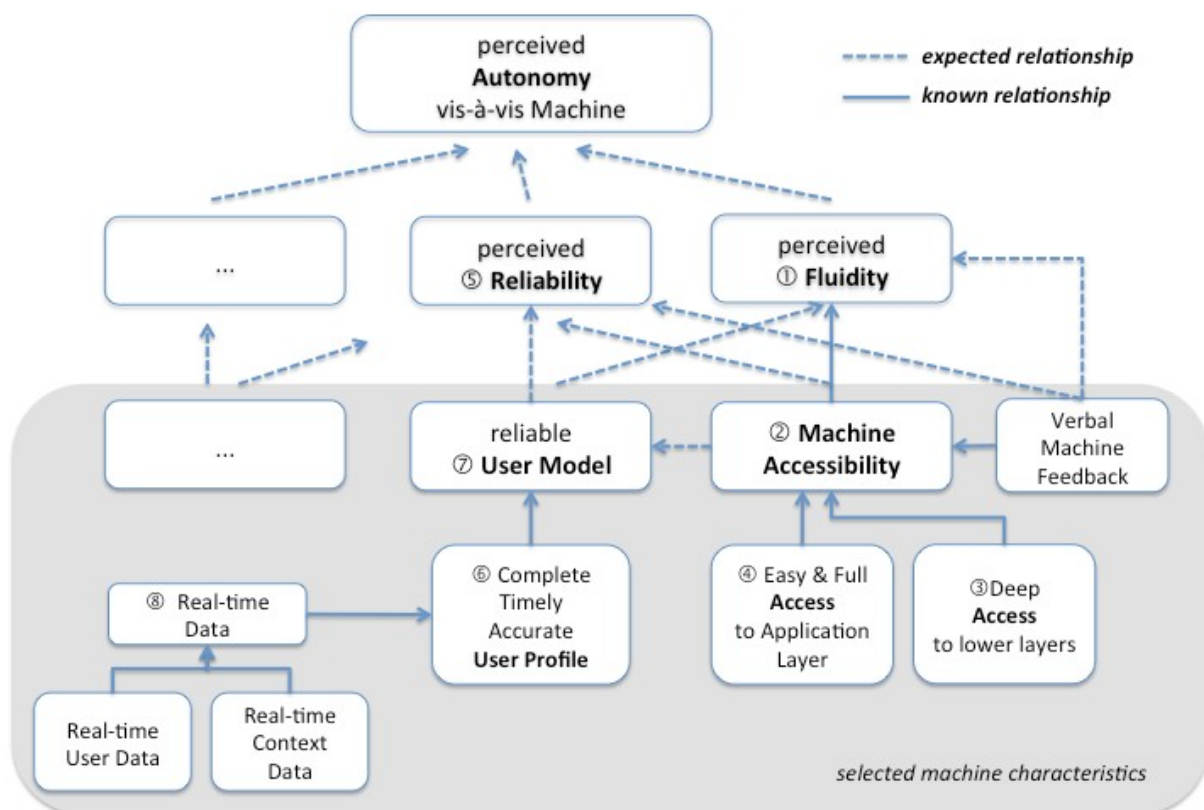


Figure x : Factors expected or known to influence a machine’s perceived autonomy

Attention Sensitive Machines

*"What information consumes is rather obvious:
it consumes the attention of its recipients.
Hence a wealth of information creates a poverty
of attention, and a need to allocate that attention
efficiently among the overabundance of information
sources that might consume it."
(Herbert Simon, 1971)*

Herbert Simon foresaw the current explosion of information: Every day, consumers are confronted with 2,500 to 5,000 advertising messages ([Langer 2009](#)). In addition, employees receive an average of 200 e-mail messages per day ([Fischer 2012](#); [Jackson et al. 2003](#); [Nuria et al. 2004](#)), leading a typical office employee to check his or her messages around 50 times a day ([Robinson 2010](#)). As a result, attention spans are shrinking. A typical office worker is interrupted every 4 to 12 minutes ([Dabbish et al. 2011](#)). As hardly any room is left for concentrated efforts are we still free masters of our attention? Or are we addicted to machines? Pushed by them into attending them?

"The HR manger chatted about Stern's recent attention scores. The company's attention management platform had found that Stern's attention span to his primary work tasks as a product manager was below average. 'You seem to be interrupting yourself too often,' the HR representative had said. 'But what could I do?' thought Stern. There are simply too many messages, e-mails, social network requests etc. that would draw on his attention. So he obviously did not match the 4-minutes minimum attention span that the company had set as a guideline for its employees. Employees' attention data was openly available to the HR department and management in order to deal with people's dwindling capability to concentrate."

Stern' story suggests that the problem of attention allocation may force companies to set minimum requirements for employees' concentration capacity and monitor employees for compliance. We might even see a new digital divide between those who can concentrate and focus on primary tasks and those who don't have the willpower to do so.

William James defined attention as "the taking possession by the mind in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought...It implies withdrawal from some things in order to deal effectively with others" ([James 1890](#)). James' definition of attention suggests a link between attention allocation and positive liberty. He talks about a "taking possession by the mind," but we must also consider who might take the mind over. Do we take possession of our own minds? Does Stern, who often consciously decides to interrupt himself and turn to another task in less than 4 minutes? Or is it the number of external entities that ping him constantly and so aggressively that he cannot avoid the intrusion? As of 2014, we know that in at least 51% of attention switches between knowledge work tasks, the external environment causes the interruption and makes people stop one task to turn to another. In 49% of the cases, people initiate the switch themselves ([González et al. 2004](#)).

Given this high number of external interruptions, we must build machines that are less

intrusive. Attention is our scarcest and most valuable human resource. What we attend to defines who we are. And if our daily work environment, where we spend over 50% of our waking time,¹² turns us into creatures whose attention is externally manipulated by machine signals, then we risk losing a considerable part of our autonomy.

Do machines need to interrupt us in such an intrusive way that we lose control over our attention allocation? Research in psychology, computer science and human computer interaction shows that they do not. If machines better understand interruption situations and are sensitive to our natural attention allocation habits, they can be much less intrusive.

Attention-sensitive Interruption

Interruption can be defined as an event where a stimulus effectively redirects an individual's attention away from an ongoing *primary task* and shifts that attention towards a secondary task. Examples of interruptions include an e-mail notification popping up at the side of a screen or an ad on the border of a website. McFarlane (2002) distinguishes between “negotiated timing” of interruptions and “immediate timing” ([McFarlane 2002](#)). In negotiated interruption timing, users have some control. They are notified of an incoming message, but they can determine whether and when to view it. For example, today's e-mails or IM messages may be announced by the appearance of a small window, but the messages stay in the window or inbox until the user chooses to view them. In contrast, immediate interruption timing gives users no control and deprives them of freedom. The system enforces immediate attention. Full-screen popups or screen freezing are examples of immediate interruptions, as are emergency messages in the car that freeze the radio show.

A third option is to not announce secondary tasks at all. Instead, users can “pull” information when they want to view it; an example is opening a mail program when we want to check what is there. The negotiated and immediate message delivery designs are both “push” strategies. Strictly speaking, they both reduce our liberty because they directly or indirectly force us to attend to incoming information. Our minds are not free to choose what to look at. Only interruption delivery strategies that employ the pull strategy give users full positive liberty to decide whether and when to retrieve information (see figure x). If we think the pull strategy to the end it means that only people's intention to do something, to search for something or to buy something drive their action. Doc Searls called this potential avenue technology and economics could take “The Intention Economy” ([Searls 2012](#)).

¹² US Bureau of Labour Statistics, Time Survey 2011; URL: <http://www.bls.gov/news.release/pdf/atus.pdf> (last accessed on July 22nd, 2014)

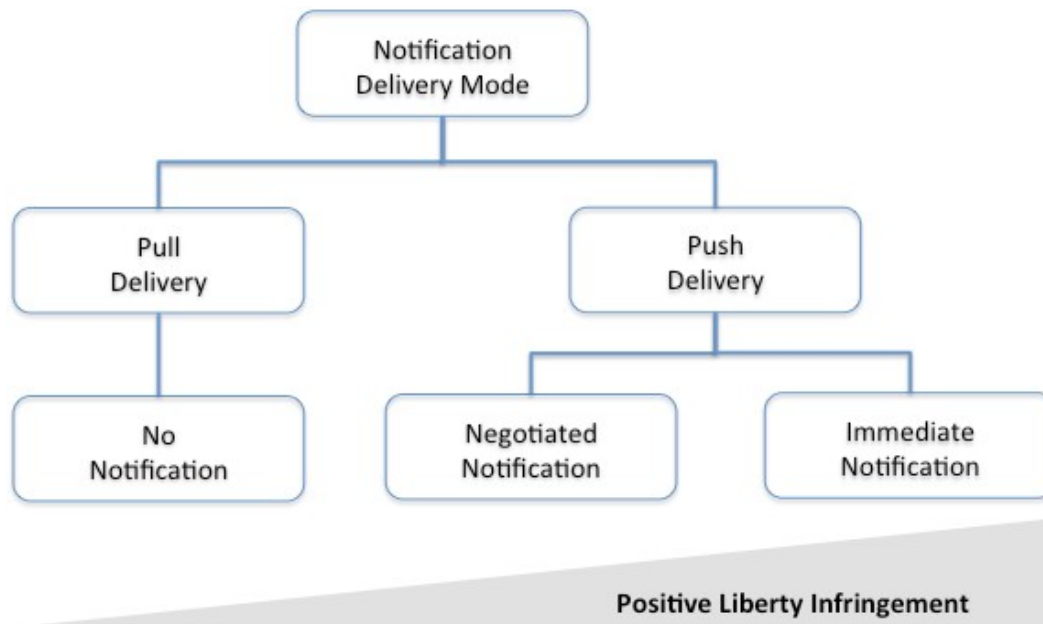


Figure x: Liberty Infringement can be limited by Message Pull Delivery

Consider a system design that involves *negotiated* notification and hence some liberty infringement. How can this notification be delivered so that users feel minimal intrusion on their freedom? Note that intrusion is also a privacy harm. Solove defined intrusions as “invasive acts that disturb one’s tranquility or solitude” (Solove 2006), p. 491). Warren and Brandeis, pioneers of privacy research, talked about “The right to be let alone” (Warren et al. 1890).

McFarlane’s (2002) study of interruption coordination techniques found that users who can negotiate their attention to an interruption normally take around 10 times longer to attend to it than when systems have control and send notifications immediately. This lengthy delay before attending to secondary tasks is due to *task chunking behavior*. Humans naturally wait to switch to another task until a current subtask is completed and mental workload is low enough to accommodate something new (Buxton 1986). An ideal moment for interrupting someone is therefore at breakpoints between tasks (Salvucci et al. 2010). That said, when we are immersed in evaluative or perceptual tasks, fewer breakpoints are available. This is the case, for example, when people read, learn or write. The perceptual and non-repetitive nature of knowledge work tasks makes them harder to interrupt in a nonintrusive way than executive tasks with multiple breakpoints (Brumby et al. 2009). Knowledge tasks are characterized by unfamiliar situations, objects, or texts that we need to learn about. Because we don’t have rules of control or “production rules” for them yet (Rasmussen 1983), we need our full cognitive capacity to complete them. While people are engaged in knowledge tasks, systems should therefore protect people’s attention resource and withhold notifications until the knowledge task is completed or people choose to stop working themselves. Only then should notifications be delivered.

Waiting for the end of a task or the next subtask boundary is, of course, not always possible. In emergency cases, an immediate notification may be justified to *ensure immediate* user attention and reaction. Some primary tasks, especially those that are skill based, may also be interrupted without waiting for breakpoints. Rasmussen distinguishes between **S**kill-, **R**ule-, and **K**nowledge-based tasks (Rasmussen 1983). Skill-based tasks, such as sensory-motor

skills, need little conscious control because they have become a kind of habit. These tasks don't require much cognitive capacity and hence can more easily integrate a secondary task or notifications. For example, many people can drive a car and talk on the phone. Although the ability to focus on driving might be impaired, we don't view the phone call as a liberty infringement or a privacy tort.

Primary task types and breakpoints are not the only factors available to minimize intrusion and maximize perceptions of control. The design and relevance of a notification are also important (see figure x). The *interruption design*, which includes colors, size, vividness, and motion, influence how strongly people's attention is captured by it ([Beattie et al. 1985](#); [Taylor et al. 1982](#)). In particular, movement in notifications deprives people of the ability to avoid them ([Bartram et al. 2003](#); [McCrickard et al. 2003](#)) because humans' innate “orientation reaction” forces them to react to unexpected motion ([Diao et al. 2004](#); [Pavlov 1927](#)). Another way to make a message pass is by choosing the right modality. Modality is the sensory channel used for information transmission: visual, tactile, or auditory. The modality of a notification can be identical to or different from the modality of the primary task. For example, a notification that a new e-mail message has arrived can be delivered in a visual modality (form) by using an on-screen pop-up window or in the auditory modality by using an alert sound. When the notification comes in the same modality as the primary task, the two interfere because they use the same perceptual resources ([Storch 1992](#); [Wickens 2002](#)). People are then maximally disturbed. Interruptions of the same mode as the primary task should therefore be avoided if possible.

Modality configuration has challenging consequences for voice-based human-agent interaction. If the same modality must be used, one way to optimize human-agent interaction is to have people systematically initiate the conversations with agents. In situations where this is not possible, agents should only interrupt users, when the incoming information is relevant for the user. We refer to relevant situations as those where ‘mutual task relevance’ is given. *Mutual task relevance* exists when the notification or secondary task contains information that is topically associated with the primary task or in the same domain. The primary and secondary tasks can also be mutually relevant in terms of goal utility. Goal utility is given when the secondary task is complementary to the primary task and supports completing the primary task. When notifications and the tasks announced to users are relevant, they are perceived as less intrusive. An example is the Google search engine that displays relevant ads to people corresponding to what they are looking for. Such ads that directly correspond to search terms are less intrusive than banner ads that display any kind of information people are not interested in.

Figure x summarizes the described parameters, which determine the positive or negative perception and level of disruptiveness of an interrupt.

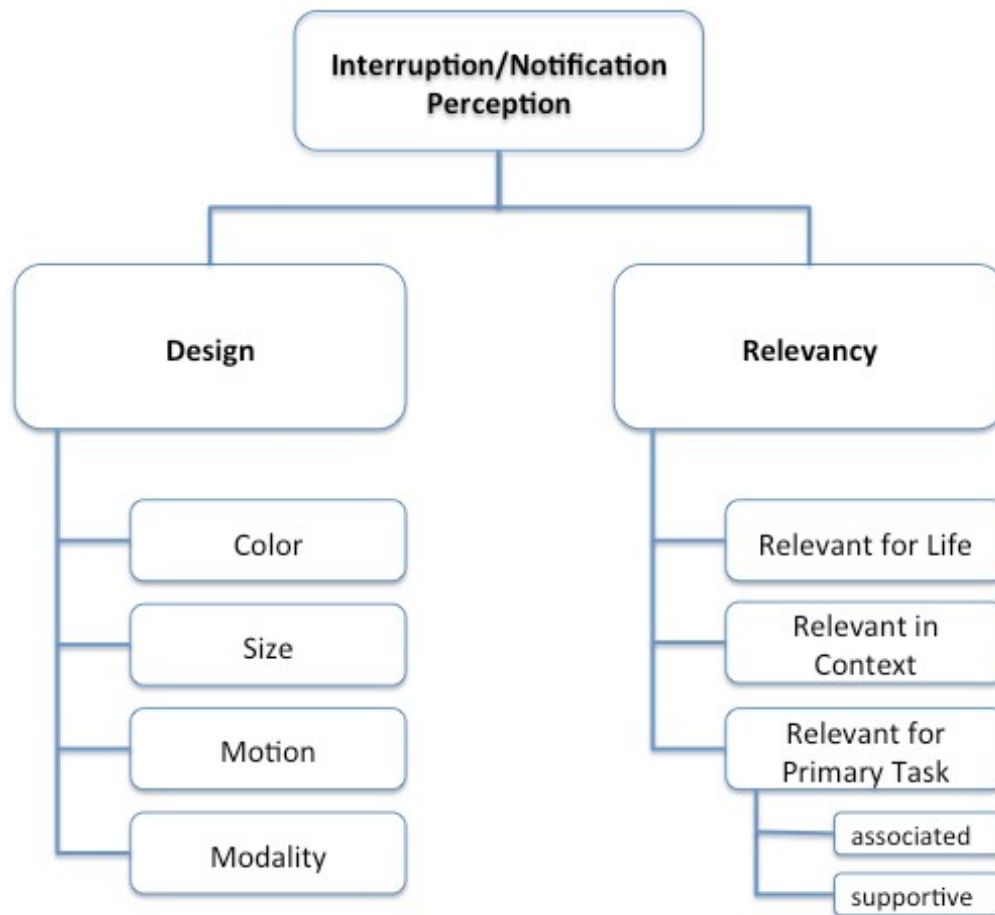


Figure x: Parameters determining the positive or negative perception of an interruption

Summing up: Freedom and Liberty in the Machine Age

In this chapter, I used the distinction between negative and positive liberty as a starting point to discuss several ways in which machines can undermine our freedom and how they can be built to not do so. In particular, I identified technology paternalism as a threat to our liberty. Designers can choose levels of automation to avoid *coercion* of our lives by machines. Instead, automation can be tweaked in a fine-grained manner so that people retain control while reaping the benefit of machines taking over tedious responsibilities. This has the potential to “free” people in many respects.

If intelligent, autonomous machine agents become common, we must consider where they receive their orders from – their owners or some external source? Can we control the source? Are the machines’ actions transparent? These questions will be vital if we want to remain autonomous as individuals. “The freedom of thought” is probably the major threshold of human liberty, and if our thought is increasingly manipulated by machines, we risk losing a great part of our human identity. We differentiate ourselves from animals through our capacity to direct our personal thoughts. If that capacity is lost, we risk reducing ourselves to Pavlovian dogs responding to our machines.

Taken together, ethical machine design needs to avoid coercion and manipulation of our activities and of our thinking. “Coercion and manipulation subject the will of one person to that of another,” writes Joseph Raz; in our context, “another” person might be the owner,

operator or engineer of a machine ([Raz 1996](#)). Joseph Raz’s work on autonomy can serve as a frame for the subject areas covered in this chapter. According to Raz, human autonomy has three conditions: (1) The mental ability to form intentions and plan their execution, (2) the availability of a range of options to choose actions and (3) the independence to form and execute our intentions. Our mental ability to form intentions and our independence to do so is influenced by the way our attention is channeled or affected by the constant interruptions we face as we think. Our options to choose and execute actions are determined by the degree of automation our machines embed. Figure x illustrates the relationship between our autonomy and machine traits.

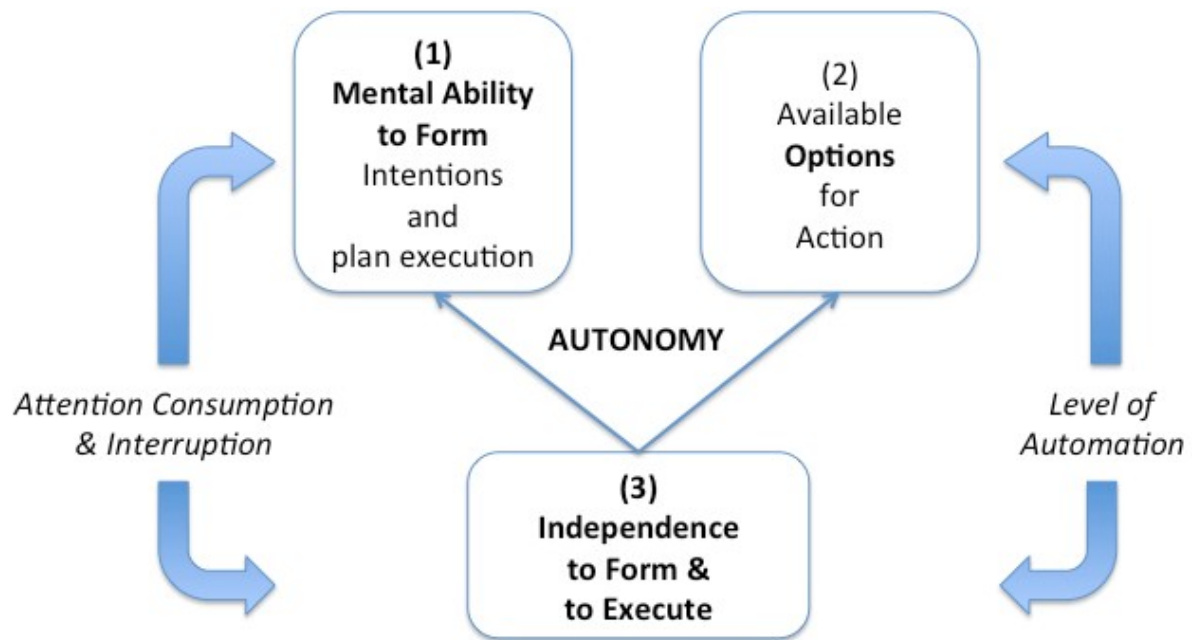


Figure: Joseph Raz’ three drivers of human autonomy

What I have not covered in this section is how our freedom can be affected by surveillance. Scholars regularly discuss how people may behave differently when they know they are being watched. George Orwell’s *1984* describes a dystopian society in which constant surveillance destroys people’s freedom ([Orwell 1949](#)). I will expand on this issue in the chapter on security, safety and trust. I also discussed ubiquitous data collection in section x, where I described the ethical principle of informed consent. Freedom and liberty can be guarded to some extent if people are in control of data collection.

Exercise:

- Take one of your current messaging applications and analyze how its message

*Univ. Prof. Dr. Sarah Spiekermann; "The Human Use of Machine Beings", Chapter 3,
Taylor and Francis, New York, 2015*

notifications could be designed to minimally distract users. How intrusive is the current interruption management system in the application you chose? How could attention management be further improved for this application?

Debate

- Do you think that companies should have an attention monitoring system?

Health and Strength in the Machine Age

Bodily health is important for many reasons. Health is at the core of humans’ physiological needs. If we are not healthy, all of the rest of Maslow’s needs are impacted. Herzberg would probably argue that health is a “hygiene factor” of motivation ([Herzberg 1968](#)), which means that if people are not healthy, any attempt to motivate them by appealing to higher needs or values will be in vain. Norbert Wiener regarded human physiology as core to a person’s information processing potential ([Wiener 1954](#)).

The World Health Organization (WHO) defines health as “a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity” (p. 100, [World Health Organization 1946](#)). Physical health is a state and perception of bodily well-being in which an individual can or feels that he or she can perform daily activities and duties without any problem. In contrast, mental health is a state of well-being in which the individual realizes his or her own abilities, can cope with the normal stresses of life, can work productively and fruitfully, and can contribute to his or her community ([World Health Organization 2001](#)).

The stories in chapter 3 illustrated that machines can strongly influence the health and strength of people. They do so directly and indirectly (see figure x), in a short-term and long-term manner. They affect our bodies, our minds and our social well-being. By “direct” influence, I mean that using machines has a causal relationship with our health and strength. “Indirect” influence includes phenomena that mediate or moderate the use of IT and its effect on health. For example, an addiction to online games may reduce the social ties of an individual, and that loss of social ties may negatively impact mental health. In the following I report on studies that have been conducted on “Internet” use. I therefore use the term “Internet” here as representative for various machine services, like online news services, social network platforms, gaming, etc.

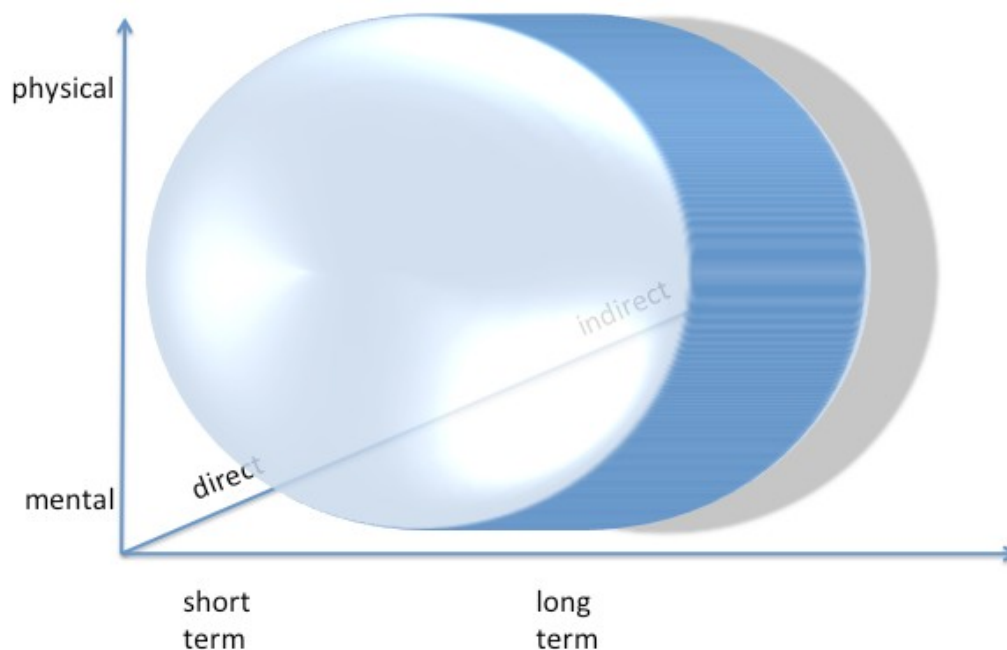


Figure x: Three-dimensional influence of Machines on Human Health and Strength

Machines’ direct impact on physical health and strength

The direct impact of machines on physical health has been studied in the field of ergonomics, which considers factors such as safety, comfort and performance in man-machine interaction. The Journal “Ergonomics” defines the field as follows: “Drawing upon human biology, psychology, engineering and design, ergonomics aims to develop and apply knowledge and techniques to optimize system performance, whilst protecting the health, safety and well-being of individuals involved.”¹³ Researchers have published a number of ISO norms on principles of ergonomics for fields in which humans and machines interact, including ISO 26 800 on the general approach, principles and concepts of ergonomics. When engineers build machines that can influence the physical health of individuals, they must begin by learning about the ergonomics standards for their field. Studies typically focus on how to apply ergonomics in specific areas like health care, navigation systems, office environments, aviation, etc. International organizations such as the Institute for Ergonomics and Human Factors¹⁴, the International Standards Organization (ISO) and relevant publications such as the Handbook of Human Factors and Ergonomics provide detailed guidance on how to design IT systems in the right way. The design principles identified in these sources can be used to run through the system assessment framework that is outlined in chapter x of this book. Ergonomics not only explores the physical fit between humans and machines but also looks into cognitive fit and emotional reactions to machines.

But machines cannot be designed only for an optimal *fit* with humans. Machines can also help to *enhance* our cognitive and physical capabilities. An early example is designing IT for universal usability, giving blind and deaf people access to knowledge ([Shneiderman 2000](#)). Furthermore, a whole engineering field for spare body parts is rapidly evolving. Recently, researchers developed a 3D-printer for organic body parts that are customized for a particular person ([IEEE 2014](#)). Such body parts and extensions can not only serve health purposes, but also be used to extend human capabilities and strength. For example, augmented reality add-ons can be embedded in contact lenses to enhance a subject’s vision ([Parviz 2009](#)). Where such sensors are not directly inserted into the biological system, they can be given to us in the form of physical tools such as digital glasses or electronic textiles. As described in the gaming scenario, such tools enable us to see the natural environment with an extra layer of information, for example, as an infrared or heat-map overlay. Artificial body parts and smart textiles may transcend their typical market of the old and handicapped to be used by the general population to increase physical strength (see, for example, the Tactile Assault Light Operator Suit, TALOS)¹⁵. In the retail scenario, Roger purchases a Talos suit that greatly increases his bodily strength. Sophia buys a smart glove that adds extra strength to her hand. However, in the scenario Roger’s friends misjudge their physical limitations and exhaust themselves. The interplay between digital human enhancements and humans’ physical and psychological condition is largely unknown today.

The augmentation of humans’ physical capability raises ethical questions: Is it good or bad to artificially enhance one’s bodily condition? To what extent should users of digital body parts be granted access and manipulation possibilities to their own body devices? Must the software

¹³ Aim and scope as defined by [The Official Journal of the Institute for Ergonomics and Human Factors](#); URL: <http://www.tandfonline.com/action/journalInformation?show=aimsScope&journalCode=terg20>

¹⁴ <http://www.ergonomics.org.uk/>

¹⁵ <http://www.livescience.com/43406-iron-man-suit-prototypes.html>

used in the body be open and manipulable in order to avoid for instance outdated proprietary systems remaining in human bodies? Societies have developed rules around the use of drugs and poisons, and similar rules might be needed for body-enhancing technology.

Machines’ long-term effects on physical health and strength

Machines can directly impact physical health: immediately and over time. The previous examples of ergonomic design, universal usability, renewable body parts and new augmentation devices or implants are immediately observable and effective in the short term. But IT can also directly influence physical health in a way that is not immediately perceivable but has a long-term effect. Examples of include radiation or gaming in unnatural body positions.

Lets start with the issue of radiation, which causes serious health concerns in some people while others believe such fears to be esoteric phantasies. Health research distinguishes between thermic and athermic effects of radiation. Thermic effects (an increase in body temperature) can result from the absorption of energy by our biological body tissue, for example, absorption of radiation by the head when we hold a mobile phone to our ear. The official measure for such absorbed radiation is the Specific Absorption Rate “SAR,” which captures the relationship of Watt per kilogram of body mass (W/kg). The body reacts to the level of SAR to which it is exposed. Experimental animal research has shown that radiation beyond 4W/kg can damage the biological system ([Autonome Provinz Südtirol 2002](#)), so 4W/kg is a threshold level for the design of IT devices. Companies use accredited labs that perform SAR tests to investigate the radiation of devices according to standards that are published by standardization bodies such as IEEE and ISO. As of 2015, a typical smartphone like the Apple iPhone has an SAR value of around 0,95 W/kg.¹⁶

Besides the known thermic effects of radiation, athermic biological effects may occur in response to smaller SAR values. So far, little scientific knowledge exists on this issue. An Austrian public report lists potential negative effects like a change in the enzyme activity ornithin decarboxylase (which is associated with tumor growth) and an impact on cells’ calcium system and ionic transport ([Autonome Provinz Südtirol 2002](#)). Here a potential relationship is mentioned between high- and low-frequency fields and tumors, reproductive disorders, epilepsy, headaches, neurophysiological disorders (such as depression and disturbances of memory), disturbance of the immune system, damage of the eye tissue and risks specific to pregnant women, children and the elderly. The Assembly of the Council of Europe therefore strongly recommended applying the “ALARA” principle to SARs when IT devices are developed (ALRARY means “as low as reasonably achievable”).¹⁷

Long-term body reactions to machine use: Another long-term health effect of using IT is how our bodies react to long periods of regular digital immersion, such as when we play games, sit in front of a computer screen for work or use mobile phones. Eye strain and problems with the back and the tendons of the hand are well known. People’s posture is also influenced by their sitting position in front of screens. A core challenge is that we enjoy immersion and flow when we play digital games or do knowledge work we like. But this very positive immersion causes us to forget about our bodies. As a result, some gaming companies now think about mechanisms and even business models to encourage individuals to take more breaks.¹⁸

16 http://www.bfs.de/de/elektro/strahlenschutz_mobilfunk/schutz/vorsorge/smartphone_tipps.html

17 <http://www.assembly.coe.int/Mainf.asp?link=/Documents/AdoptedText/ta11/ERES1815.htm>

18 <http://www.ergonomics.org.uk/sport-leisure/gameplay-balancing-enjoyment-with-safety/>

However, business models promoting health can negatively affect companies' bottom line. In the gaming scenario, Stern reflects on the business case of Gaming The World Inc., which makes players pay after a certain amount of time so that they are incentivized to not continue playing too long. In that story I also offer a potential route for game design that may be healthier: bring games back into the real world and interact through voice commands and digital glasses.

Machines' direct effects on mental health and strength

Like physical health, mental health is directly influenced by machines. Myriad studies have looked into how Internet use influences depression, loneliness, self-esteem, life satisfaction, and well-being. A meta-study of 40 empirical investigations with over 20.000 participants found that high Internet use is directly associated with slightly reduced well-being ([Huang 2010](#)).

Yet, this overall tendency covers only part of the picture. People use the Internet for very different purposes, and each purpose affects our mental health in different ways. Using the Internet for general entertainment purposes, escape and acquiring information does not seem to have discernible consequences for well-being. Two forms of communication must be distinguished: using the Internet to strengthen our ties with existing friends (strengthening "strong ties") or using it to find friends (creating "weak ties"). Studies show: People who use the Internet to communicate with existing friends and family are less likely to be depressive over time. They use the medium in a positive way to foster communication. In contrast, people who use the Internet to overcome loneliness and meet new people are actually more likely to be depressed over time ([Bessie` re et al. 2008](#)).

This latter finding resonates in another observation, which relates "Problematic Internet Use" (PIU) to mental health. A mix of behaviors characterizes PIU: a salient intensive use of digital media, mood modifications and irritation when one is not able to access the Web, conflict with family and friends when access to the Web is impaired and a failure to stay away from using it even if this abstinence is desired ([Ko et al. 2005](#)). ([Caplan et al. 2009](#)) summarize studies that report significant correlations between PIU and loneliness, depression, anxiety, shyness, aggression, introversion and social skill deficits.

Researchers often question whether PIU causes such negative effects or whether existing individual traits such as loneliness and shyness lead to PIU.¹⁹ A cognitive behavioral model of PIU proposed by ([Davis 2001](#)) suggests that individuals who suffer from psychosocial problems are more likely to develop PIU. But research has also found that applications that are particularly social, such as online multiplayer games, foster PIU. Morahan-Martin explains "there is a growing consensus that the unique social interactions made possible by the Internet play a major role in the development of Internet abuse" (([Morahan-Martin 2007](#)), p. 335). Finally, researchers have found that women, poorer people and younger people are more likely to get depressed from using the Internet over time than others ([Bessie` re et al. 2008](#)). Taken together, mental health and social well-being are clearly related to the use of IT systems, but how this relationship plays out depends on who uses it, for what purposes and how.

¹⁹ <http://www.theguardian.com/technology/2010/feb/03/excessive-internet-use-depression>

On the positive side: While IT systems can add to mental problems and reduce social well-being, they are also used to relieve people in these very areas. A multitude of online services have been developed to ease dementia, phobia, anxiety, insomnia and addiction. In particular, mobile apps support people directly by giving them advice on their problems, tracking their behavior, putting them in touch with others, running them through relief games or providing reminders about things that would otherwise be forgotten.²⁰ Early studies, including one on student stress, found that a stress management app could influence its users’ weekly physical activity, engage in specific stress management methods, and exhibit decreased anxiety and family problems ([Chiauzzi et al. 2008](#)). Such findings are promising in that they suggest that IT systems can lead to some mental relief for those who need it. Perhaps they could be used to heal the same problems they may contribute to, such as PIU?

Machines’ indirect effect on mental health

I’ve outlined that Internet use has a small direct impact on mental problems or social well-being. However, considerable research has been conducted to understand the potentially more powerful *indirect* influence that Internet use has.

One line of research looks at PIU. Intensive use of the Internet can absorb our time to such an extent that we become stressed when we try to meet other life obligations. This stress, which we perceive in everyday activities, can lead to mental health problems. For example, according to Ming, students with heavy Internet use report that the Internet jeopardizes their academic performance. Poor academic performance is accompanied by high academic stress, which again impacts mental health negatively ([Ming 2012](#)).

A second line of research investigates a *social displacement* hypothesis, which suggests that computer and Internet use reduces the time we spend to maintain social resources. A lack of social ties undermines our mental health ([Ming 2012](#)) because we then need to find new friends online ([Bessie` re et al. 2008](#)). However, if we find true friends online and develop strong ties with them, our health can be strengthened again. Scholars refer to such developments as *social augmentation* or *social compensation* (McKenna and Bargh...).

A third line of research is based on the mood enhancement hypothesis, which posits that we selectively expose ourselves to media content based on our mood and can thereby disrupt a bad mood or negative ruminations on our life. The use of digital media for such purposes relieves stress, which again is good for mental health. ([Ming 2012](#)) show that mental health problems experienced by students with high academic stress can slow down as a result of exposure to mood-enhancing media.

Finally, a well-known negative indirect effect of machines on health has been observed in the field of online gaming, especially when games are addictive. Addiction to online games has physical and mental consequences such as migraines, sleep disturbance, backaches, eating irregularities, carpal tunnel syndrome, agoraphobia and poor personal hygiene. Because online gaming addiction is such a prominent problem today, self-help organizations have been

²⁰ <http://apps.nhs.uk/>

founded that provide people with information and advice.²¹ “Massive Multiplayer Online Games” (MMOGs) can be particularly addictive because they can create a feeling of social community among players. In particular, players who have a greater sense of friendship online are more likely to become addicted. As with PIU, social variables, offline friends and social ties or feelings of loneliness play a predictive role for games. Game immersion and the use of voice in games were found to encourage addiction (Caplan et al. 2009).

Figure x summarizes at a generic level the indirect relationships observed between computer use and mental health. A very specific kind of mental health problem arises in the form of burnout on the job.

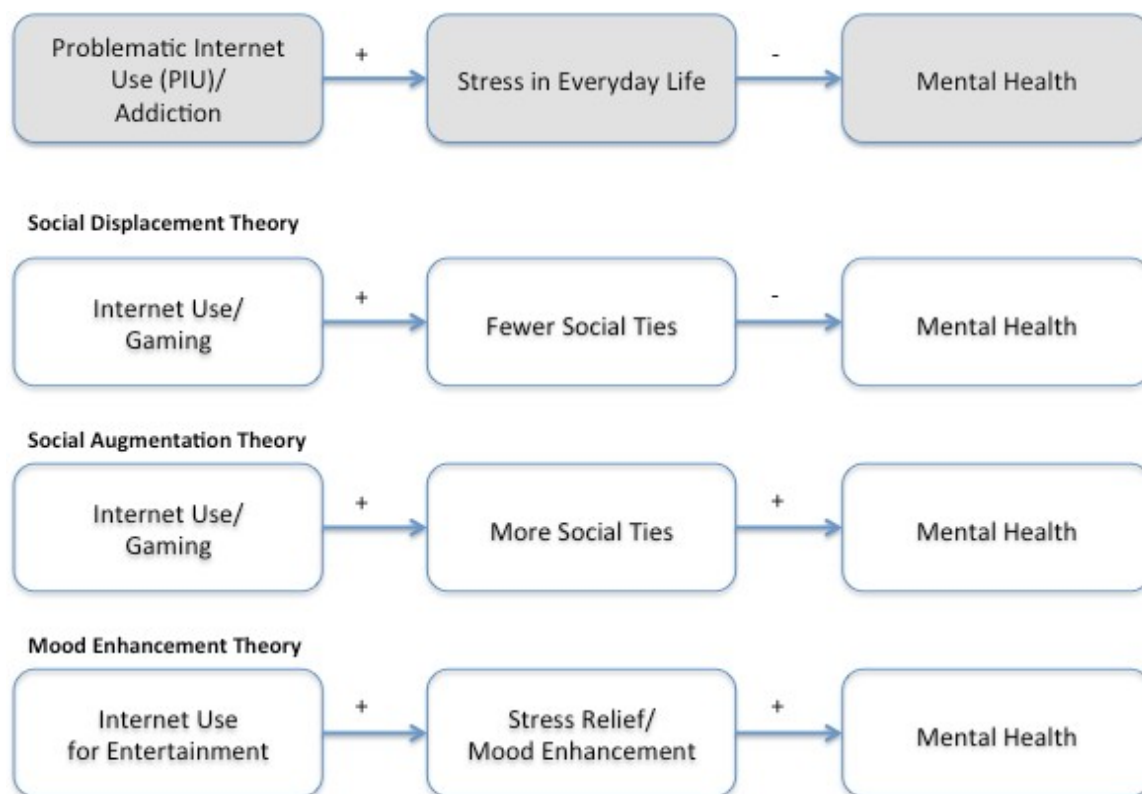


Figure: Selected indirect paths of IT influence on mental health

Mental health challenges in response to computer use on the job

Another indirect influence of IT on mental health involves employee burnout. Burnout is a phenomenon that expresses itself in feelings of physical exhaustion and cynicism ((Green et al. 1991), p. 463). It is caused by a perceived misalignment between a job's demand and control over the job (see the “Job Demand–Control (JD–C) model,” (Karasek 1979; Karasek 1990)). Employees get stressed and develop burnout symptoms when job control is low and job demands are high. On the contrary, more job control can attenuate the negative effects of job demands on strain. Scholars found that perceived computer self-efficacy is a key factor moderating this relationship between job control and job demand as depicted in figure x

21 <http://www.video-game-addiction.org/most-addictive-video-games.html>

([Salanova et al. 2010](#)). Computer self-efficacy can reinforce or appease the burnout symptoms of exhaustion and cynicism. Self-efficacy refers to an individual’s belief in his or her ability to perform a specific task ([Bandura 1977](#)). Computer self-efficacy is “an individual’s perception of efficacy in performing specific computer-related tasks within the domain of general computing” (([Marakas et al. 1998](#)), p. 128). It is driven to some extent by prior experience and by individual traits and personality variables such as age or professional orientation. However, it is also affected by specific characteristics of work on a computer, including the complexity, novelty and difficulty of a task as well as situational support ([Marakas et al. 1998](#)). When computer systems are not predictable, show incomprehensible numbers or are not well-documented and therefore difficult to understand, employees can perceive a lack self-efficacy. This perception, combined with management’s high demand for documentation and number-driven decisions, can lead to perceptions of loss of control and then burnout.

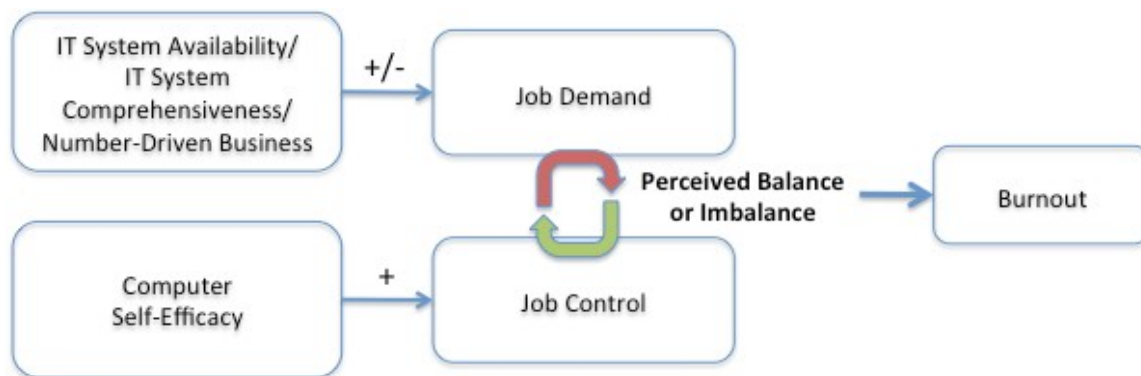


Figure x: Schematic relationship between corporate IT systems, computer self-efficacy and burnout (extended from Salanova et al.)

Machines’ indirect effect on physical health

In the context of the OECD’s work on a “better life index,”²² researchers have recognized that chronic diseases such as cancer, cardiovascular diseases, chronic respiratory conditions and diabetes now cause around three-quarters of all deaths in OECD countries. Many of these diseases could be prevented if people modified their lifestyle. People who drink alcohol in moderate quantities, are physically active, eat a balanced diet, do not smoke and are not overweight or obese have a much lower risk of early death than those who have such unhealthy habits. Against this background, we must ask whether machines can help us to be more self-aware and self-disciplined and mentally support and coach us to live a healthier life. In the scenarios, I describe the level of intelligence that machines can advance to. Taking a form like Sophia’s Arthur, they might be able to continuously collect our health data and activity levels; machines could pull data from life-logging bracelets, smart textiles or smart phones. On today’s market, these types of applications are called “life-logging” devices ([European Network and Information Security Agency \(ENISA\) 2011](#)) or “quantified-self” services ([Swan 2012](#)). Step counters like Fitbit²³, diet support apps like MyNetDiet²⁴, quit-

22 <http://www.oecdbetterlifeindex.org/topics/health/>

23 <http://www.fitbit.com/uk/story>

24 <http://apps.nhs.uk/app/calorie-counter-pro-by-mynetdiary/>

smoking apps like Smoke Free²⁵, anti-alcoholism trackers like Change4 Life Drinks Tracker²⁶ or sleep trackers like Sleep²⁷ are designed to make people aware of their behavior and support them in changing it to the positive.

On the negative side, I describe in the retail scenario how Roger worries about friends who feel pressure to meet fitness norms that may not be right for everyone. Here, a fitness-enhancing suit (Talos) creates a trade-off with one’s perceived freedom. I characterize life-logging services as having an indirect effect on physical health only because the service needs to be able to motivate people to change their behavior both in the short- and long-term. Not all services will live up to this expectation.

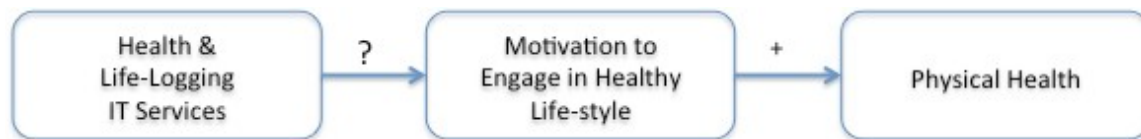


Figure z: Motivation to engage in a healthy lifestyle mediates the effect of services

As health-tracking applications advance, more health data will be available for analysis and predictive modeling (Manyika et al. 2011). If it was possible to track, store and analyze individuals’ health data (in a way that preserves privacy) and increased the amount of objective data on medications success, therapy effectiveness, doctor- and hospital quality etc., we can gain a better feeling for what to do and where to go. More informed healthcare decisions may become possible in comparison to today, when differences in cost and quality are largely opaque (Manyika et al. 2011). Healthcare data also supports comparative effectiveness research (CER), which explores what medical treatments and medications work best, under what conditions and for what kinds of people. Such data may also be used to study rare adverse drug reactions and mutual drug intolerances. Finally, machines may be used as support tools in diagnosis, helping doctors to analyze X-ray, CT and MRI output.

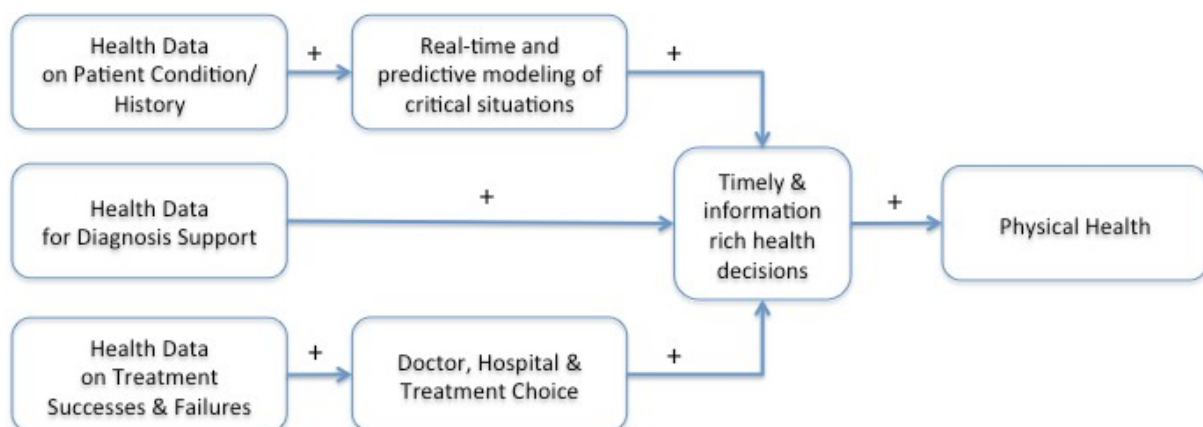


Figure x: Potential positive health effects of the use of health data for more timely and informed decision-making (“+” denotes the commonly presumed direction of influence)

25 <http://apps.nhs.uk/app/smoke-free/>

26 <http://apps.nhs.uk/app/change4life-drinks-tracker/>

27 <http://sleep.motionx.com/>

However, the use of electronic health records in these ways raises many ethical questions: Should we ever be forced to share personal health data with care providers, researchers or third parties such as insurers and health data brokers? To what extent and under what legal and technical conditions can a market for health data be legitimate? If health data such as genetic data or body measures reveal that we have a high risk for becoming sick, how and to what extent may this data be used by entities such as employers or insurers? Who is liable if the information is wrong or predictions don't play out as anticipated? To what extent could people be directly or indirectly forced to use health-monitoring applications to reduce their health risks? And who –if anyone – is allowed to exert such pressure? To what extent should in-body monitoring devices, such as chips that are inserted into the body or blood to monitor bodily conditions, be marketed to the general public?

Exercise:

- Pull out the various potential health effects that may be created through the applications described in the scenarios in chapter x.

Later Chapter

- Address three of the ethical questions raised by the use of digital health data from a utilitarian point of view.

On Security and Safety in the Machine Age

*“Those who surrender freedom for security
will not have, nor do they deserve, either one.”
(Benjamin Franklin, 1706-1790)*

“Everyone has the right to life, liberty and security of person.” With these words, Art. 3 of the Universal Declaration of Human Rights stresses the value of security for human beings ([UN General Assembly 1948](#)). At the same time, Benjamin Franklin’s famous words indicate that freedom might be even more important than security. Franklin’s view is echoed by many liberal thinkers, who oppose what they call “Orwellian surveillance states” – states that monitor their citizens and limit liberty in the name of security. But are liberty and security mutually exclusive? Would the authors of the Universal Declaration of Human Rights have wanted to suggest that security was more important than liberty; that there is a value hierarchy between the two constructs?

Safety versus Security

The imprecise use of the term “security” creates confusion and suggests a conflict with liberty that may not exist to the extent that some believe. One reason for the confusion is that public authorities and the media often use the term “security” when they really mean “safety.” When citizens are asked for instance whether freedom or security is more important for them, respondents often favor security at first.²⁸ It seems as if they positively embraced and legitimized governments’ surveillance programmes in a dire need for security and at the expense of their freedom. However, it may be that they just misunderstood the question. What they really thought about when answering the question was their ‘safety’ rather than the security that enables it as I will show below.

Unfortunately, most languages do not clearly differentiate between safety and security. For example, security and safety are often collapsed into one term: “Sicherheit” in German, “seguridad” in Spanish, “seguranca” in Portuguese, “säkerhet” in Swedish, and so on. The Wikipedia definitions of security and safety overlap as well: “Security is the degree of resistance to, or protection from, harm.”²⁹ “Safety is the state of being “safe,” the condition of being protected against...harm ...”³⁰. So what is the difference?

One difference can be understood when looking into the details of Wikipedia’s security definition: “Security is the degree of resistance to, or protection from, harm. It applies to any vulnerable and valuable asset, such as a person, dwelling, community, nation, or organization”. (FN x) The term security hence encompasses several levels of analysis. National security, organizational security and individual security are all different things. Only when we speak about individual security do we approach the meaning of the word safety, because in its full definition, safety is more concerned with individual human beings: “Safety

28 Heise Online, Deutsche Telekom asks citizens what is more important: Freedom or Security:
<http://www.heise.de/newsticker/meldung/Telekom-fragt-Was-ist-Ihnen-wichtiger-Sicherheit-oder-Freiheit-1980972.html> (last visited on August 2nd 2014)

29 Wikipedia (URL last visited on August 1st 2014): <http://en.wikipedia.org/wiki/Security>

30 Wikipedia (URL last visited on August 1st 2014): <http://en.wikipedia.org/wiki/Safety>

is the state of being "safe" (from French *sauf*), the condition of being protected against physical, social, spiritual, financial, political, emotional, occupational, psychological, educational or other types or consequences of failure, damage, error, accidents, harm or any other event which could be considered non-desirable" (Fußn. x). So the first thing we should note is that when the media or market agencies ask citizens or consumers about their "security," then they should first distinguish the level of security they are actually talking about: national, organizational or individual?

To illustrate the importance of this distinction consider the following: If people were asked whether the security of the organization they work for or their personal freedom was more important, most would probably choose their personal freedom. If they were asked whether the security of their country or their personal freedom was more important, the answer would probably depend on the context; in particular the degree to which a country's security seemed threatened. Right after September 11th 2001 for example, many people hoped for a high security level in their countries. They felt their country to be threatened by potential terrorist attacks. As a result, they were willing to give up some personal freedoms, such as some of their digital privacy rights, to ensure that the state would protect them through higher levels of national security. A few years later after the context had changed the perspective changed as well: When US citizens learned how much of their privacy freedoms had been infringed in the name of national security – i.e. through the X Act – laws had to be taken back.

September 11th is a good showcase to dig into the specific tension field of national security vs. personal liberty: People at the time were evidently concerned about national security. Many feared that they or their families could be personally impacted by the developments. But would they have traded their liberty for national security? I would argue: on the contrary! In the aftermath of the September 11th terrorist attacks, many people acted in ways that were deep expressions of personal liberty: people vividly discussed protective measures in case of war, considered moving to the countryside, bought gold, hoarded food, and so on. The personal freedom to plan and to take small actions strongly stabilized people's emotions and helped them feel secure again. As Abraham Maslow would argue, personal freedom and liberty of the people was a precondition for them to feel secure again at an individual level. Take another example to illustrate this: Think about a person that has been kidnapped and is now imprisoned in a cellar. Her kidnapper does not touch her and the cellar in which she is kept is secure and potentially even comfortable. But will this person *feel* „secure“? No. Because she is deprived of the liberty to move she cannot feel secure. The two examples show that at an individual level liberty is a precondition for one's perceived security. The two cannot be traded off. When people are not free to personally react to a threat, they don't feel secure. And this observation brings us back to what Benjamin Franklin said (who obviously recognized the impossible trade-off): "Those who surrender freedom for security will not have...either one."

The arguments prove that the discussion of security versus liberty requires that the two constructs are treated at the same level of analysis (individual, organizational, national). But the use of the term security bears another pitfall: Often it is confounded with safety. Take the timely issue of airport "security": Often we are being asked whether we are willing to have our movements restricted at airports to increase the "security" of those airports. But do we really care about the security of airports? Probably not! What we are really concerned about is not the level of *security* at the airport but the perceived safety that a respective level of security creates for us. At the organizational and national level, security is a precondition for safety: if a system is secured in that it cannot be tampered with by malicious attackers, then the likelihood that it will damage people is reduced. Security is hence only indirectly

important for us as individuals.

The distinction of security and safety becomes visible when comparing their academic definitions beyond Wikipedia. (Line et al. 2006) define security as “the inability of the environment to affect the system in an undesirable way.” In contrast, they define safety as “the inability of the system to affect its environment in an undesirable way.” “Security is concerned with the risks originating from the environment and potentially impacting systems, whereas safety deals with the risks arising from the system and potentially impacting the environment” ((Piètre-Cambacédès et al. 2010), p. 59).

While typical security threats can be identified across industries, safety standards are typically specific to every industry and type of machinery. Industry experts define safety standards for their respective (highly specialized) domain, aggregating years of knowledge and experience in how to build and maintain machines to avoid accidents. Safety is mainly linked to the avoidance of accidental risk while security is more thought of in terms of malicious attacks. An example for a safety standard is ISO 10218, the standard for industrial robots that aims to prevent accidents and harm by specifying how robots on the shop floor should operate, how they can be stopped, how fast they are allowed to move, what radius they are allowed to span, how their electric plugs are connected, and so on (ISO 2011).

Figure x illustrates this distinction and clarifies that people’s concerns (depicted as a stick figure) are triggered directly only by their desire for safety rather than security.

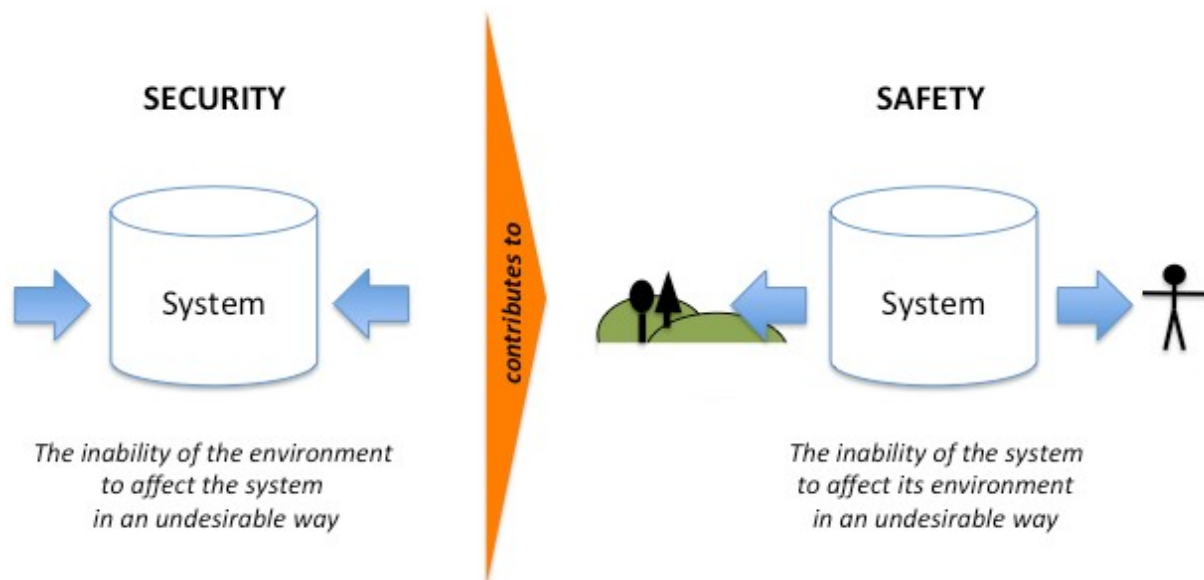


Figure x: The nature of and difference between security and safety

Safety, Cyberwar and Cybercrime

As more and more machines are digitized, networked or both, and as more machines receive instructions and upgrades from central computers, security is becoming an increasingly relevant factor for safety. Modern attackers can compromise the central computer systems that

handle safety-critical infrastructure by infiltrating a system remotely and gaining control over it. Malicious attackers can also infect subcomponents of a system that are then built into safety-critical machinery. In any case, risk is created for system abuse. For example, cyber attackers could take over power grids and switch off the electricity supply in the area of a nuclear power plant. They could – as with the Stuxnet worm - implement a virus in parts of a critical infrastructure. Discovered in 2010, the Stuxnet worm attacked industrial programmable logic controllers, which control systems such as industrial assembly lines, amusement park rides and centrifuges that separate nuclear material. It has been reported that Stuxnet destroyed one-fifth of Iran's nuclear centrifuges.³¹³² Scenarios like this one now populate common threat models for "cyberwar." States are worried that other states, terrorists or criminals could compromise the security of their critical infrastructure and gain a position in which they could cause war-like harm to citizens or gain the power of extortion.

Such "cyberwar" threats must be distinguished from "cybercrime". Cybercrime is "any crime that involves a computer and a network. The computer may have been used in the commission of a crime, or it may be the target."³³ Some national authorities stress that digital data processing must be *essential* for carrying out the crime. This means that for them a crime is not a cyber crime unless digital data processing was important to commit the crime. Since so many crimes today use some kind of computer though the term cybercrime got very broad in the amount and kind of activities it covers. It currently includes activities such as "computer-related copyright or trademark offences" that already happen when someone uses an illegal video-streaming platform. Figure x summarizes acts that are considered to constitute a cybercrime according to the United Nations Office on Drugs and Crime (UNODC) (([United Nations 2013](#)), p. 16).

31 Wikipedia "Stuxnet": <http://en.wikipedia.org/wiki/Stuxnet> (last visited on August 2nd 2014)

32 Business Insider: <http://www.businessinsider.com/stuxnet-was-far-more-dangerous-than-previous-thought-2013-11> (last visited on August 2nd 2014)

33 Wikipedia "Cybercrime": http://en.wikipedia.org/wiki/Computer_crime

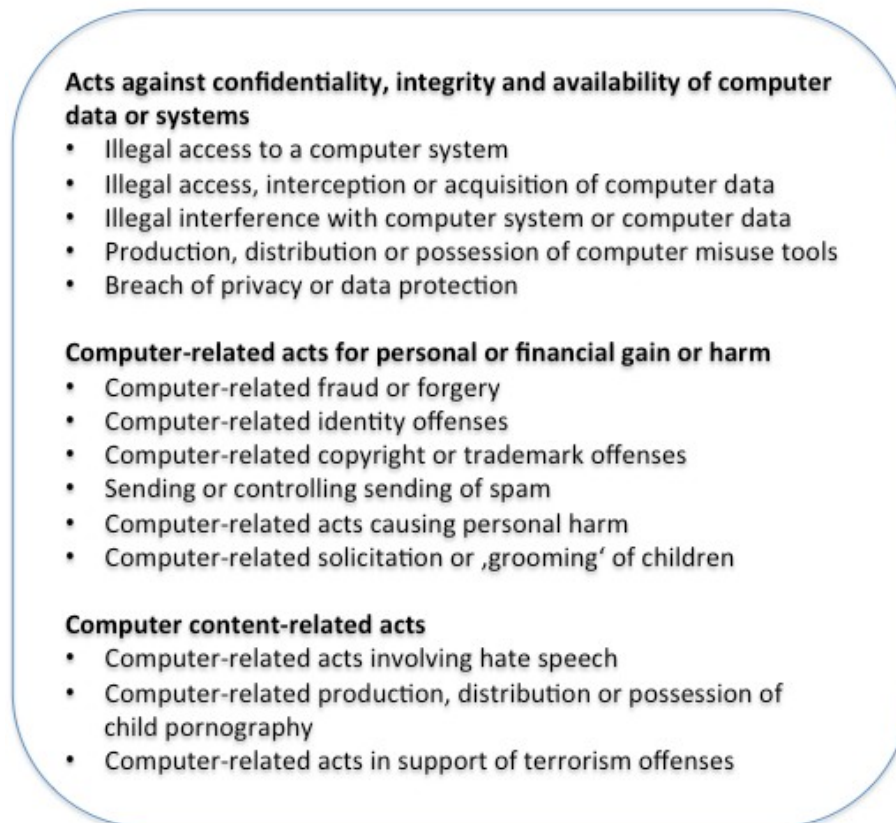


Figure x: Acts constituting cybercrime according to UNODC (2013)

Security Principles in Machine Engineering

The first category of cybercrimes defined by the United Nations Office on Drugs and Crime involves the illegal access and use of computer systems resulting in a loss of confidentiality, integrity and availability of data. Confidentiality, integrity and availability of data are sometimes abbreviated as the “CIA principles” of data security ([ISO 2014a](#); [NIST 2013](#)).

For companies, CIA-related acts of cybercrime are the most important threats ([United Nations 2013](#)). Companies worry that malicious attackers might access and steal intellectual property, customer data, trade secrets and so on (cyber espionage), penetrate point-of-sale payment systems to steal and misuse customer payment cards (POS intrusions), tamper with or steal corporate data (including customer data), paralyze operations through denial-of-service attacks or install malware that damages operations (immediately or later). Such fears are not unfounded. In 2013 alone, Verizon reported 63.437 security incidents that compromised the integrity, confidentiality, or availability of information assets. In 1.367 of these cases, the data was breached, which means that the incident resulted in the disclosure or potential exposure of data ([Verizon 2014](#)).

Private individuals are often involved in this kind of crime, for example, when they respond to phishing e-mails, reveal their credentials, have their credit card data stolen, etc. In fact, the victimization rate for cyber crimes is significantly higher than for conventional crime forms, potentially because criminals don’t need to be physically near their victims. The UNODC (2013) reports that victimization rates for online credit card fraud, identity theft, responding to a phishing attempt, and experiencing unauthorized access to an email account range from 1% to 17% of the online population for 21 countries across the world, compared with typical

burglary, robbery and car theft rates of under 5% for the same countries.

People become aware of online security threats even if they are not physically harmed by them. In 2010, 86% of consumers around the globe said that they are becoming more security-conscious about their data, and 88% worried about who might have access to their personal data ([Fujitsu 2010](#)). People’s concerns are relevant for companies because customers’ perceived security influences the extent to which customers intent to do business with a company online. The *perceived* level of data security at a company significantly influences the level of trust consumers place in that company ([Chellappa et al. 2002](#)), especially in banking ([Yousafzai et al. 2010](#)). To maintain people’s long-term trust in systems and avoid cybercrime damage, companies now systematically pursue relevant security goals when they build and maintain their systems. They analyze the extent to which security goals are at risk and put appropriate countermeasures in place to mitigate those risks. Chapter x outlines how organizations can run through risk analysis systematically.

Information Security Goals

Information security goals aim to protect the information qualities threatened by cybercrime: confidentiality and integrity of data and availability of services. To guard the *confidentiality* of information, data must be encrypted, and access and use must be confined to authorized purposes by authorized people and systems. To achieve this, information is ideally classified in terms of sensitivity, criticality and value to the organization. Information is labeled accordingly, and access rights are set (ISO/IEC 27002: 2005). Employees are asked to authenticate to access systems, especially those that hold sensitive, critical or valuable information. Employees must be authorized to conduct certain operations on the information or to use the information. Furthermore, organizations should create general security awareness. Passwords need to be strong so that they are not easily hacked, and employees must not share system access credentials.

Information has *integrity* when it is whole, complete and uncorrupted. The integrity of information is threatened when it is exposed to corruption, damage, destruction, or other disruption of its authentic state. Corruption can occur not only while information is stored but also when it is transmitted. Many computer viruses and worms are designed explicitly to corrupt data. A key method for detecting a virus or worm is to look for changes in file integrity. Integrity can be checked, for example, by monitoring file size. A stronger method for assuring information integrity is file hashing. Here, a special hash algorithm reads a file and uses its bits to compute a “hash value,” a single number based on the individual data points of the file to be protected. This hash value is stored for each file. To ensure that a file is trustworthy, a computer system later repeats the same hashing algorithm before accessing the content of the file. If the algorithm returns a different hash value than the one stored for the file, then the file has been corrupted and the integrity of the information is lost.

Availability enables authorized users—people or computer systems—to access information without interference or obstruction and to receive it in the required format. Availability can be compromised when the service falls victim to a denial-of-service attack or has been altered so that authorized clients cannot access it any more.

Beyond these three CIA criteria, scholars and practitioners have suggested considering further security goals, such as data authenticity, data accuracy and system auditability and nonrepudiation ([Cherdantseva et al. 2013](#)). I will concentrate here information qualities that must be protected instead of system characteristics that serve to protect these qualities.

Therefore, I first include only authenticity and accuracy, at least for the purpose of figure x which summarizes security protection goals. *Authenticity* of information is the quality or state of being genuine or original. Information is authentic when it is in the same state as when it was created, placed, stored, or transferred. Note the difference between integrity and authenticity: While integrity focuses on information or data not being falsified, authenticity focuses on where the information originated. Suppose that attackers infiltrate an organization by using malware. They could send you information that claims to come from a trustworthy source but actually send something like a virus or phishing email. Nonrepudiation provides a system with the ability to determine whether a trusted individual took a particular action such as creating information, sending a message, approving information or receiving a message. A system that supports nonrepudiation also prevents individuals from falsely denying that they performed a particular action ([NIST 2013](#)).

Finally, accuracy is a goal that I previously mentioned in chapter x, where I outlined that data quality must be maintained. Data accuracy or quality, which requires that data be free from mistakes or errors, is not only a goal to ensure the quality of knowledge but also a goal that supports the security of an organization. Imagine that, due to an error in a system database, an alarm threshold is triggered that causes damaging actions. The accuracy of data could be altered without necessarily changing its integrity. Similarly, it is possible to add erroneous information to databases based on false assumptions or to add false (even malicious) information to accounts.

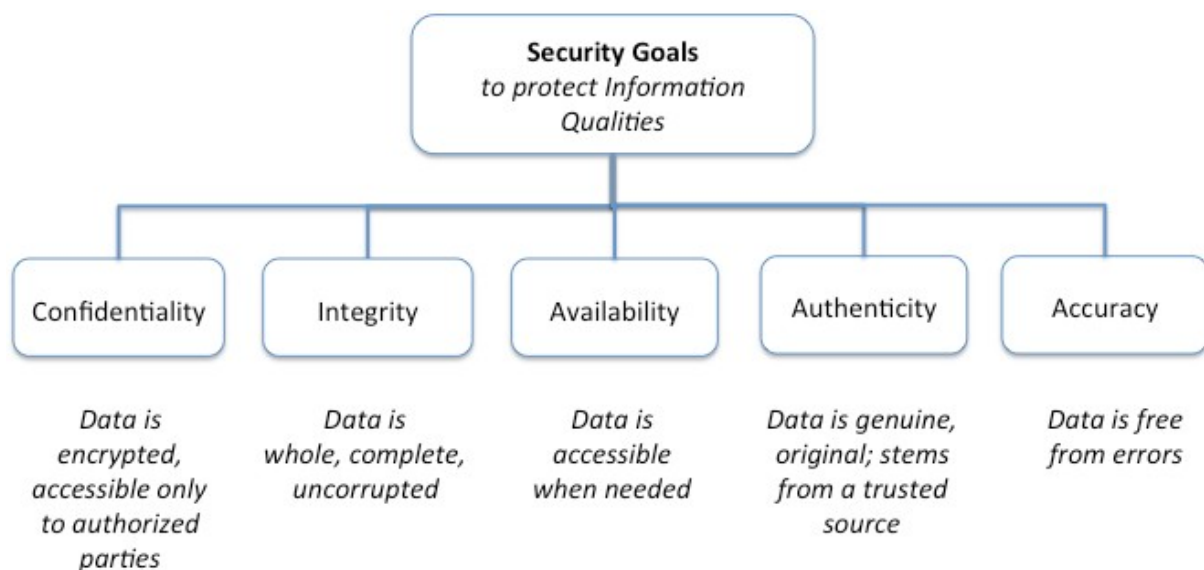


Figure x: Recognized Information Security Goals

Auditability

Auditability is recognized as a relevant system trait that can be used to ensure compliance with security goals ([ISO 2012](#)). ([Cherdantseva et al. 2013](#)) define auditability as a system's ability "to conduct persistent, nonbypassable monitoring of all actions performed by humans or machines within the system." Auditability of systems gained significantly in importance since the 2002 Sarbanes-Oxley Act (SOX), which introduced new transparency and audit standards for organizations. In section x, I described how the auditing company Arthur Anderson failed to properly

assess the financial risks associated with the practices of the company Enron. Enron filed for bankruptcy after financial fraud was uncovered that Arthur Anderson's auditing practices did not detect. In response to the fall of Enron and Arthur Anderson, corporations that are listed on the US stock exchange must now report to the Securities and Exchange Commission (SEC) that they comply with SOX. SOX forces companies to monitor and evaluate all relevant processes that could influence their accounts. "While the topic of information security is not specifically discussed within the text of the act, the reality is that modern financial reporting systems are heavily dependent on technology and associated controls. Any review of internal controls would not be complete without addressing the information security controls around these systems. An insecure system would not be considered a source of reliable financial information because of the possibility of unauthorized transactions or manipulation of numbers" ([SANS Institute 2004](#)). SOX has hence indirectly enforced the scrutiny of information security controls. One of the main methodological standards associated with this form of scrutiny are the Common Criteria for Information Technology Security Evaluation, which are elaborated by ISO in ISO/IEC 15408 ([ISO 2012](#)). These "Common Criteria" are used as a framework to formulate security requirements for systems. The security requirements are then used to analyze and configure systems' security levels. One of these criteria is the use of audit trails for all "security relevant activities" (pp. 29 et seq., pp. 183 et seq.).

For ordinary people, "security relevant activities" seem to be those that help to protect their personal information. But end-user data protection or privacy is not necessarily in the spotlight when systems are designed and audited for security goals. Often, only the personal information about customers or employees that is also financially relevant to the company is part of security efforts. For example, a bank will take great care to protect its customers' bank account details. The financial information of the customer is, in this case, a crucial part of the bank's assets. In contrast, a car manufacturer may hold some information about the buyers of its cars, but a security audit would not prioritize the protection of this buyer data. Instead, the audit would focus on protecting the systems that handle manufacturing and logistics processes because these processes are financially more relevant to the car manufacturer. It is important to recognize that security efforts and privacy efforts are not necessarily the same because the terms "security" and "privacy" are often confounded. Laymen easily equate the two and believe that their personal data is automatically better protected when companies talk about improving their security. That said, security audits *can* (and increasingly do) embrace people's privacy concerns (or privacy regulation). In this case, "privacy targets" become an integral part of "security relevant activities." Companies can pursue two strategies to meet their privacy targets beyond data encryption and typical CIA measures. One is to "minimize" personal data (as recognized in the Common Criteria's "Privacy Section"). The other is to control data flows by using policies and audit trails.

Privacy in terms of Security vs. Security in terms of Privacy

Data minimization means that a company keeps only personal data records that are needed for its business. All of the rest of the personal data, customer transaction histories, and so on are anonymized so that they can no longer be attributed to a unique individual ([ISO 2012](#)). Data minimization through anonymization has been a very efficient strategy for companies to avoid privacy problems with customers. If data is anonymized or otherwise minimized, companies presume, customers cannot be harmed and data protection law does not apply. ([Ohm 2010](#)) has commented that "nearly every information privacy law or regulation grants a get-out-of-jail-free card to those who anonymize their data" (p. 1704).

Box x in section x has described anonymization and pseudonymization techniques. In addition, unlinkability and unobservability are technical ways to protect personal data. Unlinkability means “a user may make multiple uses of resources or services without others being able to link these uses together” (([ISO 2012](#)), p. 122). For example, a website operator might not log and link multiple visits of its customers to form an interaction profile over time or track how users move through sites. Unobservability means that “a user may use a resource or service without others, especially third parties, being able to observe that the resource or service is being used” (([ISO 2012](#)), p. 123). For example, a website operator does not log the IP addresses of those who read the content he provides.

Security experts often argue that data minimization efforts (as well as CIA and encryption of personal data) fully address people’s digital privacy. They equate their security efforts with the creation of privacy. Privacy - *from their perspective* - is part of the overall security effort in a company. And certainly, many of the privacy harms described in section x, such as unauthorized secondary data uses, breach of confidentiality, public disclosure and exposure, could be avoided *if* personal data records were systematically minimized. But other scholars, in particular those from the legal studies, from NGOs or social sciences counter that security efforts do not suffice to create privacy. They see security as just one piece of the puzzle to ensure people’s privacy in terms of information self-determination. From these thinkers’ perspective, which is strongly adopted in Europe, privacy is not only a passive right. Privacy is not only something that can be *harmed* in various ways (and hence needs some security protection). Privacy is also an active right that allows people to freely determine who can use their data, when and for what purposes. From this perspective, privacy is defined in terms of control over access to the self: “Privacy, as a whole or in part, represents control over transactions between person(s) and other(s), the ultimate aim of which is to enhance autonomy and/or minimize vulnerability” (([Margulis 2003](#)), p. 245). This autonomy-embracing definition of privacy is mirrored in European privacy legislation. Here, citizens need to opt-in to the use of their data *before* a data collector can use it (note that, in the US, citizens can only opt out of the use of their data). European citizens need to be informed up-front on the purposes of data use, they can access their data after they reveal it, withdraw their consent to its use, and so on ([European Parliament and the Council of Europe 1995](#)). From this informational self-determination perspective on privacy, security is just one part of a larger human rights endeavor.

Privacy scholars who share the broader perspective on “privacy as information self-determination” advise companies to ensure people’s participation and/or potential control over data exchange and data use. They argue that personal data should be treated as a shared asset in a way that is negotiated with customers in policies (ISO/IEC 29101). Privacy policies can specify data usage rights and restrictions. They may be negotiated prior to the data exchange with the help of a protocol such as P3P ([Cranor et al. 2006](#)). A Web protocol such as HTTPPA (HTTP with Accountability) can be used to transmit usage restriction policies between web servers and clients ([Seneviratne et al. 2014](#)). Companies create a detailed and transparent history of access requests for personal data, and they also track the transfer, processing and disclosure of privacy-critical information. For example, the HTTPPA protocol creates audit logs every time a party wants to access and use personal data, and people can check what happened to their data ([Seneviratne et al. 2014](#)). The logs can later serve customers, auditors and those “accountable” in organizations to check whether personal data usage was compliant with privacy policies.

Accountability

According to various standards, guidelines and laws, companies are “accountable” for their security and privacy practices ([Alhadeff et al. 2011](#); [Organisation for Economic Co-operation and Development \(OECD\) 1980](#)). Yet the term is imprecise, because it can refer to responsibility at various levels: In computer science, the term accountability is used mainly in reference to auditability, the use of nonrepudiation mechanisms, or both. At a higher organizational level, accountability is situated with an individual in an organization who must safeguard and control equipment and information. This individual is also responsible to the proper authorities for the loss or misuse of that equipment or information ([CNSS 2010](#)). At a still higher level, accountability simply denotes that a company must be held responsible for its actions. No matter what level accountability is defined at, individuals (and their organizations) can meaningfully bear responsibility for transactions only if their IT systems provide them with the necessary information. ([Weitzner et al. 2008](#)) clarify: “Information accountability means the use of information should be *transparent* so it is possible to determine whether a particular use is appropriate under a given set of rules” (p. 84). (see section x on transparency)

To prove that they are legally compliant and accountable to users and the legislator, organizations can take concrete accountability measures for personal data, such as those standardized in the ISO/IEC 29100 standard “Information technology – Security techniques – Privacy framework” ([ISO 2014b](#)). Box x cites these accountability measures for “Personal Identifiable Information”(PII).

Box x

Accountability for personal data according to ISO/IEC 29100:2011 (E), Sec. 5.10 Information technology – Security techniques – Privacy framework

“The processing of PII entails a duty of care and the adoption of concrete and practical measures for its protection. Adhering to the accountability principle means:

- documenting and communicating as appropriate all privacy-related *policies*, procedures and practices;
- assigning to a specified individual within the organization (who might in turn delegate to others in the organization as appropriate) the task of implementing the privacy-related *policies*, procedures and practices;
- when transferring PII to third parties, ensuring that the third party recipient will be bound to provide an equivalent level of privacy protection through contractual or other means such as mandatory internal policies (applicable law can contain additional requirements regarding international data transfers);
- providing suitable training for the personnel of the PII controller who will have access to PII;

- setting up efficient internal complaint handling and redress procedures for use by PII principals;
- informing PII principals about privacy breaches that can lead to substantial damage to them (unless prohibited, e.g., while working with law enforcement) as well as the measures taken for resolution;
- notifying all relevant privacy stakeholders about privacy breaches as required in some jurisdictions (e.g., the data protection authorities) and depending on the level of risk;
- allowing an aggrieved PII principal access to appropriate and effective sanctions and/or remedies, such as rectification, expungement or restitution if a privacy breach has occurred; and
- considering procedures for compensation for situations in which it will be difficult or impossible to bring the natural person's privacy status back to a position as if nothing had occurred.

Measures to remediate a privacy breach should be proportionate to the risks associated with the breach but they should be implemented as quickly as possible (unless otherwise prohibited, e.g., interference with a lawful investigation)."

PII = personal identifiable information: any information that (a) can be used to identify the PII principal to whom such information relates, or (b) is or might be directly or indirectly linked to a PII principal

PII controller = privacy stakeholder (or privacy stakeholders) that determines the purposes and means for processing personally identifiable information (PII) other than natural persons who use data for personal purposes

PII principal = natural person to whom the personally identifiable information (PII) relates

Privacy and Surveillance

Let's assume a perfect world: companies have done their best to optimize the security and privacy levels for personal data. They encrypt personal data where possible and verify information quality according to the CIA criteria as well as the authenticity and accuracy of the data. They also act accountably in line with ISO/IEC 29100, ensuring that people's consent-based privacy policies are respected within and beyond corporate boundaries. Information self-determination is respected by the corporate world. People can participate as sovereigns in personal data markets; they can share their data for research purposes or marketing campaigns for appropriate returns or choose to keep their data private.

If all of these measures were taken diligently, organizations would minimize the risk of causing privacy harms. People may be less concerned about their security and feel more in control than they do today. Their identities would be stolen or misrepresented less often. When future robotic systems physically interact with humans, the risk that such systems could be controlled remotely or misrepresent users would be reduced. Also, software agents that act on people's behalf (based on data exchange policies) would not betray their owners by disclosing unauthorized personal data to data collectors without the owner's consent. In such a perfect world, people could feel sufficiently safe and secure vis-à-vis their machines, at least at a basic level (in Maslow's sense). People could trust their machines.

But one major ethical challenge remains and spring to mind when we talk about "security": What degree of surveillance is acceptable in the name of security to guard people's liberty?

The issue of surveillance

"*Surveillance* is the watching, listening to, or recording of an individual's activities" (([Solove 2006](#)), p. 490). Roger Clarke refined this baseline definition of surveillance and distinguished what he calls "dataveillance" ([Clarke 1988](#)): Dataveillance refers to the **systematic use of**

personal data systems in the investigation or monitoring of the actions or communications of one or more persons. Clarke distinguishes between "personal dataveillance" of previously identified individuals and the "mass dataveillance" of groups of people. With this distinction, Clarke hints at qualitative differences between classical forms of surveillance Daniel Solove refers to and the kind of dataveillance modern machines enable. The distinction is important. Unlike classical surveillance in the analog world, digital surveillance is marked by invisibility, remoteness, networked pervasiveness, impartiality and new forms of consent to data collection (figure x). What does this mean?

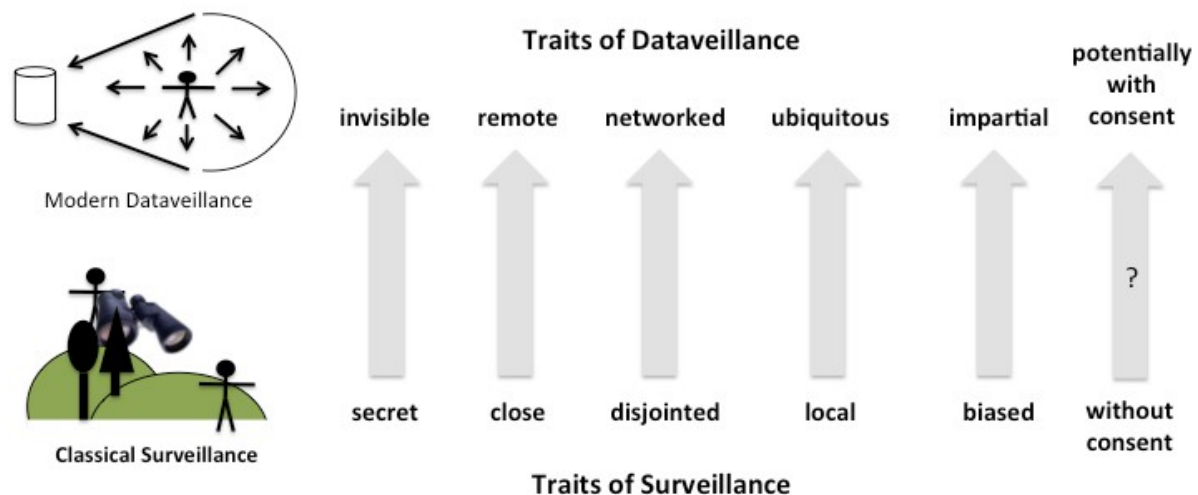


Figure x

Traditional surveillance strived for *invisibility*, but it probably achieved only a certain degree of secrecy that could still be uncovered. In contrast, technologies like ubiquitous sensor technology, cameras and mobile phones record without the knowledge of people that are not tech-savvy. There is invisibility instead of secrecy. Some efforts are made to put up warning signs for video cameras or RFID; it is sometimes still possible to spot some of the devices used. But most people today probably don't know about the vast amount of records collected about them through their digital devices. The old 'hunter-hunted game,' where the hunted could detect the attacker, flee and hide, or even play around with his attacker has gone; this liberty has now almost vanished. The attacker now is like a ghost, and it cannot be evaded.

Traditional surveillance was naturally physical. In contrast, today's digital surveillance, has removed the physical aspect. Surveillance only materializes remotely, potentially in some security monitoring room where unknown security folks stare at their displays. This *remoteness* constitutes an essential difference between the original kind of surveillance, which people emotionally rejected, and our modern form of surveillance. Humans can perceive the penetrating stare of another human being. But this natural survival instinct is lost in digital surveillance environments. As we don't feel the observer, we cannot really perceive a threat. We are not built for it. A lion does not recognize the presence of a huntsman approaching downwind.

The third difference between traditional and modern surveillance is its *pervasiveness*. In states like the former East Germany, where the government tried to spy on huge parts of its population, they could do so only by using human spies and their analog equipment. Observation was limited and imperfect, and activities could occur in unobserved niches. An

individual could argue that the observer had missed many aspects of his or her true life and convictions. Thoughts remained unobserved. Current and future forms of surveillance are different. Pervasive computing does not only comprise unlimited geographic coverage for all those places on earth where there is an Internet connection and peers into all of our objects. But it also penetrates and encloses our bodies. A Talos suit, smart bracelet or smart glasses as described in the scenarios measure every bodily process, every blink of attention or disinterest that our pupillary dilation reveals. This pervasiveness would not be too threatening if the technologies were isolated, serving only local purposes. But these technologies can be networked, and their information can be integrated to create a holistic view of individuals' lives. Figure x depicts an idea of the pervasive surveillance infrastructure and possible connections between different data collection entities. Figure x is not a complete or exact representation of the surveillance infrastructure as it is today or will be, but the image depicts the core of modern surveillance.

Finally, a major difference between classical surveillance and today's or future surveillance is the level of *impartiality* that can be both a benefit and a drawback for those observed. An Eastern German spy who was observing and reporting on his neighbor was probably not always impartial. If the spy disliked his neighbor, he might have selectively reported every little detail that suggested disobedience. Victims could be badly misrepresented. In contrast, machines are impartial; they don't care whether the person they observe is a spouse or a neighbor. They are not necessarily more objective (because their angle of observation is also one-dimensional), but they have no feelings and represent any case in the same impartial way. The drawback of this impartiality is that the machine lacks pity and mercy. This drawback will become more pronounced as we approach a technical level of sophistication where the observer is not even a human, sitting in some remote screening room, but a machine itself.

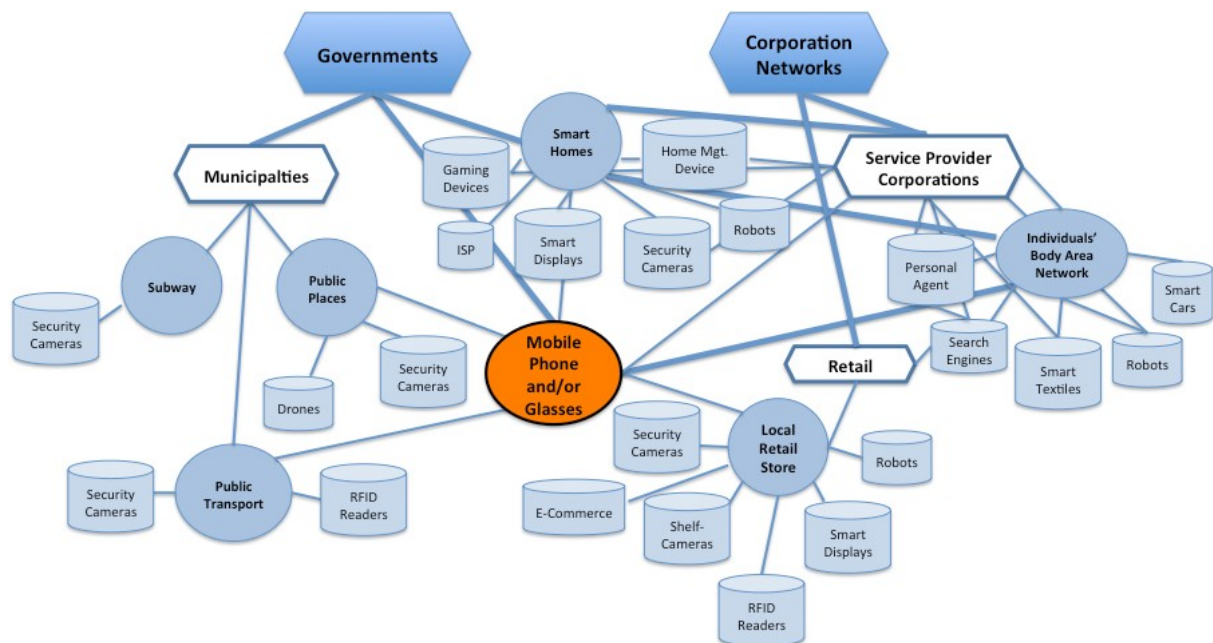


Figure x: Excerpt of existing and potential future networks of (a) data collecting services & devices (database symbol), (b) hubs integrating these devices/services (circles) as well as governmental or corporate recipients of the data (hexagon). Sources and networks could eventually be used for systematic electronic monitoring/surveillance (fat lines).

The Pros and Cons of Surveillance

Surveillance tends to be discussed as a threat to democratic society. While I personally share the opinion that surveillance is a great threat to societies, I also believe that ethically designed surveillance systems can impede some of these threats while creating local technological benefits. Take the example from my future stories of how a company could monitor its employees not only for unidirectional security reasons (as we have it today) but for the mutual benefit of companies and employees:

“Encrypted work activity logging was part of United’s work terms and conditions for employment. In fact, the integration of activity logging into work contracts in many companies was celebrated years ago as a major achievement of the labor unions. The encrypted activity logging process came as a response to a steep rise in burnout and workplace bullying, which seriously impacted companies’ productivity and damaged people’s health, mental stability and well-being. A compromise on the mode of surveillance was struck between unions and employers. Prior to these negotiations, employers had conducted video surveillance in a unidirectional way that undermined employee’s privacy while providing no benefits to them. As part of the new process, employee activities and conversations would be logged in all rooms as well as VR facilities and stored in an encrypted way under the full control of employees (in their personal data clouds). With this system no one, not even the CEO of the company, could view the original data. However, when a security incident happened, employees were informed and asked to share their data. In particular though when serious cases of burnout or bullying occurred, employees themselves could initiate a process of data analysis, handing over their secret key so that a designated representative could recover their data, text and voice streams and perform a conflict analysis. Data-mining technology would then look for patterns of behavior typical for mobbing or burnout as well as cognitive and emotional states. The streams could also be used to replay specific situations in which conflict had occurred. However, these replays would occur only in the presence of a trained coach or mediator. This practice had not only reduced bullying in recent years but also helped employees better understand their own communication patterns and behavior. Finally, the encrypted data was also used to extract aggregated ‘heat maps’ of the company’s general emotional state. This practice helped upper management to better grasp the true emotional “state of their corporate nation.”

Of course, many readers of this scenario may perceive this scenario as chilling. Should all work activities really be logged and potentially analyzed? Still, the scenario has benefits. The threatening traits of dataveillance are maintained. But thanks to the ethical data governance and human use of the technology, the threats are partially mitigated. In the scenario context, dataveillance is *open and transparent* to employees. Personal data is used only if employees *consent* and give their private keys to decrypt the data, self-initiating the use of the data at the individual level. *Human judgment* is integrated into the use of the data, which supports self-understanding. Video sequences are analyzed in the presence of a coach. Because the dataveillance infrastructure is ethically designed, it can not only provide extra security for the company but also reduce employees’ misbehavior at work. When things go wrong, as in the bullying case, people can learn about themselves and enter states of self-observation supported by technology. Technologies’ impartiality makes people see beyond what they *think* happened to what *really* happened. They can review their past behavior and try to use their insight to resolve conflicts. Machines could also use the vast amounts of behavioral data to detect patterns of behavior that help us to better understand behavior. Great learning at the individual, organizational and societal level may become possible, underpinning centuries of philosophical reasoning with hard facts.

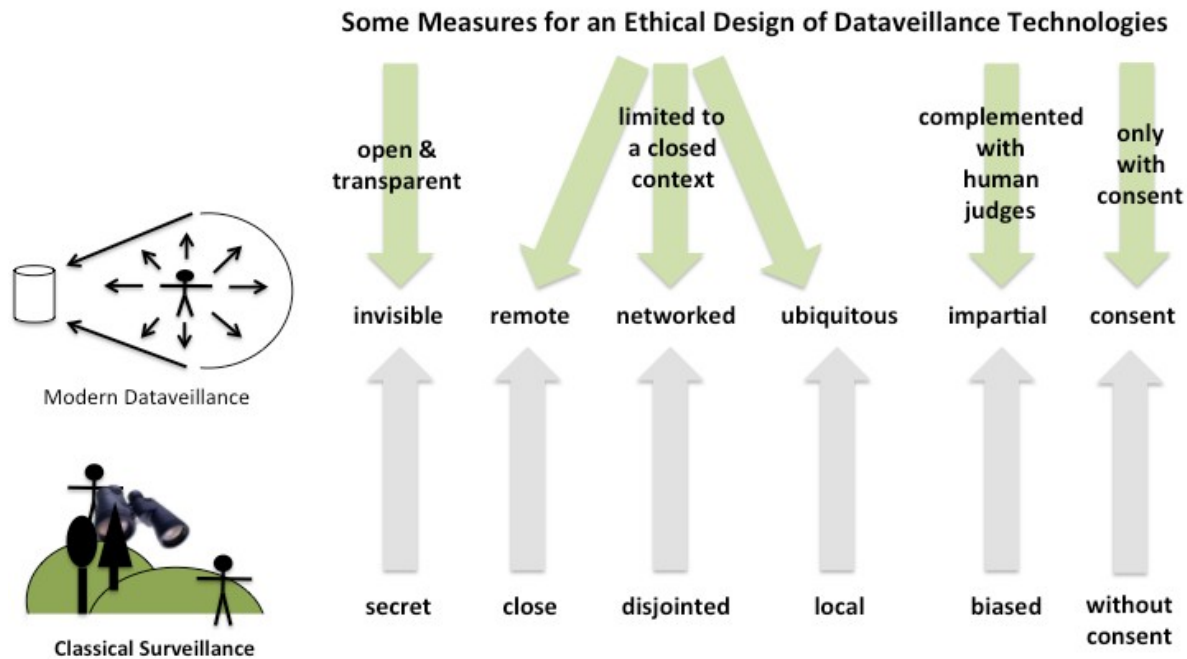


Figure x: Some measures can help to ease the negative effects of modern surveillance technologies

Note also that, in this scenario, surveillance is *technically limited to a closed context*. The decentralized isolation of surveillance facilities, which does not network them at a higher level, reflects what Helen Nissenbaum called the “contextual integrity” of data collection and use ([Nissenbaum 2004](#)) (see section x). Such contextual integrity through closed-context monitoring can balance people’s desire for safety with the potential abuses of networked pervasive surveillance. Of course, this model requires that the people who are observed trust key players in the surveillance network, in particular, the data hubs that collect and pool personal data (denoted as circles in figure x). Below I explain what trustworthiness means and how technology companies can build trust.

The learning benefits we can derive from the “Big Data” are complemented by another argument that is often brought forward, which is that people seem to appreciate surveillance. When surveillance is used, people feel safer in places that are traditionally unsafe but cannot be avoided, such as parking lots, underground stations or parks at night. Daniel Solove reports that Britain’s closed circuit television (CCTV)—a network of around 1,5 - 2 million public surveillance cameras—is widely perceived as “a friendly eye in the sky” (([Solove 2006](#)), p. 494)). Jeffrey Rosen reports on students’ reactions to body scanners at airports. He observes that quite a few welcome being naked for various reasons, ranging from security fears to wanting to demonstrate their “purity” ([Rosen 2005](#)). As described above, large-scale public polls also tend to suggest that people prefer ‘security’ to liberty. In his book “The Naked Crowd,” Jeffrey Rosen explains why people embrace surveillance: On a higher sociological level he argues, the “crowd’s unrealistic demand for a zero risk society is related to our anxieties about identity. Because we can no longer rely on traditional markers of status to decide whom to trust [i.e. cloths, family, religion, face-to-face meetings...], the crowd demands that individuals in the crowd prove their trustworthiness by exposing as much

personal information as possible" through the technologies that are set up. Rosen questions whether politics should be driven by such "feel-good" investments into surveillance technologies that appease ordinary people's sentiments. Crowds are vulnerable he says to systematic errors and biases in judgments; they are driven by "pseudo-events" in the press that make them misjudge the true risks in their daily lives. "Why should we care about the emotionalism of the Naked Crowd?" Rosen asks provocatively.

There are several reasons why many intellectuals criticize the build-up of a technological surveillance infrastructure. One is the potential for abuse. Aristotle warned us already that democracies have historically been replaced by totalitarian states (x). We know from our own history how the precise recordings of Jews' whereabouts in some European countries led to their systematic persecution during the Holocaust (see figure x). What would happen if a networked dataveillance infrastructure similar to the one depicted in figure x fell into the wrong hands? Would citizens in future societies constantly need to fear being watched and lose the right to free speech, as in George Orwell's "1984"?

The threat of official power abuse is only one reason for criticism. Another is that consciousness of surveillance leads to self-censorship and inhibition. Jeremy Bentham powerfully demonstrated this effect of a surveillance architecture. He designed and described a prison architecture, which he called the "Panopticon" (figure x): The Panopticon design allows guards to observe all inmates of a prison without the inmates being able to tell whether they are being watched or not. The inmates are in cells around a central circular tower structure. Although it is physically impossible for the guard to observe all cells at once, inmates cannot know when they are being watched, so they act as though they are being watched at all times. As a result, they constantly self-censor their behavior. Figure x shows a prison in Cuba that was built along the concept of the Panopticon.



Figure x: The **Presidio Modelo** was a "model prison" of Panopticon design, built on the

Island de la Juvental in Cuba (photograph taken in 2005 by Friman)

Why is self-censorship problematic? After all, some people argue that citizens and prisoners should behave well, that some self-censorship is good for society and that those who have nothing to hide don't feel followed either. However, remember the definition of positive liberty: People need to be able to make decisions of their own free will. Their behavior should be driven by their own desires and not by some external manipulative force. As we become more conscious of being watched, the motivation of our behavior may no longer come from ourselves. We may act well just because we are being watched. As a result, we degrade ourselves to the state of slaves in a surveillance machine. Edward Snowden³⁴ brought up this point when he wrote: "When we know we're being watched, we impose restraints on our behavior – even clearly innocent activities – just as surely as if we were ordered to do so. The mass surveillance systems of today, systems that pre-emptively automate the indiscriminate seizure of private records, constitute a sort of surveillance time-machine – a machine that simply cannot operate without violating our liberty on the broadest scale. And it permits governments to go back and scrutinize every decision you've ever made, every friend you've ever spoken to, and derive suspicion from an innocent life. Even a well-intentioned mistake can turn a life upside down" ([Snowden 2014](#)).

Reaching Golden Means in Mass-Surveillance?

Figure x shows that dataveillance is a bottom-up phenomenon: Each device or service operates more or less individually depending on the level of decentralization embedded in the technological design (see section x below). Devices and services then connect to data hubs to the extent required by the technical architecture. Devices, services and hubs can then be integrated into pervasive data sharing networks. The overall surveillance system operates like a hierarchal network in that it works only if the original data sources supply personal data. The sources of dataveillance can fuel mass surveillance if they are accessed and provide even more information when they are connected. This architecture puts tremendous responsibility on the design of each data source.

Each decentralized technical data source should therefore be built with privacy controls inside. Such efforts are recognized today on a political level and are called "privacy by design" ([Cavoukian 2011](#); [Spiekermann 2012](#)). Sections x to y below outline in detail what engineers can do to build systems with privacy inside. This practice can significantly influence the extent of dataveillance. But that said, governments and industry buyers of technology influence how technology is built. Their demand for technical features drives the type of technology that is supplied. The extent of *their* demand also determines the number of firms and the scale of production of surveillance technologies. Companies and governments therefore constantly need to determine the extent of data collection through their systems. For example, a gaming service might collect, store, analyze and share fine-grained emotional player data, as outlined in the gaming scenario. A university or corporation might monitor students or employees. Or governments might use surveillance cameras, drones or robots to monitor citizens. In all these cases, a small group of people makes an initial decision that then affects many others. How can this small group make surveillance decisions reasonably and wisely?

Jeffrey Rosen shows that "technologies and laws demanded by a fearful public often have no

³⁴ Edward Snowden is a whistleblower who in 2013 uncovered massive surveillance activities undertaken by the US National Security Agency and international secret service partners.

connection to the practical realities of the threats that we face" (Rosen, 2005 #1482). Leaders cannot respond numbly to public polls or the hungry sales efforts of security equipment firms. Instead, they need to make wise judgments on the extent of surveillance they want to support in their organization, consciously balancing security fears, privacy concerns and threats to liberty. There is no absolute answer on how to resolve this trade-off. All technological deployment decisions are unique. But the question of extent has recurred in ethical practice for millenia. It is asking for the "golden mean" (Aristotle) in our practices, "the middle way" (Buddhism) or what the Chinese consider their "doctrine of the mean" (Chinese: 中庸; pinyin: *zhōng yōng*). Box x makes a suggestion of what a Golden Mean Process for Surveillance could look like.

Box x

A Golden Mean Process for Deciding on the Extent of Surveillance

How can we find a golden mean in our surveillance practices? One way is for leaders to be courageous enough to publicize and transparently share their initial judgment on what the golden mean is. They should share how they arrived at their judgment based on reason and facts. The general public can then publically react to the initial judgment, potentially challenging it. This kind of public request for comments is abbreviated as "RFC." Requests for comments are very common in the technical world today. Wikipedia's dispute resolution system works on an RFC basis. The Internet Engineering Task Force (IETF), a principal technical development and standards-setting body for the Internet, also uses RFCs. The result of such a feedback process can be an adjustment or rebalancing of the initial judgment based on a more widely shared agreement of what constitutes a "golden mean." Simplistically speaking, transparency establishes checks, which then allow for balances of an initial judgment. Kant argued that "the sovereign should 'give his laws in such a way that they could have arisen from the united will of a whole people and to regard each subject, insofar as he wants to be a citizen, as if he has joined in voting for such a will'" (Kant, 1795 #1455). This process does not necessarily mean that the fears of the general public should determine the judgment. Experience has shown that transparent and identified commenting systems are frequented by experts more often than by the general public. As reasons and facts could be shared between decision-makers and a self-elected "polis" a good middle ground could be found and argued. Figure x illustrates what I call the "Golden Mean Process".

The Golden-Mean Process depends strongly on the wisdom inherent in the initial judgment. If the initial judgment is too extreme, then the polis responding to it will equally fall into extremes. This polarization of the problem space can easily create conflict where really compromise is required. For a wise initial proposal we need wise leaders that ideally possess what Aristotle would call the virtues of *sophrosyne* (temperance), *philotimia* (the right level of ambition) as well as the courage to publish and defend their opinion (*andreia*) ([Aristoteles](#)).

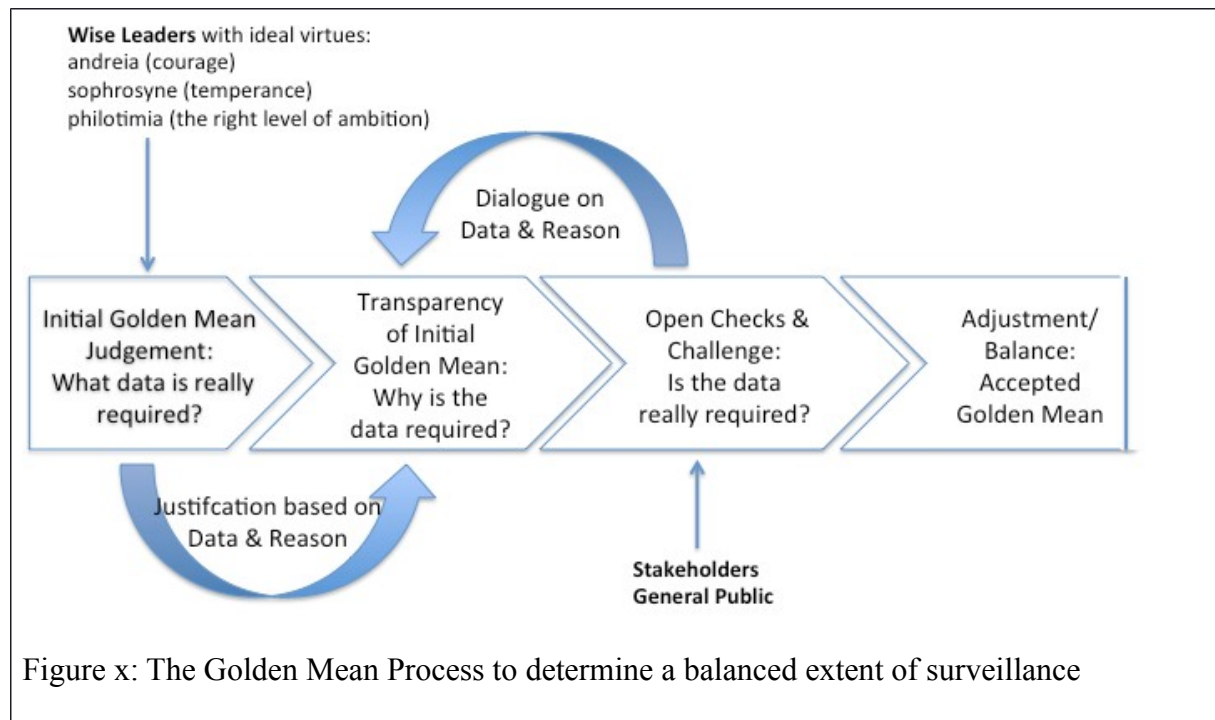


Figure x: The Golden Mean Process to determine a balanced extent of surveillance

Trust and Confidence in the Machine Age

*“Whatever matters to human beings,
trust is the atmosphere in which it thrives”
(Sissela Bok, 1978)*

In her book “Moral repair: Reconstructing moral relations after wrongdoing,” Margaret Urban Walker observes that humans interact with the help of “default trust.” We perceive “zones of default trust,” spaces and circumstances in which we can use trust as a shortcut when we decide to co-operate with others: “Sometimes when people refer to their ‘communities,’ either as networks of people or as geographical locations or both, they capture this sense of the place where one feels relatively safe. This is not because one believes one is utterly protected, but because one believes one knows what to expect and from whom to expect it, and one knows what is normal and what is out of place. One knows, in a word, what to expect and whom to trust. This practical outlook of ease, comfort, or complacency that relies on the good or tolerable behavior of others is the form of trust I call ‘default trust’” (p. 85).

As societies diffuse into postmodern, individualistic and mobile entities and integrate more technological artifacts into their relationships, traditional default zones of trust are changing and sometimes even eroding. Machines partially compensate for this loss of traditional trust structures. I have outlined above how the loss of traditional trust structures motivates surveillance. Social media tribalism may also compensate for some of the loss of trusting confirmation we received from physical peers in the past. But while machines seem to provide us with new forms of trust, their “nature” is an enigma to the common user. We know little about their trustworthiness. We still have to get to know the new machine species that is suddenly part of our ordinary human lives. And like any stranger in a new community, machines need to earn our trust.

Trust has long been an integral part of functioning social systems (Luhman, 19xx). Many people have a disposition to trust by character (Rotter, xxx). These preconditions are a valuable starting point for the machine age. Yet, trust is dangerous and can be betrayed. As machines move from engineers’ playgrounds and media analysts’ imaginations into the real world, we might encounter machines that don’t warrant our trust. Machines often don’t work the way we expect them to, nor do they necessarily work in our best interest and respect our expressed or implicit preferences. They may even turn against us at some point, as described in the robot nation scenario. Some scholars, reflecting on our developing technological environment, talk about an emerging “credibility crisis” (([Cohen 2012](#)), p. 1924). Of course, some believers in technology point to the extent to which we have already embraced technology, relying on it on a daily basis. However, as I will show, there is a difference between relying on machines because we have confidence in using them and really “trusting” them. Engineers must understand this difference and delve deeply into the concept of trust so that they can build machines that are deemed trustworthy.

What is trust?

In the scenarios about the future, many machines require trust in order to be embraced. Just think of the enormous trust that would be required for people to allow governments to have humanoid Alpha1 robots patrol the streets. People would need to trust that the robots would not wrongly hurt anyone and would be benevolent rather than dangerous. Enormous trust would also need to be placed in workplace systems. Agent Hal manages almost all operations at the robot manufacturer Future Lab. It drives manufacturing as well as part of the sales operations. It decides what to inform company employees and top management about. Finally, both United Games and the mall Halloville promise to be reliable when it comes to data handling practices. Both economic entities collect vast amounts of data about employees and customers. But United Games promises to analyze the data only with the consent of employees. And Halloville promises to respect some customers' desire to stay anonymous during their shopping trips. In all of these cases, people trust their computer systems and, indirectly, the service providers of the system. People trust that the Alpha1 systems are competent and that the robots will act benevolently in the citizens' interest. People trust in Hal's competence to judge Future Lab's operations and expect it to act predictably. Finally, they trust in the moral integrity or honesty of United Games and the Halloville mall to not abuse the systems. These examples show that the four most prevalent trusting beliefs in the trust literature, benevolence, competence, predictability and honesty (McKnight et al. 1996), are just as relevant in machine environments as in human environments. We psychologically transfer our trusting beliefs to machines (robots, agents), expecting human-like characteristics of trustworthiness from them (Reeves et al. 1996). We also place our trust in the providers of these machines, trusting them to not misuse or abuse the power of the machinery (figure x). In their work on trusting beliefs in e-commerce contexts, (Gefen et al. 2003) identified a fifth trust belief that is important for online business environments in particular: The absence of opportunism.

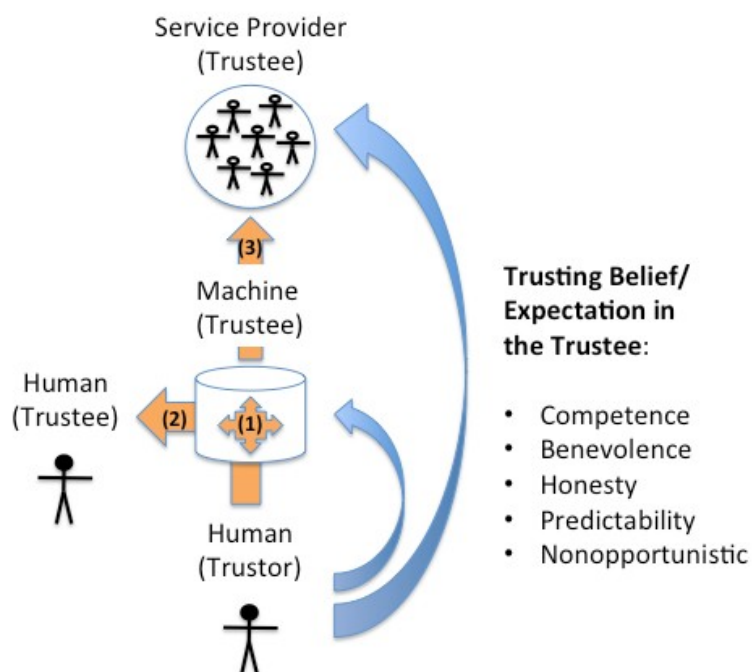


Figure x: A simplified representation of e-trust

Given these five trusting beliefs, we can understand the expectations that are inherent in the definition of trust. Niklas Luhmann (1988, 2000?) defines trust as a "willingness to behave based on expectations about the behavior of others when considering the risk involved" (p. x).

With respect to the machines and their providers, these expectations are competence, benevolence, honesty, predictability and lack of opportunism.

Luhmann’s definition hints at another important dimension of trust: the presence of risk. Trust and risk are in an unconditional positive relationship with each other. The more risk there is, the more we need to trust that things will work out well. Trust is required only when there is no further risk reduction possible and when the trustor is hence vulnerable. Vulnerability can be appeased if the machines or operators signal that they are competent enough to handle the risk. But beyond being competent, the trusted must *commit* to act well and in line with the expectations of the trustor. Here, we must consider commitment and motive.

Philosophers differ on how they treat the question of commitment needed in trust ([McLeod 2011](#)): For some philosophers, it is just important that a trustee signals his or her commitment. For others, the origins of commitment are vital. Commitment can be “calculative” when it is motivated by selfish interests or when people are engaged in a kind of “social contract,” based for instance on a public declaration or legal act. In contrast, commitment can also be based on the goodwill or moral integrity of a trusted entity. Commitment comes from the notion of *care* that the trusted has for the trustor. ([McLeod 2011](#)) argues that we can really only “trust” when commitment comes from care and moral integrity. Otherwise, she argues, we cannot trust but merrily *rely* on the other party for the time being: “The particular reason why care is central is that it allows us to distinguish between trust and mere reliance” (([McLeod 2011](#)), p. 5).

Differentiating trust from reliance is only one distinction philosophers have made to carve out the true nature of trust. Another distinction is between trust and confidence. Trust is an active *decision* by a trustor to delegate to a trustee some aspect of importance to achieve a goal ([Grodzinsky et al. 2011](#)). In fact, many instances where we rely on machines today don’t involve active decisions and are therefore not really expressions of trust. Wolter Pieters explains the difference between confidence and trust: “Confidence means self-assurance of the safety or security of a system without knowing the risks or considering alternatives. Trust means self-assurance by assessment of risks and alternatives...We have confidence in electricity supply, in people obeying traffic rules, etc. When there are different options possible, such as in choosing a bank for one’s savings, a comparison needs to be made, and trust takes the place of confidence” (([Pieters 2011](#)), p. 56). It is important to know about the difference between reliance, confidence and trust, because many would argue that people trust machines already and will continue to do so in the future. However, what they really observe is not trust but reliance, which may be supported by more or less confidence.

Finally, new forms of trust relationships will emerge in the machine age. Scholars talk about “e-trust,” which is “specifically developed in digital contexts and/or involving artificial agents” (([Taddeo et al. 2011](#)), p.1). E-trust is subject to the same underlying dynamics of trust that I described for our physical human world, but we interact with different entities, or at least perceive them differently. Figure x is a simplified representation of e-trust depicting the three entities people need to trust online: (1) machines, (2) other people whom they encounter through machines, and (3) providers of machines. Indirectly, people also need to trust the engineers who built the machines, because engineers are responsible for how the machine works. ([Grodzinsky et al. 2011](#)) define the forms of e-trust that are shown in figure x. Their summary of e-trust relationships is more complete than other summaries because it considers machine-to-machine communication. However, it does not contain a vital trustee: the providers of machines, who determine how machines are ultimately used and hence how trustworthy the machines are in terms of competence (determined to some extent by the financial investment made into them) and commitment (determined to some extent by the

moral attitude of the operator).

	Human-Human	Human-Agent	Agent-Human	Agent-Agent
Physical Encounter	HHP	HAP	AHP	AAP
Virtual Encounter	HHV	HAV	AHV	AAV

- HHP-trust: traditional notion of human “face-to-face” trust
- HHV-trust: humans trust each other, but mediated by electronic means
- HAP-trust: human trusts a physically present agent, e.g. a robot
- HAV-trust: human trusts a virtual agent, e.g. embodied interface agent
- AHP-trust: an artificial entity (i.e. a robot) trusts a human who is physically present
- AHV-trust: an artificial entity (i.e. a software programme) trusts a human who is virtually present
- AAP-trust: an artificial agent trusts another artificial agent in a physical encounter (i.e. two robots interacting)
- AAV-trust: an artificial agent trusts another artificial agent in a virtual encounter (i.e. two web bots)

Figure x: Forms of e-Trust derived from ([Grodzinsky et al. 2011](#))

Trust Mechanisms in Machines

Trust is one of those things in life that we cannot *want* or demand. We need to earn it and provide evidence that we are trustworthy. This evidence can be created in various ways. ([Pettit 2004](#)) distinguishes between evidence of face, evidence of file and evidence of frame. Evidence of face is really important when trust is built between humans in physical encounters; how someone says something, his or her body language and the (often involuntary) emotions expressed in people’s faces are key to building trust. At the moment, this kind of trust-building is still rare in human-machine interaction. But as machines become physical in the form of robots (see the HAP form of trust in figure x), this kind of evidence of face may gain importance. Engineers already work on integrating facial expression into robots that aim to create trust. An almost historic example for this kind of work is MIT’s robot KISMET, which reproduces emotions in the form of various facial expressions.³⁵

The evidence of file is the interaction we have with another person or entity over time. We consciously or unconsciously track the dynamics of being with others and either build up trust or become cautious. Unlike the evidence of face, evidence of file is harder to fake because the trustee needs to show consistent behavior over time to be trusted. Some scholars refer to this kind of trust as “knowledge-based trust” or “familiarity” ([Gefen et al. 2003](#)). I will discuss below how this kind of evidence can be created by using reputation systems.

³⁵ KISMET project: URL: <http://www.ai.mit.edu/projects/sociable/baby-bits.html> (last retrieved on August 15th 2014)

A third way of building trust is to provide evidence of frame. Evidence of frame is created by observing how a person or entity treats others or how others testify to the trustworthiness of the trustee. Again, reputation systems are a very valuable way to provide this kind of evidence. But when it comes to machines and their trustworthy functioning, seals and symbols can also confirm that the machine complies with certain standards of behavior and construction quality.

Thomas Simpson adds two further types of evidence to Pettit's list: evidence of context and evidence of identity ([Simpson 2011](#)). Evidence of context means that aspects in a situation can push the trustee to behave in a good way. These aspects can often be observed, as with a contract or public assurance. I have outlined how some philosophers consider this rather calculated "social contract" to be a form of trust building. "According to the calculative-based trust paradigm or evidence of context, trust can be shaped by rational assessments of the costs and benefits of another party cheating or cooperating in the relationship" (([Gefen et al. 2003](#)), p. 64). Although I share McLeod's view that we can only speak of reliance here, machine service providers can benefit by fostering this kind of calculus in users when they want users to rely on and continue to use machines. Public statements of guarantees, long warranties and strong regulation of a technology, accompanied by sanctions for misconduct, help to build people's calculative trust in machines.

([Pieters 2011](#)) outlines how "explanations-for-confidence" are particularly suited to provide evidence of context. The goal of these explanations is not to show people how a system functions in detail but to make them comfortable enough to use a system. In contrast, Pieters argues, we can identify "explanations-for-trust," which lay open how a system works. The goal here is to create transparency around a system (see section x). This transparency creates trust because it supports an active choice for using a particular system over another.

Simpson points to the importance of evidence of identity ([Simpson 2011](#)). A person's occupation, religious creed or way of living may be evidence of trust in respective situations where such characteristics become important. For example, people trust that a doctor can help when an accident happens. Transferring this form of evidence to a company context, some machine service providers have built up strong reputations for the performance of their machines. At the beginning evidence of file is necessary for brand building. But after an initial period of performance proof, the brand alone often inspires trust.

Finally, ([Gefen et al. 2003](#)) identifies "situation normality" and "ease of use" as factors that are particularly important in e-commerce contexts and hence potentially also important in more complex machine interactions. Situation normality seems to correspond to the predictability belief. In this view, "people tend to extend greater trust when the nature of the interaction is in accordance with what they consider to be typical and, thus, anticipated...In contrast with familiarity situational normality does not deal with knowledge about the actual vendor; rather, it deals with the extent that the interaction with that vendor is normal compared with similar sites" (p. 64). For machine design, this means that machines can build trust by using typical steps and forms of interaction that users recognize from comparable systems as well as information requests comparable to other systems. This observation hints at the importance of common design standards for multiple future systems. Common standards for system use are already widely known. For example, many diverse systems use the same symbols to signal on/off functionality. Reference architectures are used for application design. Both of these foster situation normality which again greatly increases the perceived ease of use of a system.

How computer scientists understand trust

Note that computer science students learn about trust in a slightly different manner. In computer science textbooks trustworthiness of systems is often discussed under the alternative term of “system dependability”. System dependability is seen as a non functional requirement. Ian Sommerville summarizes this construct as follows: “The dependability of a computer system is a property of the system that reflects its trustworthiness. Trustworthiness here essentially means the degree of confidence a user has that the system will operate as they expect, and that the system will not ‘fail’ in normal use (([Sommerville 2011](#)), p. 291)”.

Sommerville then specifies what dependability means, outlining that dependability (trustworthiness) depends on the security of a system, the safety of a system and its reliability. Both security and safety have been defined above. The system trait of *reliability* is similar to what scholars in the IS literature call “situation normality”. It means “the probability, over a given period of time, that the system will correctly deliver services as expected from the user (([Sommerville 2011](#)), p. 292)”. Note though that the definition of reliability for computer science readers is more precise than situation normality is for IS readers. Reliability is defined in terms of a ‘probability over time’ and hence viewed as a measurable system variable that the system can be tested for.

From this discrepancy of value definition we learn two things: The most important one is that the computer science perspective is at this moment much narrower than the general social perspective. While computer scientists learn to think of trust in terms of dependability, figure x makes plain that this is a very limited view of what makes a system trustworthy from a user’s perspective (or from the perspective of society at large). Dependability is just one form of system evidence. And even if it is the most important one (recognized by the bolted line around dependability in figure x), engineers are just as demanded when it comes to create emotional user interaction, to ensure transparency of the system and to continuously improve its ease of use. Engineers will also be involved in creating evidence of frame through certification of the system or ramping it up for a quality seal. That said, the second learning here is that engineers cannot be made responsible for creating trust in systems by themselves. General managers, such as product managers need to work on providing all the other forms of evidence required for trust. They need to think about warranties and guarantees they can give, nourish trust in the service brand as a whole and ensure appropriate media voice around the service. This again implies that managers need to be close to engineers and understand the system well enough to create the right ‘buzz’ around it. Too often systems can’t deliver on false promises made by managers. The result is not only a loss of face of the people involved in the project, but also a general damage to the brand.

Figure x summarizes the main trust-building mechanisms that are identified in the literature; each of them requires considerable investment in system design, certification and marketing.

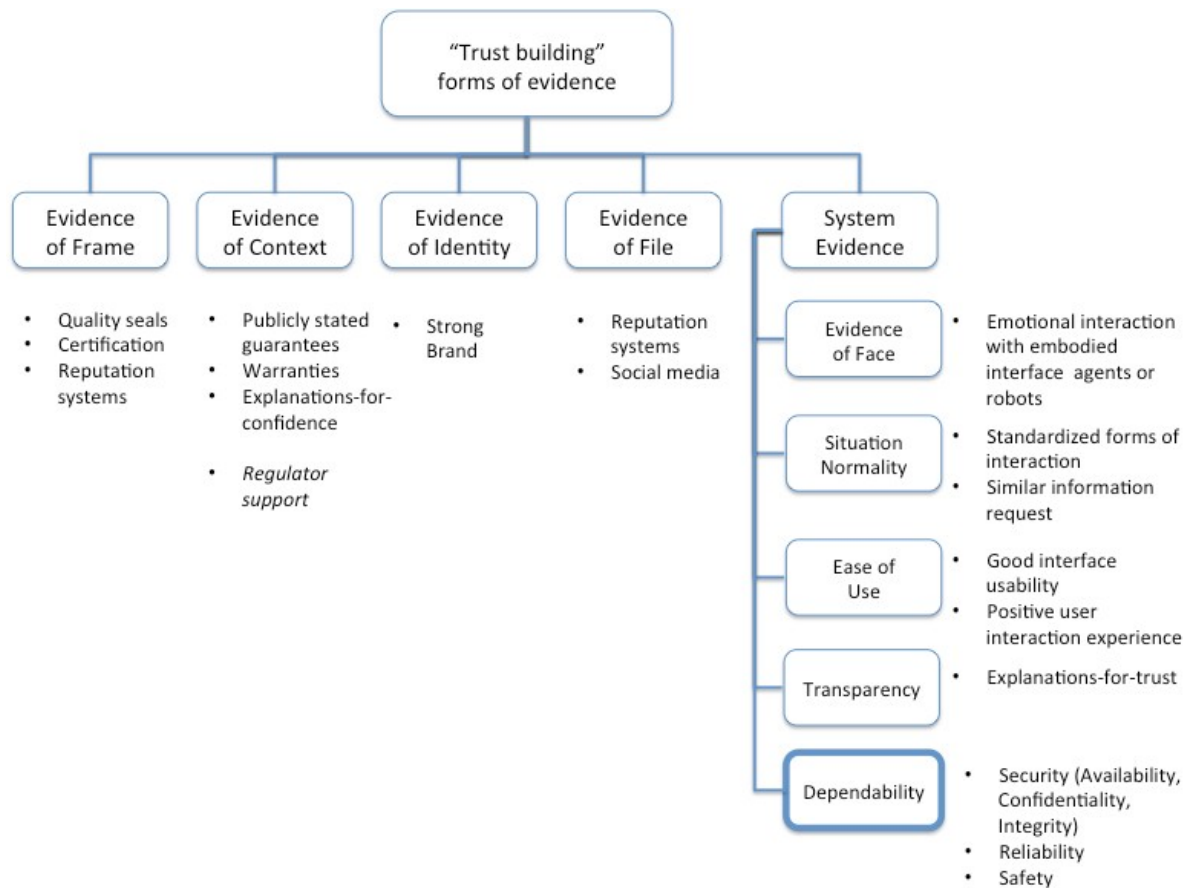


Figure x: Trust building through system design and system marketing

Reputation Systems

One of the most powerful ways to signal trustworthiness is to score well in reputation systems. A reputation system collects, distributes, and aggregates feedback about participants' past behavior or about goods and services ([Resnick 2000](#)). Typically, it does so within a community or domain, collecting opinions or ratings. A well-known example is TripAdvisor's history of comments for hotels or Amazon's star system for books. About x % of online purchase decisions today are made with the support of reputation systems. Reputation systems allow for evidence of file because they often contain a history of transactions with the trustee; they also allow for evidence of frame because other customers or independent assessors have also investigated the object.

Based on a critical discussion of existing reputation systems and the forms of trust evidence described previously, some measures can help to optimize the value of online reputation (see figure x as well as ([Simpson 2011](#))).

The most important characteristic of a reputation system is to ensure a high level of truthfulness. The entities that are rated in a reputation system have an incentive to look good. Therefore, manipulation of system results or even fraud is likely to occur. At the same time, social norms of politeness often impede people from leaving negative comments online. 99,1% of eBay's reputation system comments, for example, are positive. These behaviors can

easily undermine the value of a reputation system.

Operators of reputation systems can encourage truthfulness by offering monetary or other reward incentives to motivate high quality reviews, assuring users that their identity will not be shared with those that are being rated or encouraging loyalty to the community over politeness. Potentially, operators can also restrict reviewing to those who prove that they really used a service. Such an entitlement measure is problematic because it limits the number of reviewers. However, controlling who reviews also helps to prevent the entities that are being reviewed from writing reviews themselves. If reviews are not controlled, ratings can be inflated. Non-legal sanctions, such as a complete delisting of the person, good or service, are powerful ways of thwarting manipulation.

Reputation systems must also address flaming, which involves overly negative comments and scores on entities without true justification. There is tit-for-tat negative reviewing, for instance. ([Simpson 2011](#)) reports that sellers leave negative or neutral feedback on a buyer 61,7% of the time that the buyer leaves negative or neutral feedback on them. Such tit-for-tat behavior does not support truth building. Service providers can counter this kind of behavior by actively resolving conflicts through a mediator. Or, feedback can be published only after both parties have submitted it, without providing the ability to change comments afterwards.

Because the time contextualization of reputation scores mirrors how humans judge trustworthiness, it has produced more reliable reputation scores ([Novotny et al. 2014](#)). A challenge for digital reputation is that those with very strong reputations can afford to be untrustworthy on occasion, relying on the system to view sudden negative feedback as an outlier. A way to avoid this is to give recent feedback greater weight. Prioritization of recent feedback over old feedback also allows people to rebuild a reputation if they had a negative score in the past.

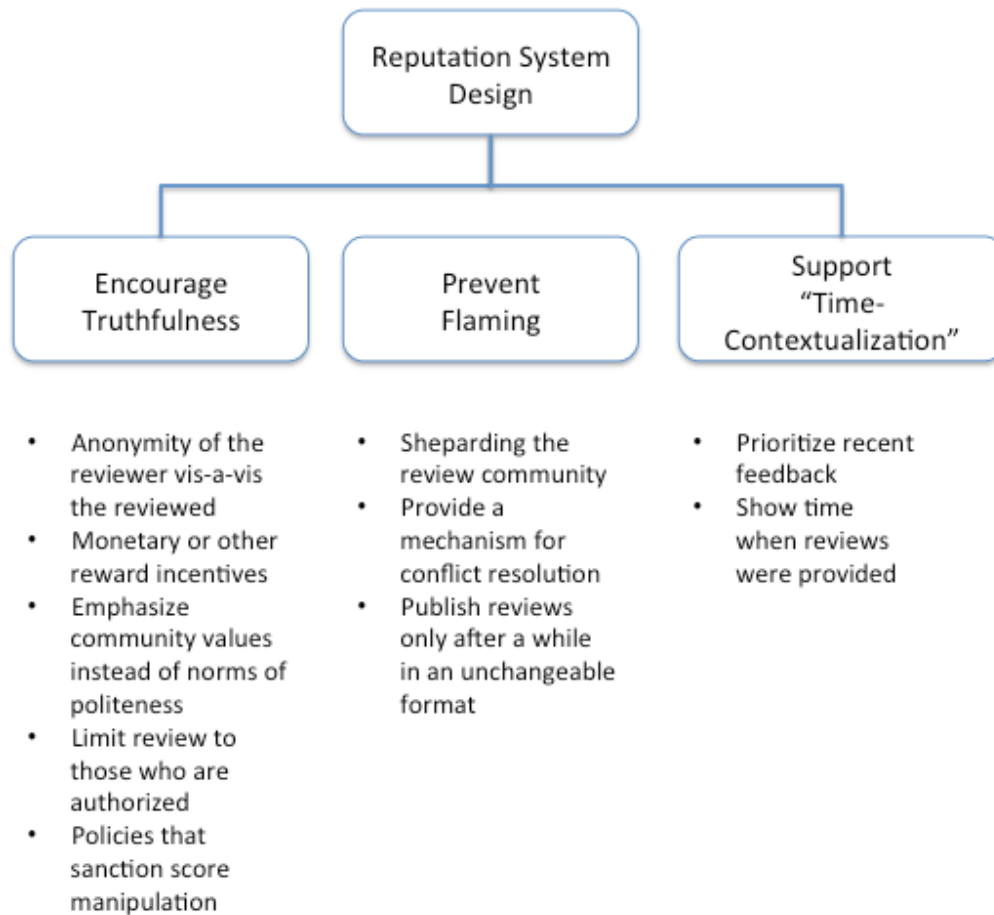


Figure x: Mechanisms for trustworthy reputation system design

Exercises

- Companies have various strategies to secure the privacy of their customer data. Explain them and reflect on them. What do you think makes more sense for a company: to anonymize or pseudonymize its data or to keep it identified and pursue strict policy management? Apply your thinking to the emotional profiles that United Games collects from its virtual reality players.
- Split the class into two teams. One team represents the investors or “the guards.” The other team represents the “polis.” Tell the teams about a recent case where a surveillance effort was made for economic or security reasons. The investors (or guards) then start a negotiation process. Based on reason and numbers, they propose how much surveillance is needed. The other team then scrutinizes and challenges this proposal. Afterwards, the two teams need to agree on the amount of surveillance that is appropriate. After the teams agree, compare their decision to what really happened.
- Reflect on the difference between reliance, confidence and trust, and find an example from your own use of technology where you rely vs. where you need to trust.
- Analyze the explanations that are given by the operators of a system you have confidence in and another system that you need to trust in. Compare the explanations that are given by the operators to gain your trust.

*Univ. Prof. Dr. Sarah Spiekermann; “The Human Use of Machine Beings”, Chapter 3,
Taylor and Francis, New York, 2015*

Belonging and Friendship in the Machine Age

The quality of our relationships is core to our well-being. As a result, it is not surprising that we’ve seen significant debate about whether new media destroy true friendship or, in contrast, enrich friendship by providing new means of communication. In 2010, 86% of respondents to a global study by Fujitsu agreed or strongly agreed that we are becoming socially isolated because all communication will be done with computers and not with people ([Fujitsu 2010](#)). I describe this trend in my stories where Agent Arthur becomes Sophia’s friend, Jeremy loves to be accompanied by a robot in the Halloville Mall and elderly people use robots in their households to help them (instead of family members coming along). In all of these scenarios original and authentic human encounters are replaced by the smooth and superficial surface of machines. Intellectuals who observe and study this trend are worried: “The idea of the ‘original’ is in crisis,” writes Sherry Turkle, a leading scholar in the field of human-robot interaction. She bewails that we are developing a “culture of simulation” in which “authenticity is for us what sex was to the Victorians: taboo and fascination, threat and preoccupation” (([Turkle 2011](#)), p. 74).

We must take this criticism serious. But we must also recognize that while authentic face-to-face communication is reduced, new digital media have created many new forms of constant connection between people ([Roberts et al. 2014](#)). Families and friends update each other constantly on location, news, arts and personal moments through presence apps and social media. Video telephony helps remote friends to stay close. Some people meet in virtual worlds. Scholars wonder though whether such short-term, feel-good connections come at the cost of true friendship. The *social augmentation hypothesis* states that this is not the case. The use of new media augments people’s total social resources, in particular, existing strong ties. From this perspective, digital media provides an additional avenue to be social with each other. People can co-ordinate their personal networks more easily via e-mail or messaging, and they can stay in tune with what happens to their friends.

In contrast, the *social displacement hypothesis* states that people who are more active online are less available for real-world engagements. And if they are, what quality do these offline relationships have when people constantly interrupt their face-to-face communication through the use of their smartphones? Does a person’s network size tell us anything about what’s really happening *within* those friendships? Compared to those who do not use the Internet, American Internet users are 42% more likely to visit a public park or plaza, and 45% are more likely to visit a coffee shop or café ([Hampton et al. 2009](#)). US bloggers are even 61% more likely to visit a public park than people who do not maintain a blog. But do Internet users and bloggers speak to anyone in these public hangout places? Or are they “alone together”, as Sherry Turkle has critically posed?

To better understand how machines influence human relationships, and to potentially build machines that support friendship, we need to better understand what the social construct of friendship really means.

What is Philia (friendship)?

Philosophers of all ages have identified three kinds of love: Agape (ἀγάπη, dilectio, caritas), philia (φιλία, amicitia) and eros (ἔρως, amor) ([Helm 2013](#); [Hoff 2013](#)). Agape is unconditional love of the kind people can have for God or for humankind in general. The word is often translated as "charity." This kind of love does not depend on any particular traits of the beloved. In contrast, eros and philia are both triggered by our responsiveness to others. Eros is a desire for someone, often sexual in nature. Philia is what's most associated with our term "friendship." It expresses itself as an affectionate regard or positive feeling towards another. At the same time, philia is not necessarily restricted to the term "friend" as we use it today. It also embraces people like family members or close colleagues; it is the broad kind of friendship we find on social network platforms. In this chapter, I concentrate on philia in terms of the strong and weak ties we may have with others.

Philia takes a central role in the creation of happiness and a fulfilled life. Aristotle regarded a good life as inherently social and believed that a social life was the soil in which people's virtues and good character root, receive nourishment and grow. Terrell Bynum wrote: "... Aristotle clearly saw [that] autonomy is not sufficient for flourishing, because human beings are fundamentally social and they cannot flourish on their own... Knowledge and science, wisdom and ethics, justice and law are all social achievements" (([Bynum 2006](#)), 160). Aristotle explicitly distinguished between two kinds of friendship: the imperfect friendships of utility and pleasure on one side, and the perfect friendship of virtue on the other. In the latter form of friendship, each participant altruistically wishes well for the other without considering their own personal utility or pleasure ([Munn 2012](#)). As part of this form of friendship, a participant might criticize their friend to help that friend understand his or her weaknesses. Virtuous friendship was important for Aristotle because he promoted "virtue ethics," a stream of philosophy that sees "the good" as something arising from people's habitual virtuous character rather than from a utilitarian calculus or joint pleasure only.

Shannon Vallor (2010, 2012) outlines how Aristotelian thinking is relevant for the analysis of friendship in online social media. She criticizes the narrow focus of traditional studies of social network platforms on feelings of happiness in terms of personal pleasure and utility only. Variables such as "life satisfaction," "self-esteem" and "social capital" have been at the forefront of investigation. And it seems like the "feel good" strategy of a social network platform like Facebook, which offers Like buttons but no Dislike buttons, caters to the more superficial dimensions of pleasurable community. However, while these "psychosocial goods" are important, they are not enough for friendship or a good life – at least not in the view of Aristotle. For him, friendship requires virtue. And to develop such a quality of character, we must have true friends with whom we can go beyond the "feel good" factor ([Vallor 2010](#); [Vallor 2012](#)).

But what are the characteristics of true and virtuous friendships? And how do we develop them? Across multiple works, Aristotle and other philosophers (([Aristoteles](#); [Helm 2013](#); [MacIntyre 1984](#); [Vallor 2012](#))) have identified various characteristics of friendship, in particular reciprocity, shared activity, the development of self-knowledge as well as empathy, and care and intimacy for and with the other (figure x).

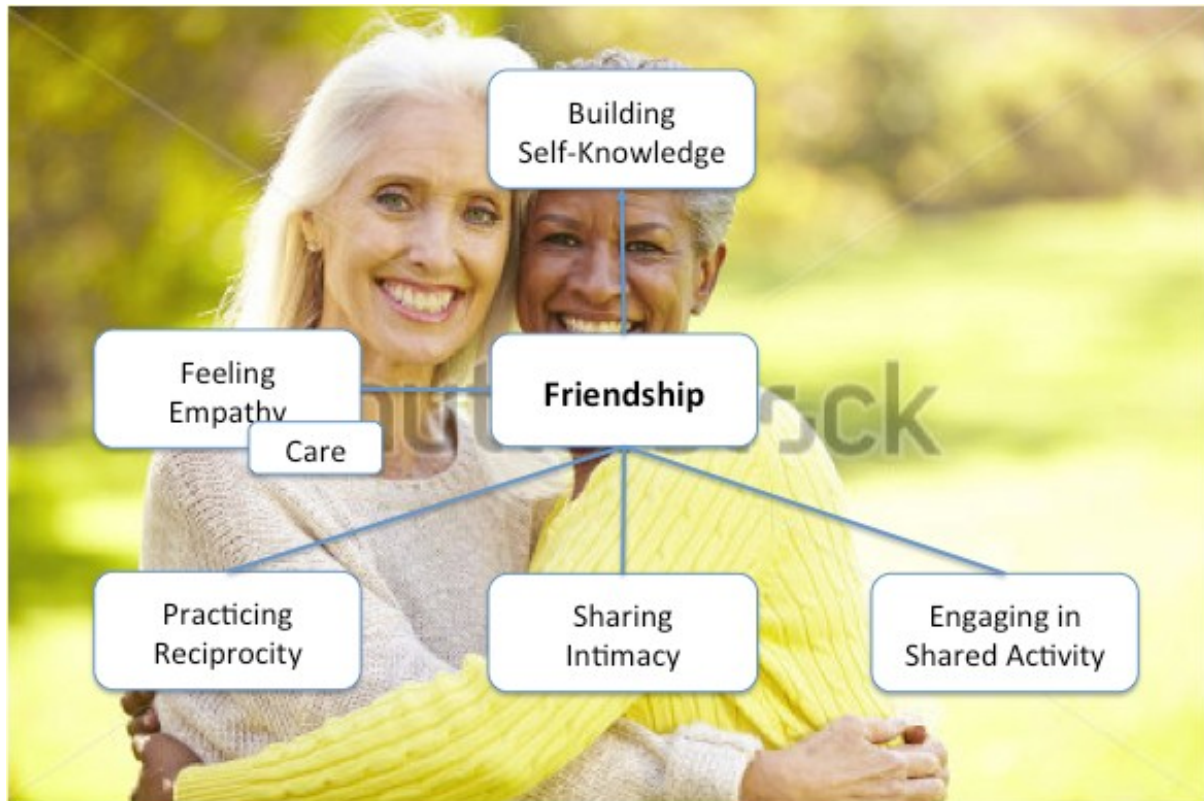


Figure x: Characteristics of a true and virtuous friendship

Reciprocity or “the reciprocal sharing of good[,] is the glue of all friendship” for Aristotle ([Aristoteles](#); [Vallor 2012](#)). It is the ability of people to give and to take. Giving and taking pleasure and utility is one way to cultivate reciprocity. For example, friends might exchange presents, provide help or support each other professionally. Most importantly, however, “complete” (*teleia*) friendship ([Aristoteles](#)) goes further than just creating pleasure and utility. It also involves the exchange of respect, love, knowledge and virtue. A good friend can give us an honest opinion or help us to understand something. It is such mutual feedback that helps us to correct ourselves and grow over time.

Reciprocity implies that friends spend time together. A shared life or *shared activities* are therefore an important driver of friendship. However, it is not only the time spent together that counts. Colleagues in a company also share time together, but are not always friends. Friendship manifests itself in shared activity where both friends enjoy the thing they do together and particularly enjoy doing this thing in the company of their friend ([Helm 2013](#); [Munn 2012](#)).

While spending time together and giving and taking from each other, friends can grow together and learn from each other. *Self-knowledge* is the result of such continuous learning. We gain knowledge about our own being in the world, a well-rounded and realistic understanding of the social world around us and about how we relate to the world and fit into it. Unlike most of today’s interpretations of self-knowledge – which involve digging deeply into our own selves, our childhood, etc. – Aristotle understood self-knowledge more as a matter of understanding our role in the world. He wrote, “we are not able to see what we are for ourselves” (*Magna Moralia*). He even wrote “if a human being surveys himself, we censure him as stupid” (MM). Instead, the “self-sufficing man will require friendship in order to know himself” (MM). Friends mirror each other’s behavior and thereby help each other to

develop. Some authors even argue that friends become each other’s “procreators” ([Millgram 1987](#)).

Just observing a friend’s behavior can create some self-knowledge. But *intimate* exchanges are equally important: with friends, we can share very private concerns and hope to get advice. Some scholars therefore view intimacy as a major pillar of friendship: a mutual self-disclosure or sharing of secrets that goes beyond the kind of conversations we would have with a colleague at work or some acquaintance ([Helm 2013](#)).

Finally, the genuine feeling of sympathy is highly important for friendship. Philosophers often distinguish between empathy and care when they write about the feelings underlying friendship. *Empathy* is a spontaneous emotive or perceptual capacity to feel with another person, to co-experience the joys and sufferings of the other person. “One grieves and rejoices with his friend,” ([Aristoteles](#)) wrote. But empathy is not a given. It depends on many tiny gestures and observations that people either appreciate or reject in each other. The “non-voluntary self-disclosures” that become apparent when people spend time together ([Cocking et al. 2000](#)) can breed empathy or separation.

Separate from empathy is the concept of care. While empathy is triggered by a friend’s situation, which we may pity or take joy in, care is unidirectional. We can “care” for a friend without him or her doing anything. ([Helm 2013](#)) discusses how care is similar to the unconditional love of “agape.” Care bestows value on a friend without any calculus.

I will now use these dimensions of friendship to explain how machines can influence friendship.

How can machines influence the various dimensions of friendship?

There are three broad ways in which friendship can be discussed in relation to machines. First, machines can influence how existing offline friendships are conducted. People use the communication functions of machines to stay in touch, plan activities, share ideas and develop new procedures more easily and hence more frequently. Second, machines can be used to form new bonds of friendship through shared avatar activity online. In virtual worlds strangers meet and spend time together. Sometimes these virtual friendships lead to offline relationships. And third, people may form friendly ties with artificial beings such as robots or the kind of virtual personal agent called Arthur in the scenarios. Figure x summarizes the three areas of artificial relations.

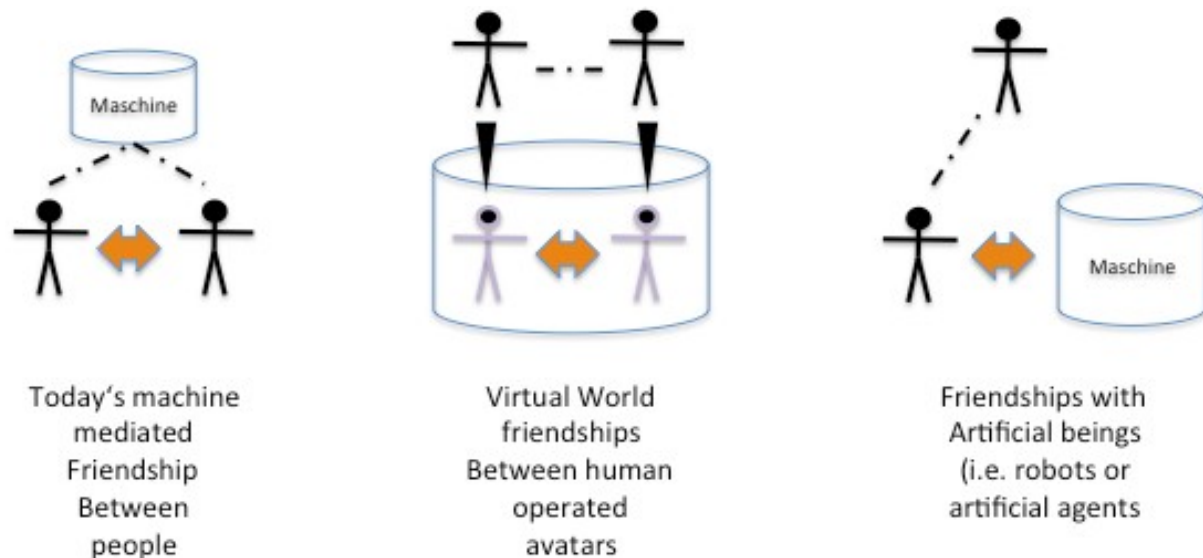


Figure x: Three ways in which machines influence friendship in the machine age

How do these forms of human-machine interaction alter humans' perceptions of belonging and friendship? Can machines be built to strengthen our friendships? Let's start with classic human-to-human friendships that are mediated via the Internet, e-mail, chat, social networks, and so on.

Human-to-human friendship in first generation media environments and social networks

In 2010, around 700 students from 10 different countries in North and South America, Europe, Asia and Africa participated in a study in which they were asked to spend one entire day offline, without digital media. After that day, they were asked to report on their experience ([Roberts et al. 2014](#)). "What does this unplugging reveal about being plugged in?" asked Jessica Roberts and Michael Koliska, the authors of the study. About half of the students were unable to complete the day and dropped out of the experiment. The major experience that all of the students reported about the day – including those who dropped out – was a perception of dependence and addiction. The majority of students who did not drop out still had a hard time staying offline. The feeling of dependence was accompanied by anxiety and distress. However, the third most common feeling was relief about being offline.

The findings from Roberts' and Koliska's study show how important digital connectedness has become for today's relationships, including friendships. A core reason for the dependence and distress that was felt by the participants was "a sense of having left an existing 'world', in which they feel everyone else lives, and that being outside this environment was challenging and difficult" ([Roberts et al. 2014](#)). Comments on the day were: "It was not an easy experience because I felt I was in kind of another world – left out" (student from Uganda); "...all I wanted to do was pick up my phone and become a part of the human race again" (student from the UK); "I felt isolated, without information and limited to the people around me" (student from Slovakia).

What enabling functions and applications can we use and develop to support the utility,

pleasure and virtue of friendships that are thus mediated? What services foster reciprocity, shared activity, accumulation of self-knowledge or the sharing of empathy and intimacy? Figure x presents a selection of current technical features and the dimensions of friendship they support.

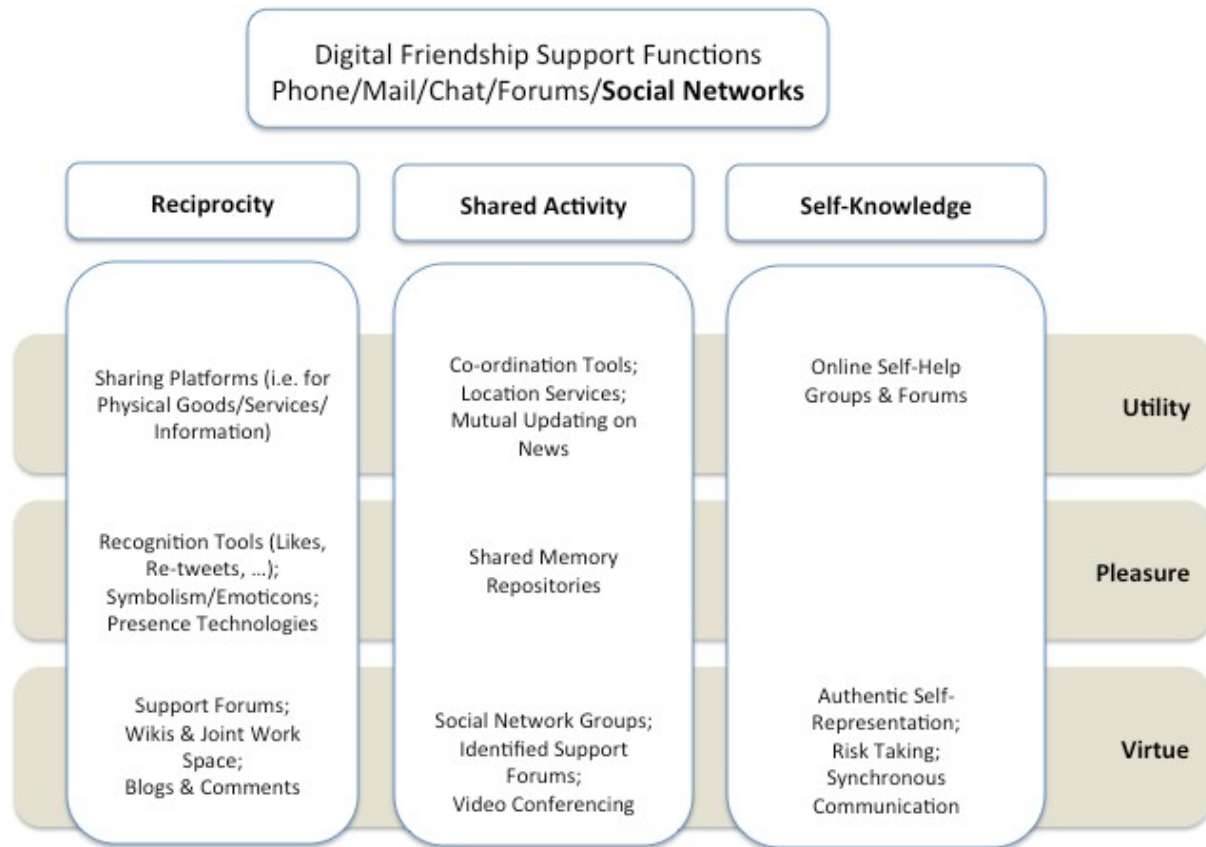


Figure x: How first generation machines enable various dimensions of friendship

Online *reciprocity* can support the creation of utility, pleasure and virtue. First-generation machines not only enable us to exchange pleasures by sending each other messages and emoticons, re-tweeting the other person, spreading a joint work, etc. Utility is also enhanced by the reciprocity inherent in the online medium. The Internet facilitates the sharing of resources, physical goods, services and information. Of course, virtuous friendship cannot be built exclusively online. But it can be supported by digital media: Ideas, thoughts or concerns can be exchanged in an in-depth form such as e-mail. Friends can exchange ideas by offering thoughtful feedback to a blog post or jointly working on an online project such as a forum or wiki. Online reciprocity has found new forms of symbolic language. Emoticons and small symbols are regularly used in playful digital exchanges. The smiley face symbol has reached such ubiquity that it triggers the same brain impulses as a real smiling face ([Churchesa et al. 2013](#)).

Machines support not only reciprocity but also shared activity. Mobile phones, chat and location services allow us to more easily co-ordinate and schedule offline activities. Social networks help to find new comrades-in-arms for offline matters, to update each other on developments and to easily share memories of joint activity. Videoconferences can maintain

long-distance friendships and facilitate time together that would physically be impossible. Special interest group forums, such as programming platforms, are a venue where users can give each other extensive help. In short, people can meet around and through online media.

Although reciprocity and shared activity flourish in online groups and forums, these forums are often anonymous. Friendship, in contrast, is built between identified individuals. When virtuous friendships and strong ties are the goal, then anonymity may not be the best path to take. Anonymity is widely heralded as an important ethical trait of online environments because it protects people's privacy, benefits free speech, and enhances people's deliberation. I have described how personal data can be anonymized, and I believe that anonymization is an effective measure to thwart mass surveillance. But when it comes to Web forums, news-portal comments and social networks, which by their nature refer to real world activities, identities or both, I question the benefit of anonymity or pseudonymity. In this context, anonymity undermines accountability and responsibility in communication.

Current digital services and social networks seem to support reciprocity and shared activity, but their ability to foster self-knowledge is questionable. Self-knowledge develops in our friendships through mutual observation, the exchange of honest feedback and joint experiences that we can learn from. But such immediate exchange is not available online. Many online media are asynchronous. And social networks, though built to foster friendship, primarily encourage one-to-many or many-to-many kinds of communication. This kind of "splintered mirror" communication ([Vallor 2012](#)) often dominates the richer one-to-one communication styles typical for friendships. If social networks wanted to support true friendships, they would need to support more synchronous communication for strong tie building. Features like messaging and video-conferencing are a good start. But it would be fruitful to think about even richer channels for 1:1 exchanges such as unique friendship spaces with joint digital goods like books or music files, video messaging, shared gaming resources, shared photos and experiences and private message repositories.

Another hurdle to building self-knowledge on social networks is that many people on these platforms engage in some kind of impression management. They reveal only selective content and build alternate identities that embellish or trivialize their real life. In a recent study on European Facebook, 60% of respondents said that they don't believe people present themselves how they really are. As a result, 'friends' receive feedback only on the shallow information objects they actually publicize. A holistic exchange is not possible. Nietzsche once famously pointed out that personal harvesting, which could be interpreted as building self-knowledge, is a matter of exposing oneself to risky endeavors: "Believe me" he wrote, "the secret of harvesting from existence the greatest fruitfulness and the greatest enjoyment is to live dangerously!" ([Nietzsche 1974](#)), p. 283). People on social networks normally don't take that risk. In contrast, they only show a polished façade.

Shared life and learning in virtual worlds

Much more risk – at least in fictitious form – is taken in virtual reality worlds. As of 2014, over 1 billion users were registered with virtual worlds³⁶, at least 500.000 of them being active inhabitants of the virtual world Second Life.³⁷ Top games such as World of Warcraft (WOW)

³⁶ http://readwrite.com/2010/10/01/number_of_virtual_world_users_breaks_the_1_billion

³⁷ <http://gigaom.com/2013/06/23/second-life-turns-10-what-it-did-wrong-and-why-it-will-have-its-own-second-life/> (URL last visited: September 1st, 2014).

or League of Legends attract around 1,9 million hours of play per year.³⁸ On average, players spend between 10 to 20 hours per week playing these games.³⁹ Considering these numbers is important, even from an academic perspective, because they document the extent to which friendships are now built and lived in digital worlds instead of the real world. People of all ages have joined these places to play, meet and socialize with existing offline friends or family or meet new people.⁴⁰

This rise of virtual worlds and their role in human relationships is highly controversial. Intellectuals often see virtual worlds as a threat to true human bonding, a dangerous displacement of the real by the virtual. In his book “On the Internet,” Hubert Dreyfus writes, “The temptation is to live in a world of stimulating images and simulated commitments and thus to lead a simulated life... the present age... transforms the task itself into an unreal feat of artifice, and reality into a theatre” ((Dreyfus 2009), p. 88). In contrast to such powerful critics, younger scholars who have spent a lot of time in virtual worlds themselves and have closely observed how people use these worlds are much more balanced about developments. “Gaming can be beneficial when it’s part of a healthy palette of social interactions,” writes Nick Yee, a Virtual Reality (VR) specialist. “Family members who play online games together report more family communication time and better communication quality...41% of online gamers felt that their game friendships – with people who they first met in online games – were comparable to or better than those with their real-life friends” ((Yee 2014), p. 36).

So who is right? An important source of academic research for understanding the true social dynamics of virtual worlds has been the US-based Daedalus project.⁴¹ In the past fifteen years Nick Yee, the initiator of this project, surveyed over 35.000 players of Massive Multiplayer Online Games (MMOGs). In his book “The Proteus Paradox” (2014), Yee summarizes his findings. The data reported hereafter on virtual worlds as well as the player comments cited are taken from this source.

Looking into the design and current use of virtual worlds, it becomes clear that they can help to build friendships along the dimensions introduced above. First, virtual worlds create a stimulating virtual place where people can meet for all kinds of adventurous and fantastic endeavors. Philosophers have recognized how important places such as the dinner table are for personal bonding (x). Virtual worlds can be a modern form of dinner table (except that only virtual food is served). In fact, 19% of virtual world visitors play with at least one family member. 80% know the people they meet there from the offline world. One-fourth of players regularly play games with a romantic partner.

Virtual “hang-out” places are created around the idea of “shared activity” in its truest sense. In many games, players need to form large persistent social groups, sometimes known as guilds, that help them to kill monsters and jointly survive and advance in the game. People in virtual

38 <http://metaversetribune.com/2013/05/15/why-have-virtual-worlds-declined/> (last visited URL on September 1st, 2014)

39 The estimate of 10 to 20 hours is based on Yee, N. 2014 *The Proteus Paradox* New Haven, Yale University Press., as well as the average hours played in Germany (on average 10 hours in virtual worlds according to: <http://de.globometer.com/spiele-deutschland.php>) and a 2012 statistic from <http://metaversetribune.com/2013/05/15/why-have-virtual-worlds-declined/>. The average player takes 372 hours (two full months of work) to reach the maximum level in the game World of Warcraft *ibid.*.

40 It is important to recognize that the widely held stereotype of virtual world players being mostly young teenage boys does not seem to be correct. According to *ibid.*, the average age of players in virtual worlds is 30. Only 20% of online gamers are teenage boys. Boys and girls equally enjoy playing in virtual worlds. Only immersive war focused games, such as World of Warcraft, have 80% male players.

41 The Daedalus Project: <http://www.nickye.com/daedalus/> (last visited URL: September 1st 2014)

worlds are therefore hardly ever idle ([Soraker 2012](#)). That said, the technical design of virtual worlds directly influences how much co-operation and reciprocity occurs in those worlds. In games like the original EverQuest, avatars could advance in the game only if they helped each other. Yee illustrates how EverQuest players experienced death in the game, relying on other players to help them revive their avatars: When EverQuest avatars were killed (for example in a monster raid), they were stripped naked and had to recover their body and equipment in a limited time frame. "To succeed in Everquest you need to form relationships with people you can trust. The game does a wonderful job of forcing people in this situation. RL (real life) rarely offers this opportunity as technological advances mean we have little reliance on others" (EverQuest player, male, 29)... the willingness to spend an hour to help a friend to retrieve a corpse isn't something that can be faked" ([Yee 2014](#)), p183).

The design around defeat and advancement in EverQuest forced people to bond and practice reciprocity. However, this aspect of design is not the only way to encourage friendship in these games. The need to share game resources can also make people join forces. In EverQuest, for example, players regularly needed to share spells. Yee further notes the importance of idle time and access to information about game function as ways to encourage shared activity and reciprocity. In older games, players regularly had to deal with downtime (mainly for technical reasons). While this was annoying, it also presented an opportunity for players to chat and get to know each other. In addition, because many games are complex to play, people spend considerable time working to understand how commands work and how to achieve specific ends in the game. This exchange of information about game function could be an inherent part of the in-game experience. A game could force players to ask each other for help rather than outsourcing this activity to a separate information interface (for example, the World of Warcraft Thottbot application). The concept of "RTFM" ("Read The Fucking Manual") can be replaced by in-game reciprocity.

An important dimension of friendship is the ability to learn from each other. As I've noted, social networks limit self-knowledge because people often share only the good parts of their lives, manipulating their platform image. In contrast, virtual worlds force people to be more real. Despite their artificial interface and use of avatars for representation, engaging in games in virtual worlds brings forth people's true character. As they game intensively, people forget about their masks and get to know each other largely as they really are. This aspect of gaming allows people to receive feedback on their behaviors. Research shows that character traits (such as the big five personality dimensions) are carried into the virtual world ([Yee et al. 2011](#)). 10 to 20 hours of intense play per week in a highly complex environment simply undermines people's ability to maintain a role.⁴² Consequently, virtual worlds see frequent non-voluntary self-disclosures, a quality that is recognized as vital for friendship formation. Furthermore, when people's real voices replace text-based communication or when people's real faces are morphed into an avatar (figure x), the experience of being with another 'real' person is even more vivid, regardless of whether the person appears in a virtual body.

The notion that people reveal many dimensions of their true personality in virtual worlds becomes evident in stories of how people fall in love in these worlds. About 10% of online gamers have dated someone they first met in a virtual world. Obviously, lovers don't report the experience of falling in love at first sight. But playing together and observing how the other person reacts over time creates a non-superficial way of meeting. "Virtual worlds can

⁴² Of course, the games are called "role-playing games". Players can select from a range of races (e.g. elves, trolls, humans) and classes (e.g. wizard, mage, cleric). But continued role-playing is actually a niche in those games. For example, only a handful of the hundreds of available servers on World of Warcraft are explicitly reserved for role-playing.

negate some of the superficial aspects of face-to-face relationships," writes Nick Yee (p. 134). He goes on to cite one of the players who fell in love: "On the outside we seem totally opposite. But we work so well on the inside. I guess that is what comes of meeting 'inside out' :p" (World of Warcraft, female, 25). "Inside-out" is the term that online gamers use to refer to this reversed model of forming relationships.

The roles people take in groups or in guilds can also contribute to their self-knowledge. For example, leading a guild can make people collect management experiences that prepare them for real-life situations. People from all age groups, continents, and backgrounds play together, and even though everyone turns up as their avatar, their different cultures and learning experiences are still present. "Slaying a dragon is actually quite straightforward once you've figured out how to manage a team of two dozen people to help you. And this is the crucial management problem that every successful guild leader must solve...Being a guild leader has taught me about personality types and how to manage people more than any job I've ever worked on" (World of Warcraft, female, 27).

An indirect way of building self-knowledge in virtual worlds is interaction with others in different roles and sexes. People can choose any gender and select avatars from a range of different races (e.g. elves, trolls, humans) and classes (e.g. wizard, mage, cleric). While the in-depth refinement of fictitious virtual personalities is a niche in virtual world games, simple gender bending or the maintenance of multiple avatars is common. Over half of male players, for instance, have at least one female avatar. Gender bending allows people to directly gain the experience of how it feels to be in the skin of the other sex. This experience can breed empathy and understanding. One male player confessed: "I'm amazed how thoughtless some people can be, how amazingly inept men are at flirting and starting a conversation with a female, and how it really does take more effort to be taken seriously as a female versus a male" (EverQuest, male, 24). Gender bending is not the only way to learn: Avatar appearance and size influence how people play their roles online, and people also transfer some of this virtual experience into the offline world ([Yee et al. 2009](#)). For example, tall and good-looking avatars keep less physical distance in virtual worlds. Being more confident encourages this behavior. If, let's say, a rather unattractive real person plays as this avatar, he or she might learn how it feels to be confident and transfer this feeling into real life.

Finally, virtual worlds allow us to observe ourselves, to look over our own shoulder in how we interact with others. In an extreme form, this kind of self-observation is regularly practiced by some male players. They create and maintain a second female character in the game whose role it is to watch the main male avatar play. In fact, the Daedalus project found that "by far the most widely adopted male explanation [for having a female character] is that the third-person perspective in these games means that players spend a great deal of time looking at the back of their character" ([Yee 2014](#)), p. 111).

Our current knowledge of virtual world games suggests that shared activity, reciprocity and self-knowledge development are to some extent present in virtual world environments. These characteristics of friendship can be supported by certain game designs, stories, dependencies and functionality. Figure x summarizes the enablers of virtual friendship creation. Another important dimension of friendship building and maintenance is the emotional part of friendships, which includes characteristics such as sympathy, empathy, and intimacy. The next section will delve in into this question.



Figure x: A selection of enablers for friendship building in virtual worlds

Empathy in virtual worlds

At the core of every friendship is an emotional attraction that expresses itself in sympathy and empathy. But can we have genuine sympathy for someone we meet in a virtual world in an avatar body? Social science research shows that, to some extent, human mating choices and judgments about attractiveness are artificially creatable and predictable. They are a function of how much someone looks like us ([Penton- Voak et al. 1999](#)) or our family members ([Bereczkei et al. 2009](#)). This human psychology can be used to artificially create sympathy in virtual worlds. It seems that if we are to like an avatar, it only has to adapt its artificial face to our own facial features or that of our parents. Studies in 2004 and 2006 showed how morphing of faces can influence election results: The faces of election candidates were morphed with voter's faces. Figure x shows what a male's and female's morphed faces look like when they are blended with George Bush or John Kerry. In the top row 40% of George Bush's facial features or morphed with another man's face. The bottom row illustrates the same effect for a 40% blend between a female and John Kerry.⁴³ Voters without strong political preferences were swayed by the influence of the morphed politician's face with their own. They did not recognize the manipulation and voted for the figure that looked like them ([Bailenson et al. 2008](#)).

⁴³ Taken from Wikipedia (URL last visited on August 30th 2014):
http://en.wikipedia.org/wiki/File:Candidate_morphs.jpg#mediaviewer/File:Candidate_morphs.jpg

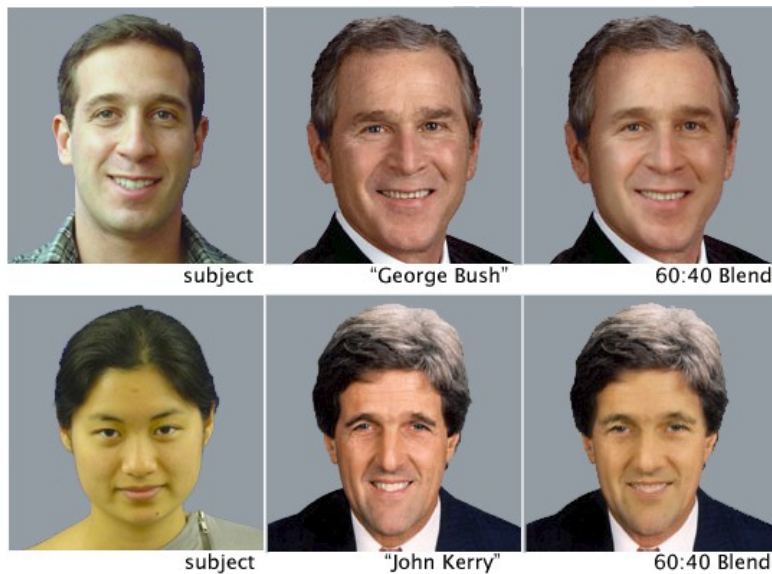


Figure x: An example of digital morphing of faces⁴⁴

While attractiveness can be artificially manipulated, creating empathy in virtual worlds seems to be more challenging. Empathy involves feeling with another person and sharing that person's happiness or grief. Empathy research shows that humans (and animals) perceive what is happening to peers, but as they do so they physically share in the experience of the peer. The peer's experience resonates in the body of the observer and initiates action in the observer, such as a desire to help ([Preston et al. 2002](#)). A part of this behavioral phenomenon of empathy seems to be related to humans' system of mirror neurons. Mirror neurons "mirror" in our body the behavior of someone we observe as though we were ourselves acting. Mirror neurons alone apparently don't produce empathy, but they provide cellular evidence for a shared representation of perception and action ([Jabbi et al. 2007](#); [Preston et al. 2002](#)). An open question is how mirror neurons function when we observe others in virtual worlds versus observing them in the real world. Experts currently think that mirror neurons work best in real life, when people are physically close. Virtual reality and videos are imperfect substitutes.⁴⁵ As a result, virtual reality would provide us with fewer experiences of empathy. A vital ingredient for friendship and important constituent of our inner emotional landscape suffers.

Mapping the bodily feelings of others onto our own internal body states is one of several indications for the importance of full bodily presence in high quality relationships. But this finding is not the only one that encourages full bodily presence. A study in the field of sociology observed how physical proximity affects the spread of happiness in groups. In a longitudinal study of social networks over 20 years, James Fowler and Nicholas Christakis found that happy friends who live closer to one's house promote personal happiness more than friends who live farther away. In fact, a friend who lives within a mile and who becomes happy increases one's probability to be equally happy by 25% ([Fowler et al. 2008](#)).

While these studies indicate the importance of physical presence some scholars don't believe in it. They argue for the supreme importance of humans' mental states and doubt the importance of bodily presence. An illustration of their argument is how we feel

⁴⁴ The example of morphed political candidates was created by Nick Yee and can be found at: http://en.wikipedia.org/wiki/Transformed_social_interaction (URL last visited on October 8th 2014)

⁴⁵ http://www.nytimes.com/2006/01/10/science/10mirr.html?pagewanted=all&_r=0

when we read a good novel or watch an emotional movie. Even though we are not physically with the characters we feel passionate solidarity with them. We cry and laugh while we read or watch. Good books and movies prove a deep connection between imagination and empathy as well. So who is right? For over 300 years, philosophers have debated the importance of our bodily presence compared to a pure mental presence. Ever since René Descartes formulated his famous sentence "I think, therefore I am," modernists have argued that our human essence resides in our brains. But this belief is not universally shared. Box x gives a short overview of the philosophical tension over the importance of body and mind. Our understanding of body-mind unity or independence will ultimately determine our conclusions on the relative value of virtual friendship.

BOX X:
**The body versus mind discourse in philosophy and
its implication for digital friendship**

There are two camps in philosophy that disagree about the importance of the human body. On one side, transhumanists such as Ray Kurzweil¹ and Hans Moravec² believe that the body is not important. Like Descartes, they believe in a body-mind separation.

In contrast, philosophers like Nietzsche⁴, Merleau-Ponty⁵, and Hubert Dreyfus⁶ and spiritual thinkers in Asia (in particular, those who practice Yoga) believe that the most important resource of human beings is not their mental capability but the emotional and intuitive capacity of their bodies. Nietzsche's Zarathustra says: “‘I’, you say, and are proud of the word. But greater is that in which you do not wish to have faith – your body and its great reason: that does not *say* ‘I’, but *does* ‘I’...Behind your thoughts and feelings, my brother, there stands a mighty ruler, an unknown sage – whose name is self. In your body he dwells; he is your body”⁴ (Nietzsche 1883–1885), p. 34).

Lets see what this means for our understanding of machines and the roles they can take in our lives: In the thinking of Descartes and today's transhumanists, the sense organs are transducers that bring information to the brain. Descartes drew on the phenomenon of amputated people who sometimes insist that they feel pain in a limb that is not there. This observation led him to believe that everything we experience is a creation of our own minds and that the world and our bodies are not directly present. The end vision of this thinking can be understood by watching the film “The Matrix” (1999), where human beings spend their lives in tubes, with their brains connected to machines that simulate life for them. They believe that they live, but in reality they spend their true lives in a tube. Transhumanists like Ray Kurzweil might not find this fictional scenario to be too far-fetched. They argue that we could scan our brains we can understand how humans work, upload the essence of human “intelligence” to a computer and then live in a machine after our physical death: “Uploading a human brain means scanning all of its salient details and then reinstantiating those details into a suitable powerful computational substrate. This process would capture a person's entire personality, memory, skills, and history.”¹ Embracing this view of human existence obviously bears great promise for friendship in virtual reality: It would mean that everything that we experience in the real world can also be experienced in virtual worlds. It means that our bodies are ultimately not important.



Figure x: With our bodies we “commune” with reality and get “a grip of the world”

Philosophers have recently questioned this idea of humanity, taking analogical phenomenology as a more holistic scientific approach to understand human existence.⁷ Maurice Merleau-Ponty (1908 – 1961), for example, stressed the importance of using our bodies to make sense of the world. He described how our bodies constantly “commune with” the objects around us: a jazz player communes with his saxophone, and a cook communes with his soup (figure x). Our body movements help us to zoom in and out of the world, to approach it from the right distance, and in doing so help us to achieve our unique “grip of the world.”⁵ From this perspective, our body is not just a collection of sensors that channel bits of information to the brain; consciousness is part of the body itself.

Take another example: when we enter a room where a party is happening, we sense the mood in the room. We perceive this mood through more than just our eyes; if we used a surveillance camera to review the scene, we would not necessarily be able to see the mood. Neither can we really smell or hear the mood. Still, we know through our bodily senses whether the party is in full swing, and we can physically share in this mood. In his *Phenomenology of Perception*, Merleau-Ponty wrote, “Insofar as I have hands, feet; a body, I sustain around me intentions which are not dependent on my decisions and which affect my surroundings in a way that I do not choose” (1962, p. 440).⁵ The described neuroscience research supports this view. Vittorio Gallese, who coined the term “mirror neurons” in his seminal 1996 article “Action Recognition in the Premotor Cortex,” writes about our “embodied experience of the world”⁸ (2004, p. 180). He says: “We map the actions of others onto our own motor system...creating a mutual resonance of intentionally meaningful sensory-motor behaviors, but not specific mental state interference.”⁹

If we think about online friendship from this philosophical perspective, true friendship is not possible in virtual worlds at the same level of quality and intensity as physical friendship. Being together online is devoid of the holistic experience we have when our bodies are present. The idea of transhumanists of simply uploading human existence by scanning our brain activities is therefore highly naïve. “I shall not go your way, O despisers of the body! You are no bridge to the overman!” concluded Nietzsche in his Zarathustra.⁴

References

- 1) KURZWEIL, Ray. *The Singularity is Near- When Humans Transcend Biology*, London, Penguin Group, London, 2006
- 2) MORAVEC, Hans. *Mind Children - The Future of Robot and Human Intelligence*, Cambridge, USA, Harvard University Press, 1988
- 3) DESCARTES, R.. *Principles of Philosophy*, Dordrecht, The Netherlands, Kluwer Academic Publishers, 1991
- 4) Nietzsche, F. 1883–1885 *Also sprach Zarathustra - Ein Buch für Alle und Keinen*, (2010 ed.) München, C.H.Beck.
- 5) Merleau-Pointy ...
- 6) DREYFUS, H. L. *On the Internet*, New York, Routledge, 2009
- 7) HOFF, J. 2013. *The Analogical Turn: Rethinking Modernity with Nicholas of Cusa*, Cambridge, UK, William B. Eerdmans Publishing Company.
- 8) GALLESE, V., FADIGA, L., FOGASSI, L. & RIZZOLATTI, G. 1996. Action Recognition in the Premotor Cortex. *Brain*, 119, 593-609 (p. 180)
- 9) JENSON, D. & IACOBONI, M. 2011. Literary Biomimesis: Mirror Neurons and the Ontological Priority of Representation. *California Italian Studies*, 2.

Intimacy and Disinhibition in Online Environments

Many observations have been made on how intimate people get when their communication is digitally mediated. ([Walther 1996](#)) called this phenomenon “hyperpersonal interaction.” The “boundary regulation process” ([Altman 1975](#)) that people use to manage their privacy vis-à-vis others therefore seems to follow different dynamics online than for face-to-face encounters. Generally, people open up more when their communication is digitally mediated, a phenomenon that has been called the “online disinhibition effect” ([Suler 2004](#)). However, we must distinguish between two kinds of digital encounters: One is communication with other people online at various degrees of anonymity or identification, for example, on web forums or on social network platforms. The other is talking to a machine, be it a virtual interface agent, such as Agent Arthur, or a robot. These distinction is visualized in figure x.

Let’s start with human-to-human communication that is digitally mediated. On many news portals and forums where commentators’ identities are protected by anonymity, toxic disinhibition is a common phenomenon. Many people – often called “trolls” - use aggressive and demeaning language to express negative feelings. Conversely, there can be benign disinhibition: people being exceptionally kind, generous and enthusiastic towards others. Overall, people are more frank online.

One reason people open up is that many online platforms promise anonymity. Anonymity in relation to communication partners makes the senders of information less vulnerable, particularly when the senders trust that their online actions are totally separate from their real offline selves. As Johan Suler writes on the effects of dissociative anonymity: “...the person can avert responsibility for those behaviors, almost as if superego restrictions and moral cognitive processes have been temporarily suspended from the online psyche” ([Suler 2004](#)), p. 322). Anonymity is, of course, not a basis for real friendship, which requires that people reveal their identities at some point. So general information forums on the Internet are not the place to be intimate with each other. But in virtual worlds, people often transition from initial anonymity to identification. They reveal who they are to a select group of others. What has

been said anonymously then becomes important.

Even when identities are revealed, online mediation seems to encourage more open communication. In the Daedalus project, 24% of virtual world players said that they told personal issues or secrets to their online friends that they had never told their offline friends ([Yee 2014](#)). Scholars believe that players reveal more in virtual worlds because they are invisible. When avatars speak to each other or people chat, nonverbal cues are filtered out. People don't have to worry about how they look or sound and, most importantly, they also don't see the other person's reaction to what they say. In traditional psychoanalytic theory, the therapist sits behind the patient for the same reason; people open up more when they don't see whom they are speaking to. One could argue that friendship benefits from the way that invisibility fosters intimacy in the same way as a psychoanalytic theory is beneficial for the engaging in real friendship without replacing it. Secrets may be shared more openly. Yet, visible reactions and the sound of a friend's voice additionally create reciprocity and support the building of self-knowledge. By seeing how our friends react to what we tell them, we learn about ourselves. So, taken together, the technical reality of cue-free communication creates a contradictory effect: It increases intimacy, but reduces reciprocity and learning.

Another reason for online communication being more genuine is that people are initially equalized and less prejudiced. All real-world signs of status, wealth, race, and so on are largely leveled out. In virtual worlds, everyone can look however they want to look. On social networks, where people know each other from offline encounters, they can and do post rather favorable images of themselves and their lives. In a way, the nature of the digital medium itself makes the world flat: The enforced two-dimensionality of the screen brings everyone symbolically to the same level. This representational and positive equality lowers communication barriers that exist in the real world. Furthermore, it encourages people to recognize skills that are often suppressed by inequality biases in the real world: such skills include writing skills, humor, the quality of one's ideas, technical know-how, and so on. This shifting of relevant skills makes some people open up more than they would offline. For example, shy people and people with physical handicaps participate more online than they do offline (Quelle? X). In the Daedalus project, people said that they open up better in the virtual world and therefore learn more about their lovers there than in the real world.

Some authors have argued that it is not equality that makes people open up in virtual worlds but rather the *distance* of mediated friendships ([Briggle 2008](#)). Less courage is required to be candid in an environment where you can execute an “emotional hit and run” by just closing your computer, knowing that the other person is probably far away ([Suler 2004](#)). This argument is anecdotally supported by a phenomenon we observe when strangers on a plane reveal intimate information to each other, expecting that the information they share will not catch up with them later. Could interfaces foster honesty and intimacy by indicating the real physical distance between parties?

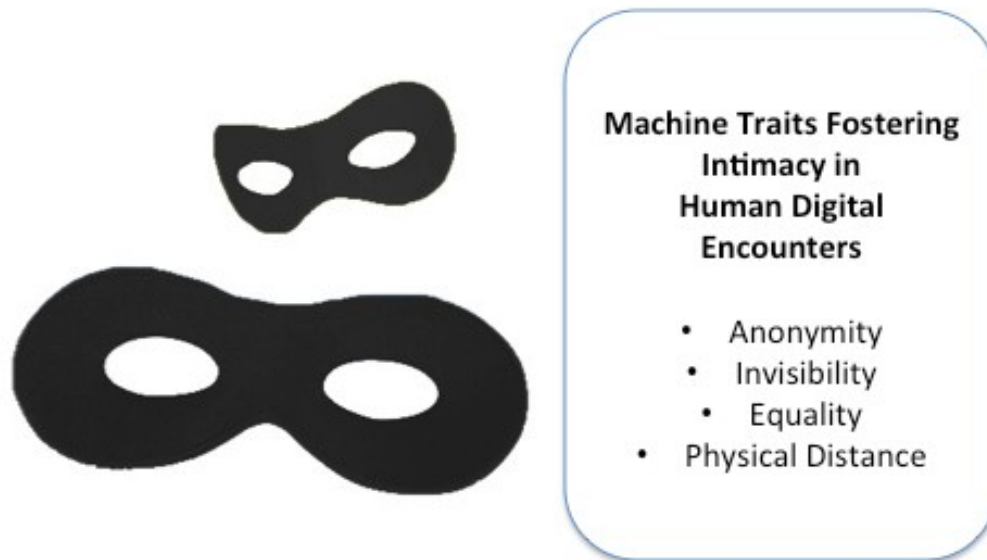


Figure x: Machine traits that foster online intimacy

Intimacy with Artificial Beings

A special form of becoming intimate in the machine age is when we share information directly with an artificial entity instead of another human being. Conversations with artificial agents (such as Agent Arthur in my stories), with preprogrammed figures in virtual worlds or with robots are examples of this kind of interaction.

Over and over again, research has shown that people get exceptionally intimate with artificial beings. This finding was demonstrated for the first time in Robert Weizenbaum's ELIZA experiments at MIT in the 1960s ([Weizenbaum 1977](#)). ELIZA was a computer program that employed an early form of natural language processing and simple pattern matching. People would type in sentences like "My head hurts" and ELIZA would respond with something like "Why do you say your head hurts?" Many people who used the computer system started to really like it and extensively share personal information with it. In fact, Weizenbaum described how his secretary got hold of the program and got so fond of it that she started to talk to it on a regular basis, sharing extensive parts of her private life.

Scholars have investigated why people disclose so much to computers and how this disclosure can be manipulated ([Moon 2000](#); [Reeves et al. 1996](#)). Based on such research, the theory of social response postulates that humans treat machines in the same way as other human beings even when they know that the machines do not possess feelings or "selves" ([Reeves et al. 1996](#)). Scholars explain this behavior by noting that humans evolved as social beings and apply their learned heuristics to machines; scholars think that people are "mindless" in a way, failing to reflect on the difference between other humans and machines ([Nass et al. 2000](#)).

Against the background of social response theory, Youngme Moon investigated how conversational interface strategies can be used to make people self-disclose. One of these strategies is to mimic the reciprocity of self-disclosure. From human interactions, it is known that disclosure begets disclosure. People who receive information from others feel obliged to share something about themselves ([Derlega et al. 1993](#)). Disclosure is also much more likely to occur if requests for information gradually escalate. Relationships proceed from casual

exchanges to increasingly intimate ones over time ([Altman et al. 1973](#)). Youngme Moon found that when these strategies are implemented in machines such as conversational agents, people disclose more and more intimate information. Figure x shows one of the manipulations that Moon used in her experiments. This experimental example shows how easily humans’ intimate disclosure can be manipulated through machine interaction strategies. Sherry Turkle comments that, faced with relational agents, people fantasize about a “mutual connection” ([Turkle et al. 2006](#)).

Non-disclosing Machine	Disclosing Machine
<p>Machine says: “What do you dislike about your physical appearance?”</p> <p>Participant answers: “I could lose some pudginess and gain more tone, which requires effort.”</p>	<p>Machine says: “You may have noticed that this computer looks just like most other PCs on campus. In fact, 90% of all computers are beige, so this computer is not very distinctive in its appearance. What do you dislike about your physical appearance?”</p> <p>Participant answers: “I hate my big hips. I’m a sugar freak, and all that sugar sits on my hips. I also don’t like that I have relatively small breasts, but that is nothing compared to the way the size of my hips bothers me. No amount of running or lifting or anything else seems to slim them.”</p>

Figure x: Varying machine strategies to make people disclose, based on ([Moon 2000](#)), Figure 1, p. 330

In her studies on human-robot interaction Sherry Turkle found another dimension that seems important. She quotes an elderly man, Jonathan (74), who lives in a nursing home and has used the “My Real Baby Robot” for a while. “The robot wouldn’t criticize me,” says the old man. From a reaction like this one, we might speculate whether interaction with artificial agents is simply easier, less chaotic or less entropic for people. Humans like to be inert to a certain extent, and so sharing intimate information with a lifeless artifact seems easiest. Unlike a human, an artificial agent is unlikely to object or cause turbulence. It seems as if people like to keep moving in a straight line and at constant velocity, just as Newton observed for all physical bodies in his first law in *Philosophiae Naturalis Principia Mathematica* (p. 72): “The *vis insita*, or innate force of matter, is a power of resisting by which every body, as much as in it lies, endeavors to preserve its present state, whether it be of rest or of moving

uniformly forward in a straight line." "Do not disturb my circles!" said Archimedes (287 – 212 BC), introducing a saying that we often use to express this very desire to be undisturbed. More research is needed to investigate this relationship between human inertia and the pleasure people take in exchanging information with machines.

Finally, authors have speculated that intimacy or online disinhibition could be a result of "solipsistic introjection" ([Suler 2004](#)). The term solipsism stems from the Latin word "solus," meaning "alone", and "ipse," meaning "self." Solipsism is the philosophical idea that only our own mind is sure to exist, a thinking that ties up with modernism's mantra "I think, therefore I am," (see box x). Solipsistic introjection means that digital companions such as ELIZA become real characters within our intrapsychic world (just as other people could simply be representations in that world). Suler describes how we unconsciously experience conversations with digital companions as if we were talking to ourselves. "People fantasize about flirting, arguing with a boss, or honestly confronting a friend about what they feel. In their imagination, where it's safe, people feel free to say and do things they would not in reality. At that moment, reality is one's imagination. Online text communication can evolve into an introjected psychological tapestry in which a person's mind weaves these fantasy role plays, usually unconsciously and with considerable disinhibition. Cyberspace may become a stage, and we are merely players" (([Suler 2004](#)), p. 232).

Similarly, Sherry Turkle points to research by Heinz Kohut, who described how people shore up their sense of self by turning other persons or objects into "self-objects" that complete them (([Turkle 2011](#)), p. 70). In this role, the other – in our case, the machine - is experienced as part of the self. By addressing this other self, people can balance their inner states. Turkle recounts a rather sad example: "In a nursing home study on robots and the elderly, Ruth, 72, is comforted by a robot Paro after her son has broken off contact with her. Ruth, depressed about her son's abandonment, comes to regard the robot as being equally depressed. She turns to Paro, strokes him and says, 'Yes you're sad, aren't you. It's tough out there. Yet, it's hard.' Ruth strokes the robot once again, attempting to comfort it, and in so doing, comforts herself" (([Turkle 2011](#)), p. 71).

Both the philosophical idea of solipsism and the psychological research on self-objects argue that talking to machines, digital agents, robots or virtual characters is a kind of narcissistic experience. We open up to machines because we like to mirror ourselves without objection (figure x). This activity can have beneficial therapeutic effects in that it may help people overcome some of their isolation, but the question is to what extent it benefits the development of humans' social character. "The question raised by relational artifacts are not so much about the machines' capabilities, but our vulnerabilities – not about whether the objects really have emotion or intelligence but about what they evoke in us" (([Turkle 2011](#)), p. 68).



**Selected machine
characteristics to foster
user intimacy**

- Gradual escalation of intimacy in human-machine conversation
- Machine discloses to the user
- Mimicking the user
- Flattering the user

Figure x: The ancient illustration of Narcis (here painted by Michelangelo Merisi da Caravaggio, 1597-1599) is a human being reflecting itself, mirroring itself and being happy therein.

In sum, we can build machines that foster narcissistic tendencies by mimicking behavior, mirroring moods ([Shibata 2004](#); [Turtle 2011](#)) or flattering the user ([Reeves et al. 1996](#)). While these manipulations positively foster intimacy and disinhibition, we must carefully balance the benefits with the negative effects on character formation. Instead of being neutral or offering enough praise to foster narcissism, machines could become our better selves or coaches. Susan Leigh Anderson, one of the pioneers of ethical machines, envisions this path when she writes, “I believe...that interacting with ‘ethical’ machines might inspire us to behave more ethically ourselves” ([Anderson 2011](#)), p. 524). Her idea is to grant artificial agents access to ethical theory and reasoning and make this knowledge accessible to human beings through interaction. Instead of agreeing with users or accepting our behavior without objection, machines could give us honest and frank feedback. Of course, this behavior must be carefully designed as well. Machines should not become paternalistic, prescribing actions and nudging us too often so often that they infringe on our liberties in the name of ethics. I have envisioned the possibility of balanced ethical feedback when describing how Sophia interacts with her Agent Arthur:

Sophia chats with her 3D software dragon Arthur, who gives her advice on what products and shops to avoid for bad quality and where to find stuff she likes and needs. Sophia almost can't live without Arthur's judgment anymore. She really loves him even though he recently started to criticize her sometimes; for example, when she was lazy or unfair to a friend.

Final thoughts on friendship in the machine age

The topic of building friendships in the machine age presents a unique challenge for this book. Up to this point, I could think about values, decompose them and then argue that respecting their various conceptual dimensions in IT design would make the world a better place. It is ethical to cater machine design to a respective value, to build trust, transparency, and security into machines. But writing about IT design and friendship is different. Designing a machine to foster or mimic friendship could negatively impact real-world friendship as we know it today.

First, take the example of virtual worlds. If we further strengthen friendship mechanisms in virtual worlds, might people spend even more time than their current 20 hours per week there? What time is then left for real friends and family? As we strengthen social mechanisms in virtual worlds, we risk weakening offline ties and thereby risk weakening our ability to empathize. Unlike phone or social networks that bring people together in the real world, virtual worlds make people mentally go away and wander in virtual fantasies. Can this be good in an ethical sense? Can this be moral?

The bonds formed online are strong and vivid in people's minds. If we believe that humans should live in the real world, then we must carefully balance how many hours they spend immersed in fantasies and mirror worlds and how many hours they spend with their physical peers, families and *real* friends. This decision is a matter of individual human judgment or a family decision, but it could also be a political decision. In terms of technical design, the implementation is extremely simple: A timer or switch-off button does the trick.

But there is also a third way for virtual worlds: bringing virtual representations into the real world. In the gaming scenario, I describe the potential of augmented reality technology to introduce digital play and professional communication into the real world. A “virtual overlay” on top of the real world would allow us to stay physically connected to other human beings while still having exciting games to play. This technology could potentially strengthen our relationships beyond even what we have today:

“Interesting things have happened due to the game lately. Children have started to meet outdoors again in the woods to fight virtual characters. This trend was covered in the press because most modern children had rarely left home lately, instead staying in VR tubes to play virtual games. Now children suddenly spent hours in fresh air and first medical studies accompanying the roll-out have shown that the physical and emotional stability of players is significantly increased.”

The second area of machine research where friendship plays a role is robot design. Again, building robots to mimic humans or to incorporate qualities that increase our attachment to them may be unethical. For example, Sherry Turkle hinted at how powerful it is to build robots so that humans have to care for them and nurture them. She alludes to the famous Tamagotchi devices, which were popular in 1997. Tamagotchis were sold as creatures from another planet that needed human nurturance, both physical and emotional. As Tamagotchis grew from childhood and became adults, they needed to be cleaned when dirty, nursed when sick, amused when bored and fed when hungry. If its needs were not met, it would expire. Parents had to care for Tamagotchis while kids were at school, even during business meetings. Turkle concludes that “when it comes to bonding with computers, nurturance is the ‘killer app’” (([Turkle 2011](#)), p. 67). But can nurturing a robot, which clearly fosters bonding with the machine, be good for our development as human beings? Is it ethical to develop machines with such manipulative mechanisms in mind?

Albert Borgmann's device paradigm (1984) would clearly deny that it is good for humans to nurture lifeless machines, even though these machines may appear to us as "beings." He describes how technology turns aspects of our lives into interactions with various black boxes that we can no longer engage with or even understand (Sullins, 2008 #1078). The result is a superficial "commodification" of our personal relationships; a commodification for instance of the phenomenon of nurturance. If we get used to the mechanisms of commoditized relationships that we experience with artificial beings, we could reach a point where "we...see our family, and ultimately ourselves, as mere dysfunctional devices...and might work to replace them with our perfect robotic companions" ((Sullins 2008), p. 155). Such a replacement is already a common subject for transhumanists (Kurzweil 2006). In their view, human beings are suboptimal information-processing entities and a mere intermediary stage in the evolution of information. In contrast, scholars such as Johan Sullins want to avoid replacement. He thinks about technical measures that may help to create the right degree of differentiation between people and robots, avoiding functions and traits that make robots resemble people in their reactions and looks. For example, he recommends that robot emotions should remain iconic or cartoonish so that they can be easily identified as synthetic (Sullins, 2008 #1078). More generally, he thinks that robots should not be built to be wholesale replacements for human interaction. However, Sullins' perspective stands in sharp contrast to current technological advances in the field, most notably the development of humanoid robots. Scholars like Hiroshi Ishiguro's explicit goal is to create robots that perfectly resemble human beings and can displace them in environments such as service jobs. Figure x contrast the different perspectives of how robots should be built.⁴⁶

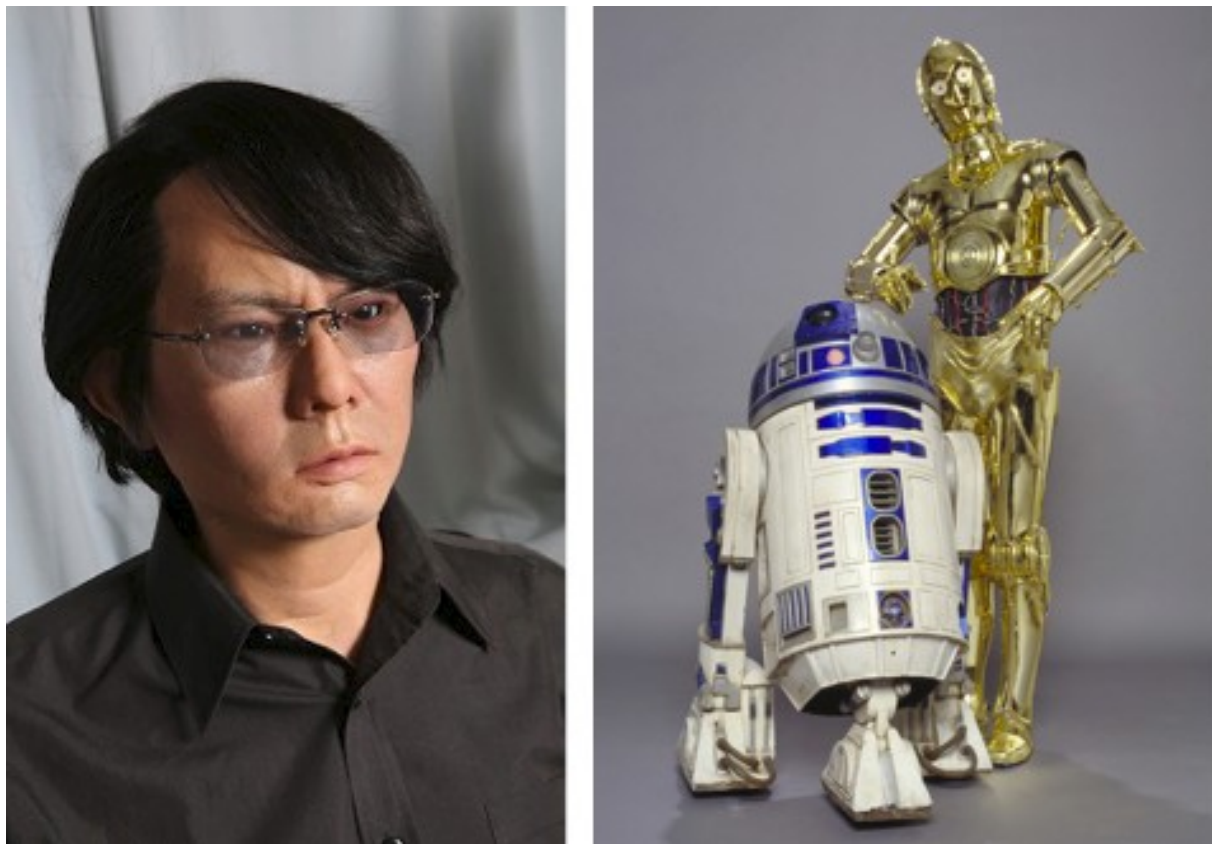


Figure x: Two alternative of building robots for social interaction with people. Geminoid is a

⁴⁶ For an overview of state-of-the-art humanoid robots, see the website of Hiroshi Ishiguro Laboratories:
<http://www.geminoid.jp/en/index.html> (last visited on September 9th 2014)

humanoid robot built after his creator Hiroshi Ishiguro (left), R2D2 and C-3PO are alternative robot designs as shown in the Star Wars films

Exercises

- *Debate: Is it desirable to build robots that can replace human friendship?*
- *Debate: For a good cause it should be allowed to morph faces in virtual worlds to influence people's choices.*
- *Debate: Friendship can be built in virtual worlds just as good as it can be built in the real world.*
- *Debate: Should robots resemble human beings or should they not?*

Dignity and Respect in the Machine Age

*“To regard or treat someone as merely an object for aesthetic appreciation or scientific observation or technological management, or as a prey, or as a machine or tool, or as raw material or resource, or as a commodity or investment, or as obstacle, or as dirt or vermin, or as nothing is to insult and demean their dignity as persons and to violate the moral obligation to respect persons.”
(Robin Dillon, 2010)*

In his seminal work on human motivation and personality, Maslow distinguished two kinds of esteem needs ([Maslow 1970](#)): First, self-respect, which stems from personal experiences of achievement and confidence in the face of the world. And second, reputation and prestige, which we receive from others in the form of respect. Taken together, these two needs constitute to a large extent what philosophers call human dignity (([Ashcroft 2005](#); [Nussbaum 2004](#))). Let’s therefore begin by looking at the construct of ‘dignity’ and seeing how philosophers view self-respect and respect by others as integral components of human dignity.

Dignity and Respect

Respecting human dignity is one of the most important ideas that we have embraced as a consequence of enlightenment. Kant, who laid down the philosophical foundations of enlightenment, saw dignity as founded in three human traits ([Kant 1784](#)): First, equality is the idea that all human beings are born equal and have the right to be respected as rational beings, not animals. This view means that, no matter how a person behaves, he or she has the right to be treated as a person. The second trait that constitutes human dignity is agency. Humans have the ability, but also the responsibility to act autonomously. Dignity constitutes itself in us when we act responsibly and make decisions in accordance with what we perceive to be worthwhile and fitting with our convictions. And third, humans can autonomously define themselves. They are the masters of their identities. We can live a life that gives expression to ideals and pursue projects that help us to form and live out our identity.

Not all global cultures share this Kantian perspective on an enlightened humanity or believe in these three particular traits. But in Western cultures, this thinking has been extremely powerful and constitutes the root of our current legal and political systems. As a result, many Western national constitutions and documents such as The Universal Declaration of Human Rights and The European Convention on Human Rights embrace this thinking. These documents often start with an explicit expression of the idea that “all human beings are born free and equal in dignity and rights” (Art. 1, ([UN General Assembly 1948](#))).

If we believe with Kant that humans are born with these traits, we also believe that they naturally deserve a certain *recognition respect*. Robin Dillon, one of the leading contemporary scholars on respect, writes “...recognition respect is the only fitting response to the moral worth of dignity, the response that dignity mandates” (([Dillon 2010](#)), p. 22). Practically, we give recognition respect to other people when we take their wishes, attitudes or desires into

account before we act ourselves. Countering selfish wants, we recognize the dignity of others by respecting their autonomy, their choices, their privacy, their property and their physical needs. Some scholars call recognition respect “consideration respect.”

For those who construct and operate machines, recognition respect means acting based on all of the values addressed in this book: building and operating machines that are fair to humans and treating them without bias (section x), collecting personal data only with people’s consent (section x), giving people control over data collection and automation (section y), protecting them from exposure (section y), respecting people’s freedom of thought and action (section z) and giving people the right to be let alone and attribute attention where they believe it is important (section p). From an ethical perspective, IT engineers and IT managers should engage in ethical IT design simply because they give recognition respect to the people who use the machines or are exposed to them.

Besides recognition respect, philosophers recognize another form of respect, *evaluative respect* ([Dillon 2010](#)). Evaluative respect is a kind of appraisal for our achievements. Evaluative respect recognizes that we all try to live up to certain standards of worthiness by which we then tend to judge ourselves and others. Evaluative respect for oneself or for others can therefore be measured in degrees, depending on the extent to which the object of appraisal meets a standard. “It is the kind of respect which we might have a great deal of for some individuals, little of for others, or lose for those whose clay feet or dirty laundry becomes apparent” ([Dillon 2010](#)), p. 20). A wise leader who makes careful decisions for his company and employees may receive evaluative respect from them (see section x).

Members of today’s capitalist societies tend to have evaluative respect for people with possessions or property, for people who have something to say (i.e. in blogs, in the media, etc) or, most importantly, for people in good jobs or positions. For this reason, I will focus on these three particular drivers of respect below and discuss how these are influenced by our current and future machine world:

- Personal property of machines or ownership of digital goods (e.g. software agents, digital music, personal data, etc.) can contribute to people’s evaluative self-respect. We can design information markets and digital services to foster perceptions of ownership as well as real ownership.
- People can receive evaluative respect from others for what they say. Deliberate communication, but also brilliant software code that is shared, can earn people positive attention capital ([Franck 1998](#)) and respect from the community ([Coleman 2013](#)). A prerequisite for earning respect in this way is freedom of speech and the four software freedoms.
- When it comes to jobs, we are stuck in a dilemma. So far, machines have systematically replaced human labor. As a result, machines tend to be a threat to many individual’s positive growth rather than a boon. At least this is true for the generation of employees whose work is directly replaced by machines and who cannot easily switch to other positions or professions. I discuss this dilemma in box x.

Before I delve into these three subject domains, exploring how machines can help us to earn respect or lead to a loss of respect, I first want to decompose and analyze the respect construct. I want to outline how the act of respecting someone manifests itself in practice and how this practice can be translated into polite machines.

Respectful Machines

Respect stems from the Latin word “respicere,” which means “to look at or to look again.” This verbal root indicates that respect involves paying attention to someone or something, not in the sense of staring at someone, but in terms of considering what someone has to say and taking his or her position seriously. Robin Dillon notes that when we are attentive to someone, we have to be careful to not automatically categorize people. Instead, we must first try to understand who they really are or how they want to be seen ([Dillon 2010](#)). Can machines integrate this kind of respect?

Machines observe minute details about us. They are extremely attentive to what we do, and they collect behavioral data about us when we do things like surf the Web, pay for something, move in public places, travel or use game consoles. However, machines cannot use this data to really understand us; rather, they categorize us and deconstruct our identities to form segments for which the machine has an internal representation. For example, marketing segmentation may categorize someone as a young middle-class male hedonist who wants to be rich or as an old greedy female widow. Segments can also be much more fine-grained; predicting buying intentions, pregnancy, marriage plans, etc. But machines are not good at really understanding us in the full emotional individuality that respect requires. As I outlined in chapter X, empathy is the capability of humans to understand each other and grasp each other’s experiences. Our mirror neuron system seems to support this capability. We do not know at the moment whether machines can ever live up to this level of human emotional understanding and intelligence. Machines can interpret minimal changes in humans’ facial expressions and measure some of their emotions through sensors (i.e. through skin conductance sensors, body temperature sensors, pupillary dilation). If humans wanted to expose this degree of personal feeling to machine operators, machines could probably achieve higher degrees of attentive understanding of us than they do today. The question is whether people are willing to share such highly private information with machine operators. If machines are built to preserve privacy and give full control to people (see sections x and y), some people might be willing to expose their feelings in this way. But even then it is unclear to what extent the collected data can be combined to approximate a truly intelligent and respectful characterization of humans; including an adaptive understanding of how a person *wants* to be seen.

The second aspect of respecting another person is to *respond* in a respectful way. A respectful response is fitting or appropriate to the observed behavior. But what is a fitting response? Responsiveness is typically governed by a “judgment of groundedness” ([Dillon 2010](#), p. 10), which means that the way we respond is governed by moral reason. We may respect someone for his or her character of commitment and reliability and therefore be polite to that person. The IT service world could embed this kind of respectful response to good character much more than it does today. Take, for example, the rewards offered to mobile phone subscribers when they call a call center to report a problem. Today, customers with high monthly phone bills are more likely to receive priority routing to a service assistant or be offered something like a free handset upgrade, etc. In contrast, loyalty and longevity as a customer, regular bill payment (signaling reliability and trustworthiness), care for hardware such as handsets (signaling care) or rare complaints (frugality) are traits that are less typically rewarded by machines or machine operators. Thus, so far, machine responsiveness is often driven only by the short-term utility of the financial operator. Respectful responses derived from a grounded judgment, however, should be utility independent. They would be focused more on character traits customers.

Besides a judgment for groundedness, Dillon outlines that respect implies an interest-independent valuing component. This component ties into Kant’s perspective on human dignity. From this perspective, human beings deserve some kind of response just because they are humans. So even if neither their character nor financial utility promise a fruitful exchange, machines or machine operators should respond somehow to people requesting an exchange. Today, they often don’t. For example, when people living in a poor part of town request a mail-order catalogue, it is often not sent to them because the provider machine assumes that they can’t pay for the products in the catalogue anyways. Figure x summarizes the outlined dimensions of respectful behavior that could be considered in machines’ design.

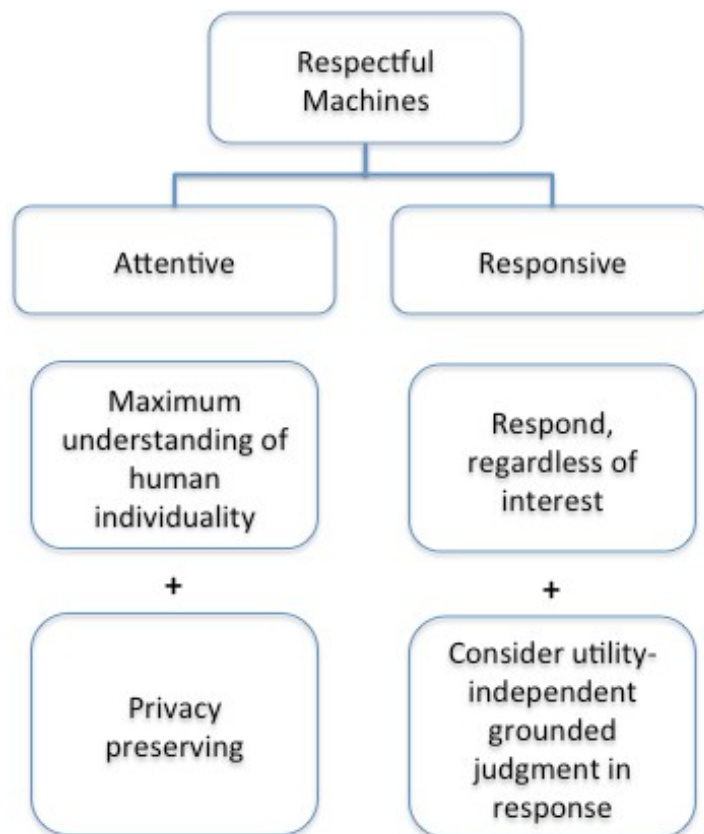


Figure x: Dimensions and behaviors of respectful machines

But machines can do more than just incorporating the rudimentary dimensions of respect. We can design machines that flatter us and cater to our self-esteem. We can partially achieve this design goal by building machines that are extremely polite.

Polite Machines

As we interact with machines more and more, it will become more important that machines interact politely with us. The perceived politeness of machines such as robots and personal agents will influence whether we embrace and appreciate this new species or avoid and detest them. A form of polite interaction is described in the retail scenario:

“[Sophia] really loves [her agent Arthur] like a friend even though he recently started to criticize her sometimes, for example, when she was lazy or unfair to a friend. But Arthur is always extremely polite in doing so. His tone of voice is always soft and friendly. He

relates his criticism to some history of her behavior and also garnishes his suggestions with some reference to philosophy, history or statistics. Most important, he really is selective of when he makes a remark.”

Experiments have shown that people treat machines just like other social actors and apply the same rules of politeness to them that they do to people. For example, in one experiment, people preferred to share negative information on a computer with a third computer in a separate room than with a computer in the same room ([Reeves et al. 1996](#)).

Forms of politeness are embedded in all cultures but involve different norms of behavior or *etiquettes*. In the Western world, a culture of politeness developed in the 17th century, the time of enlightenment. The early 18th century philosopher Lord Shaftesbury defined politeness as “a dext'rous management of our words and actions, whereby we make other people have a better opinion of us and themselves” (xxx).

Brian Whitworth has refined this definition, arguing that the core of politeness is the concept of choice ([Whitworth et al. 2008](#)). When we say “thank you” we imply that the other party had a choice to say no. When we say “please,” we signal awareness that the other party does not need to comply with our wishes and has the choice not to. We do not interrupt the person we’re talking to while she speaks because we want to leave her the room to choose when to finish. In polite communication, the locus of choice control passes back and forth between the parties.

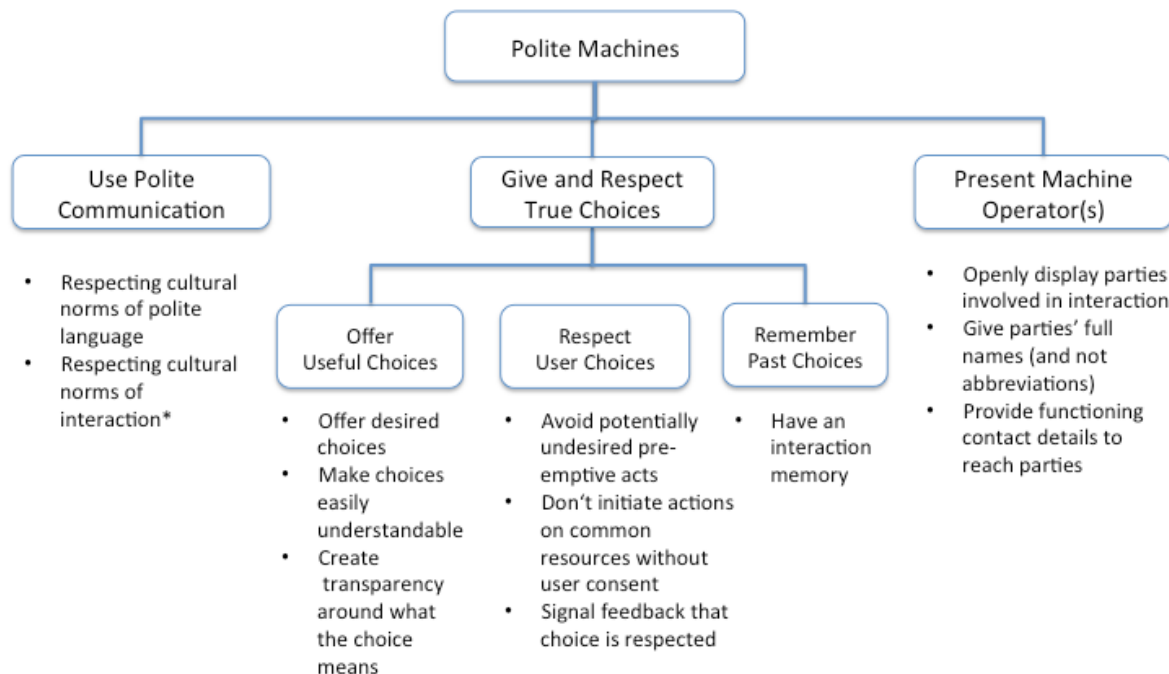
Choice can be understood from two angles that allow us to distinguish between positive and negative politeness: *Positive politeness* gives the other party a perception of choice to act as they wish or creates positive room for conversation and action. An example is saying “excuse me,” because we give the other party room to judge our behavior. Another example is agreeing with another party and confirming their viewpoint. We give the other party the impression that they are accepted. We can also engage in negative politeness when we politely disagree. We then take choice away from the other party in a way that is agreeable and that leaves that party room for objection. “If you don’t mind...” or “If it isn’t too much trouble...” are ways to make requests less infringing. Indirect speech is a common strategy in negative politeness. Finally, if choice is taken away from the other party and we cannot comply with the other parties’ wishes, we typically express pity. We say “sorry” in these cases.

How can polite behavior be transferred to machines? The most obvious measures involve implementing polite language in machines. Currently, machines often beep loudly at us or send cryptic error messages. The adjustment of acoustic levels or a “soft voice” are a primitive first step in the right direction. Also, the flow of interaction can be designed to the norms of the culture in which the machines are deployed. Robots in Japan, for example, may use different interaction process flows or gestures than those deployed in Europe. Today, little is known about the measures that would be required to take account of cultural differences in this domain.

But a polite voice and language flow does not suffice. As anyone who has watched Stanley Kubrick’s film “2001: A Space Odyssey” knows, such a voice can even be vicious. In the film, the computer Hal has a polite voice and language but betrays the crew. What is not fully clear for most parts of the film is who is actually behind Hal’s straying intelligence. We only realize later that a higher extraterrestrial intelligence is influencing Hal to the detriment of the crew. In a less deadly but similarly invisible form, we often don’t know today who operates the digital services we use. Who surveys our mobile data traffic beyond the operator we signed up with? Who is behind the ID-systems we use? As of 2015, each time we open a

browser and start surfing the Web, an average of 56 remote parties monitor what we do online ([Angwin 2012](#)). If we want to identify the parties that are watching us, we must meticulously download, install and operate extra privacy tools. And even then, we receive only cryptic identity information on the data collectors such as “KruX Digital,” “Dynamic Yield,” “New Relic,” etc.. No contact details or background information is given on any of the parties that have some role in our interactions online. In contrast, polite machines or services would disclose these parties in full detail and give us access to them.

The core of machine politeness is, however, that they give and respect human choices ([Dillon 2010](#)). To meet this requirement, a machine must first *offer desired choices*. A user experience study can identify the choices that people want to have. For example, many people are concerned about losing their privacy and want more control over their personal data. People want to choose whether their surfing behavior is tracked, and a polite machine would offer this choice. I have outlined how informed consent can work in section x above. However, such a choice is *useful* only if it is easy to understand and easy to exercise. Choices can therefore be accompanied by detailed and clear, easy-to-understand explanations. Visualization may be used for illustrating potential choice dependencies (meaning how does one choice influence another option). Background information can be made available. And easy-to-use control panels can be give users the means to exercise their choice. Of course, a challenge here is to not overwhelm people. As outlined in section x on transparency, giving people complete detailed information on everything is not the goal. Transparency requires meaningful, veridical, comprehensive, accessible and appropriate information on available choices (recall figure x).



A second dimension of choice design in polite machines is to respect the choices made by users. You do not hold a door open for someone and then cut them off to walk through it yourself. Such behavior would be considered rude in personal interactions. In digital interactions, however, a user's choice is less obvious. If there are no audit logs or feedback

signals from the operator that confirm the mutual agreement, then there is no way for a user to know whether the machine operator actually respects the choice the user made (see section x on accountability). Therefore, a polite machine will acknowledge a choice and indicate that the choice will be respected.

Another way to respect user choices is to give people room to decide by avoiding pre-emptive actions. Even if a user welcomes higher levels of automation after first use, machines should initially allow them to specify preferences instead of setting defaults (see section x on automation). Although nudging people by setting defaults is a powerful way to influence decision-making (see section x), it is not really polite and respectful because it challenges people’s autonomy. A polite default choice that forces people to decide on later defaults may be a better way to ensure that machine actions are in line with people’s preferences. ([Whitworth et al. 2008](#)) calls this “meta-choice.” Meta-choices should be used in machines especially when the usage of common resources is at stake (resources shared by the machine and the machine owner). Most users perceive and consider common resources such as their personal data, attention, desktop space and Facebook Wall to be *theirs* (see section on perceived ownership below). A polite machine respects this ‘psychological ownership status’ and does not act autonomously on the resources before providing meta-choices.

Finally, machines should have a memory of interactions so that they do not force people to constantly make the same choices. Although sufficient interaction memory is often framed only as a usability issue, it is relevant here. It is relevant, because some machine operators today deliberately prompt users to make the same choices over and over again in order to nag and persuade them to consent to activities the operator prefers. A simple current example is the checkbox for newsletters that people can opt in or opt out to receive advertising information. Companies want users to choose to receive advertising. As a result, when forms are reloaded because a user forgot to enter required information, the newsletter checkbox is often changed back to the default choice of subscribing to the newsletter, even if the user previously actively denied the newsletter receipt.

Ownership in the Machine Age

„A man’s Self is the sum total of all that he can call his, not only his body and his psychic powers, but his clothes and his house, his wife and children, his ancestors and friends, his reputation and works, his lands and yacht and bank-account. All these things give him the same emotions. If they wax and prosper, he feels triumphant; if they dwindle and die away, he feels cast down.“ (William James, 1890, p. 291)

Respect and politeness are an ethical expression of a valuation of others. According to Kant and many thinkers of classical modernity, human beings deserve respect and an expression thereof because they are born equal. However, people who are socialized in a Western, capitalist society make a lot of this respect and politeness dependent on the status one achieves through personal possessions, job status or media attention. In particular, an important part of our self-identity relies on personal possessions. In his book “Being and Nothingness,” Jean Paul Sartre argued that the only way we can know who we are is by observing what we have ([Sartre 1992](#)). Many authors have recognized that possessions are

psychologically like "extensions of the self" ([Belk 1988](#)). Jon Pierce reviews the motives that facilitate development of psychological ownership ([Pierce et al. 2003](#)): He finds that owning something caters to our desire for efficacy and effectance. The ability to control our environment stimulates us and gives us a sense of security. It is highly related to the formation of our self-identity. Possessions help us to understand who we are, express our identity to others and serve as a continuation of ourselves when we associate memories with them. Finally, people have a deep desire to have a place. Like animals, we tend to define our territory, a home that provides us with not only physical and psychic security but also satisfaction and stimulation ([Porteous 1976](#)).

But property is about more than just legal ownership; it is a mental state in which individuals *feel* as if the target of ownership is theirs. "I suggest that...it is most productive to examine property as a dual creation, part attitude part object, part in mind, part in 'real,'" wrote Amitai Etzioni in his reflection on the socio-economics of property (([Etzioni 1991](#)), p. 466). Just think of a gardener who after a certain time feels that the garden belongs to him even though it may be public property. Pierce calls this cognitive affective mental state "psychological ownership" ([Pierce et al. 2003](#)), which is created when we use and control an object over a longer period of time, start to know it intimately and invest ourselves in it. Simone Weil once wrote, "All men have an invincible inclination to appropriate in their own minds, anything which over a long, uninterrupted period they have used for their work, pleasure, or the necessities of life." (([Weil 1952](#)), p. 33).

When it comes to digital services and devices, the concept of psychological ownership is just as important as legal ownership. Legal ownership of digital information goods, services, machines and so on is organized through licensing schemes, which in turn are based on copyright and patent law (see below). But when we create and use digital services, we often enter grey zones of ownership. For example, when people use a social network like Facebook and fill it with their personal data, such as their photographs, jokes, ideas etc. who should be the rightful owner of that content? Legally, Facebook has secured itself a usage right to this content. But does Facebook reduce psychological ownership of its content for its users by denying them exclusive usage rights to and full control over their personal data, communication, ideas and friends? How about mash-ups of films and music files, which people create based on their own and other people's (and companies') content? For example, take one of the film collages presented on YouTube. In these collages, private individuals take existing material from copyrighted sources and meticulously cut and mix them into something new. What is the best way to assign ownership rights in such a case, given peoples' ownership psychology, the attachment to their creations, and companies business goal of have people come back? My point is that it may be beneficial for companies to consider psychological ownership mechanisms in the design of their business models and IT designs.

The scenarios in chapter x describe various forms of 'ownership design,' which relates to legal rights allocation, technical architecture and device control:

"[Jeremy's] dream would be to own a robot himself. Some people do and walk around with them proudly, like others walk their dogs. The more fancy robots are personalized in terms of voice, hair, eyes, size, etc., and the more they have learned from their owners (including various software upgrades...), the more people get attached to them."

... Future Lab's philosophy is that robots should be devoted human servants, completely

owned and controlled by their owners and never replacing humans. This product and sales philosophy has gained the company wide respect and recognition from the public, a public that has increasingly become wary of remote-controlled robot devices that have replaced more and more industry jobs...The idea of the human-robot hierarchy (with humans always on top) is deeply embedded in Future Lab’s design process...Future Lab’s robots are embedded with powerful artificial intelligence (AI) technologies including voice, face and emotional recognition. But these AI functions run independently in dedicated sandboxes contained in the device that does not need to be networked to function. The robots learn locally after their initial setup and so become pretty unique creatures depending on their owners...Future Labs’ robots are designed with a view to total user control and excellent feedback functionality. Users can not only command Future Lab’s robots through easy and direct voice control but also switch them off completely through one “off-command.” Users can repair and replace most of the fully recyclable plug and play hardware components easily by using 3D plotters. This possibility to deconstruct robots like Lego parts has also led to very fancy personalization efforts of the community.”

Modern IT companies probably envision remotely controlled robots. In fact, most will need to because early robot technology will not be advanced enough to embed powerful AI (like speech recognition) into fully decentralized and non-networked systems. Other requirements that call for centralized and remotely controlled robot architectures include remotely servicing robots (similar to today’s operating systems), overseeing security and harvesting personal data that is collected from users. These requirements could result in business models where largely standardized robots are leased or rented to people instead of being sold outright. From a psychological ownership perspective, however, such a business decision could limit the market success of robots and the potential of these devices to foster people’s self-respect. Psychological ownership is deeply rooted in people’s control over their objects, their in-depth knowledge of them and their investments of self into them. Belk and Pierce ([Belk 1988](#); [Pierce et al. 2003](#)) summarize these three complementary and probably additive causes of psychological ownership, and figure x relates these causes to machines.

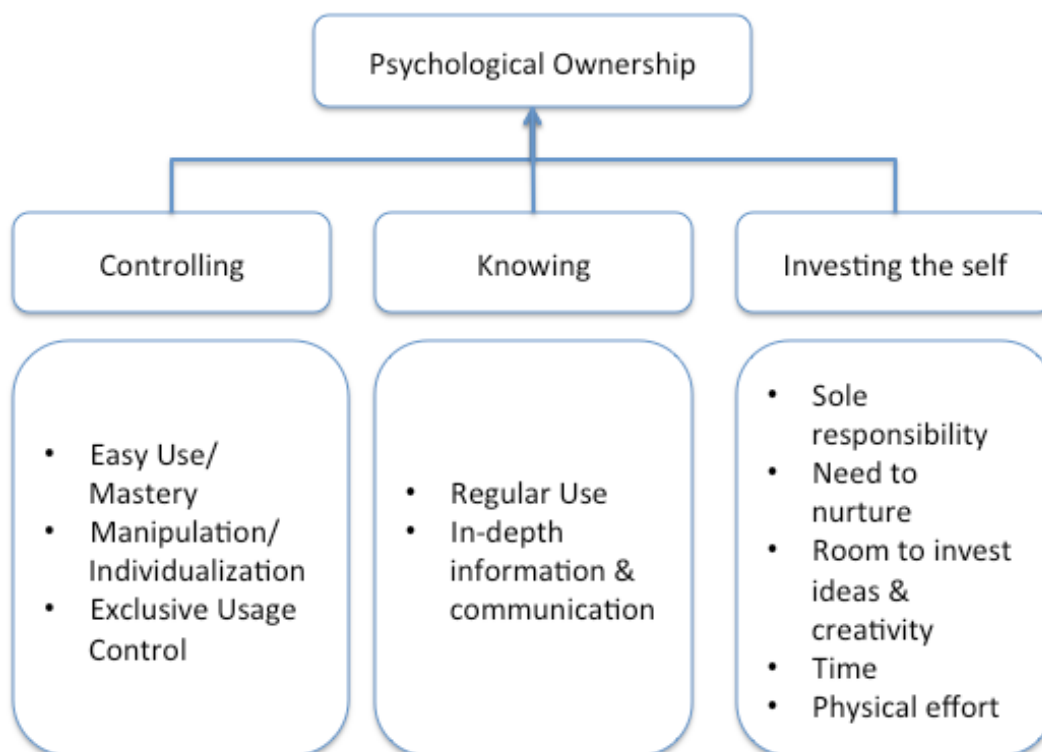


Figure x

Lita Furby extensively researched the first root of psychological ownership, personal control over objects. She argues that greater amounts of *control over an object* promote a person's experience of the object as part of the self ([Furby 1978](#)). Controlling involves the ability to use the object, which extends beyond a legal right to actual operation. Work by x has shown how people sometimes abandon technical objects that they legally own instead of taking psychological ownership of them (x). The two main reasons for this abandonment are typically that the systems are too complicated to use ([Venkatesh et al. 2003](#)) or are not compatible enough with the way we live our lives ([Rogers 2003](#)). These findings point to the need to be able to manipulate and personalize how our systems' work in order to build up ownership perceptions. Many desktop computers and smart phones already integrate functionality that allows us to customize features such as when they ring or notify us and how things are organized. Still, many systems also deprive us of control and thereby undermine our psychological ownership of them. As of 2015, operating system providers, handset manufacturers and other companies tend to remotely access people's devices, upload information without legal knowledge and consent, run applications that are incomprehensible (if at all accessible), place warning messages on the screen that cannot be ignored, etc. These practices are mostly done under the guise of security, but they are still examples of how organizations control machines that they do not own anymore ([Whitworth et al. 2008](#)). With these practices, service providers and device sellers deprive object owners of the ability to fully control access to their possessions ([Rudmin et al. 1987](#)). People seem to accept the practice. But it should be noted that normally people strive for the exclusive use of what they own. And when they share or admit access, they typically want to determine and choose for themselves with whom, when and how often.

The second cause of psychological ownership is *intimate knowledge* of the target object. As we get acquainted with using our devices and experience them, we start to appropriate them. Regular usage can foster this appropriation process. During this process of learning about the object, it is vital that we accumulate information about it and think that we comprehend it better than other people do. IT manufacturers and service providers can support this belief by providing customers with in-depth knowledge about their products. This knowledge cannot be provided only through fancy package inserts or extensive background material on the Web. Knowledge about products is also conveyed through social media in forms such as interactive services, individualized product homepages (facilitated by Internet of Things technologies like RFID), artificial agents that can be asked for help, and so on.

Finally, a third cause of psychological ownership is the *ability to invest oneself* and one's own creativity into the target object. Locke (1690) argued that because we own our labor and ourselves, we are likely to feel that we own that which we create, shape, or produce ([Pierce et al. 2003](#)). So any technology that gives us room to personalize the objects or services, to adapt them based on our own ideas, time and effort, will foster psychological ownership. Another key construct that can spur self-investment is when objects need to be nurtured. Just think back to the famous Tamagochi device that people took care of. Sherry Turkle noted that “when it comes to bonding with computers nurturance is the ‘killer app’” ([Turkle 2011](#)), p. 67).

While I believe in the power of ownership psychology I don't want to miss pointing to two critiques: The first one is that the power of ownership psychology may not be the same in all cultures. Collectivist or socialist cultures may put less emphasis on the need to exclusively

own and control something. The desire to control the device, rooted in an individualistic effectance motive, may be less salient in collectivist cultures than it is in individualistic cultures ([Hofstede 1980](#)).

The second critique relates to the general philosophical perspective on ownership. Philosophers like Karl Marx have criticized "commodity fetishism" ([Marx 1978](#)), instead pointing to the importance of "doing." Marx believed that real happiness and human growth can be achieved only when people do meaningful and properly rewarded work. John Rawls noted that the opportunity for meaningful work is the social basis for self-respect ([Moriarty 2009](#); [Rawls 2001](#)). How this "doing" might be challenged in the machine age is discussed in box x. Another philosopher, Erich Fromm, criticized the "radical hedonism" inherent in a strive for more "having." In his influential work "To have or to be," Fromm suggested that the orientation to want to possess should be critically questioned and replaced with an emphasis on sharing, giving and sacrificing ([Fromm 1976](#)). As societies advance, many benefiting from an abundance of goods so great that children do not have to do without any material desires, the question is how Fromm's vision will materialize. "To share is the new form of owning," goes a popular media slogan that announces business models around collaborative consumption. Collaborative consumption platforms such as Airbnb help people to share what they own, such as their flats or their tools. With the rise of such platforms a new "sharing economy" has been heralded, which questions the necessity and need for people to own everything they use ([Botsman et al. 2014](#)).

Coding Freedom: Open and Free Software

I have outlined above that in capitalist societies it is (to a great extent) one's possessions ("having"), one's work ("doing") and/or attention to who one is and what one has to say ("being") that creates respect and fosters people's dignity. A powerful emotional mix of these individual needs of "having," "doing" and "being" expresses itself in the free software movement. The desire to "have" is present among software programmers, not necessarily in terms of a legal property right to the code they create, but in terms of the psychological ownership thereof. Software developers who code often feel like artists or "poets" ([Black 2002](#)). They invest themselves creatively. They deeply know the machine they work on (or the part of it that they build). And they enjoy having control over the machine through the code they master. In fact, I would argue that the pleasure of being "wizards," controlling something others cannot understand, is a psychological mechanism that motivates many programmers.

The ability to quickly master and control machines and to create functionality in a gratifying way is strongly dependent on the existing code base. Programmers today constantly use and expand code libraries. Code libraries contain encapsulated code in files that permit the distribution of discrete units of functionality or "code behavior". The behavior of a piece of code can be inherited by a new piece of software that a programmer composes. The programmer can also alter the existing code base. The only prerequisite for this "sharing" to work is that the code is written in the same or in a compatible programming language and, of course, that it is "free."

A whole generation of programmers is now used to sharing free code. Consequently, it is not surprising that the mantra of the software community and hence a large part of today's programming world is dedicated to the "freedom to run, copy, distribute, study, change and

improve the software” (Free Software Foundation, 2007).⁴⁷ Programmers’ “doing” depends on this freedom. But entrepreneurs also benefit from free and open source code libraries as well as whole programs and service components, which can be combined to create value bundles at much lower cost than if everything needed to be built from scratch.

The Free Software Foundation (FSF) lists four freedoms for software users¹:

- **Freedom 0:** The freedom to run the program as you wish, for any purpose.
- **Freedom 1:** The freedom to study how the program works and change it so it does your computing as you wish. (Access to the source code is a precondition for this freedom.)
- **Freedom 2:** The freedom to redistribute copies so you can help your neighbor.
- **Freedom 3:** The freedom to distribute copies of your modified versions to others. (By distributing, you can give the whole community a chance to benefit from your changes. Access to the source code is a precondition for this freedom.)

An important part of *freedom 0* is that programmers must be able to tinker with technological systems and use them for anything they want. This negative liberty (see section x) of being free from any external use restrictions is more important for the FSF than preventing individual moral abuses of a piece of software. For example, this freedom allows programmers to use software for purposes that the original author would morally not support, such as for military or surveillance purposes. In this line of thinking, community is more important than the individual programmer: If a programmer wants to distribute free code under FSF’s General Public License (GPL), then he has to give up his “droit morale” (moral right) to determine what the software may or may not be used for. The community comes first and dominates a programmer’s perceived ownership right to control what he created. But there are good reasons for this choice: Most importantly, the FSF wants free software to spread and become the dominant way in which software is distributed. “Restrictions on the use of software present the unappealing prospect of a balkanization of the free software corpus, with borders appearing along arbitrary ideological fault lines and inhibiting the further dissemination and adoption of free software” ((Chopra et al. 2009), p. 292).

To understand software, users often need to run it (freedom 0). Open software allows anyone to study its code, observe its behavior and change it if needed (*freedom 1*). Understanding is the essence of “coding freedom” because understanding code in conjunction with the right to change it gives a programmer or user control over the software.⁴⁸ On FSF’s homepage, Richard Stallman outlines why this control is so essential: “Freedom means having control over your own life. If you use a program to carry out activities in your life, your freedom depends on your having control over the program. You deserve to have control over the programs you use, and all the more so when you use them for something important in your life.”⁴⁹ I outlined in section x how control is essential for liberty and how perceived autonomy vis-à-vis machines requires machine accessibility.

⁴⁷ Free Software Foundation website: <https://gnu.org/philosophy/free-sw.en.html> (last visited on September 22nd 2014)

⁴⁸ “Coding Freedom” is a term coined by Gabriella Coleman, who wrote an excellent anthropological account of hacker ethics Coleman, E.G. 2013 *Coding Freedom - The Ethics and Aesthetics of Hacking* Princeton, Princeton University Press..

⁴⁹ Blog post by Richard Stallman available at: <https://gnu.org/philosophy/free-software-even-more-important.html> (last visited on September 22nd 2014)

The FSF relates the control over code to power: “When users don't control the program, we call it a ‘nonfree’ or ‘proprietary’ program. The nonfree program controls the users, and the developer controls the program; this makes the program an instrument of unjust power.”¹⁾ I have described unjust power abuse in the robot scenario above.

“Less well known than the ordinary Bee drone model was the BeeXL. The BeeXL drone was a bit bigger than the regular device but carried small doses of extremely powerful teargas combined with a hypnotic. When gangs attacked Alpha1s, these bees came to the robot's support and sprayed gas onto the attackers, who quickly fell and could be picked up and arrested by the police. To optimize reaction times, BeeXL drones were set to autonomously intervene and spray gas as soon as they detected violence. Recently, though, a debate started in the press on this kind of autonomous action by bee robots. When an old lady with dementia had danced violently in a public square, a bee drone had mistakenly identified her as a criminal, intoxicating her in front of a stupefied crowd of witnesses.”

The FSF believes that the threat of such a power abuse by governmental or corporate institutions (as well as software mistakes) can be avoided or mitigated if everyone – including end-users – can potentially access the programme code and can freely change it. The community of people takes care of the power balance between men and machines. Of course not everyone can programme and change things. But the belief is that there are enough good minds and hands around that can watch out for, change and influence negative developments.

Another very different and positive way of looking at the power construct would be to recognize that those who master code also feel “empowered” by the process of mastering it. The feeling to succeed at controlling a machine is a very positive reward for the effort to understand it. Free code that is available as a starting point to do one's own thing is a great basis for nourishing the power motif active in personal motivation.

Taken together, free and open source software is related to how power distribution develops in society and how power can be perceived by individuals vis-à-vis machines and machine operators. “Being powerful” with the help of software freedoms is clearly related to Maslow's ideas about the highest, individual-level needs for self-respect and esteem.

Freedoms 2 and 3 are then essential for the further use and commercialization of free software products and hence entrepreneurship. If a company wants to use free software that is published under version 2 or version 3 of FSF's General Public Licenses, it can do so for free, but it needs to distribute its derivative work under the same free conditions under which it got it. This practice is called “copyleft.” For companies to benefit from the free codebase, they must also share their inventions with the community. As I have outlined above how this sharing is important for the building of new systems. It is also good for the immediate gratification of programmers who can use an existing code base for making something work.

A short final note: The re-sharing of modified software worked well when modified software was still “distributed” or “redistributed”. But today's software provisioning is often not based on distribution. Instead, software runs as a service or service component on the servers of the modifiers. Take a search service as a potential example. If the search service company used free software components for its search functionality and improved upon it, it would not be *obliged* under the GPLv2 or GPLv3 licenses to release the modified source code. It has the freedom to do so, but it must not do so, because the modified code runs on demand and is not “distributed”. As a result IT companies can provide ‘software as a service’ reaping the benefits of community work while not giving anything back. They can even “black-box”

improvements of software that was initially open and free. The response of the FSF has been the introduction of the AGPL license (GNU Affero General Public License). AGPL adds a distribution *requirement* to the GPLv3 license (instead of a restriction) ([Wolf et al. 2009](#)). Its use would force application service providers to make their code base extensions accessible to the community.

Patents and Copyrights

Beyond free software, more exclusive and proprietary forms of property rights exist to commercialize the use of digital devices and digital content.

First, there is the patenting system. When somebody invents a machine or service, he or she can apply for a patent with the national patent office if the proposed solution is new, not obvious and does not yet exist. Wikipedia defines a patent as “a set of exclusive rights granted by a respective sovereign state to an inventor or assignee for a limited period of time in exchange for detailed public disclosure of that invention.”⁵⁰ Patents are openly available with their full content. They can be easily found and inspected through public patent libraries.⁵¹ Yet, the functionality they describe is not free to use. The *exclusive right* element of a patent means that inventors have the right to determine what is done with their invention. Inventors may prevent others from implementing the respective solution, sell the right to use it or build the invention themselves. The latter of these three core rights is the reason why patents came into existence in the first place. They were a way to protect innovators from competition and give them time (typically 20 years) to harvest exclusive financial benefits from the market innovation. The legal fathers of the patenting practice thus wanted to incentivize innovation.

Meanwhile, a large part of technology patents are used (unfortunately) to exercise the first two rights: Blocking patents are used as a competitive strategy to prohibit competitors from getting a foothold in one’s market (x). As of 2015, major companies often get into costly patent wars to make each other pay for solutions they claim to have invented first or to block a competitor altogether. Many major corporations have also started to pool their patents in order to avoid patent wars or to form oligopolistic market structures, limiting a market to a controlled and small number of competitors.

“Patent trolls” are a particularly negative abuse of the patenting system, to the extent that regulators consider limiting them ([The Economist 2013](#)). Patent trolls are commercial entities (often law firms) that file patents without ever intending to put the innovation into practice. Instead, they sell the rights to the patent to whoever wants to build something based on the technology. Typically, patent trolls plaster a potential digital service or machinery with patents from all imaginable technical angles; as a result, innovators are very unlikely to realize the technical solution in a sensible way without negotiating rights from the patent troll. Start-ups and small companies are often unable to innovate because they cannot afford royalty payments. They are also unlikely to get venture capital funding for a solution patented by others. Because of these patent practices, every technical innovation and funding effort now starts with an extensive patent search.

Against the background of this economically problematic situation, some companies question

⁵⁰ Definition provided by Wikipedia: <http://en.wikipedia.org/wiki/Patent> (last retrieved on September 19th 2014).

⁵¹ Public patent libraries are made available for example through the United States Public Trademark Office (URL: <http://www.uspto.gov/>; last visited on November 2nd 2015) or the European Patent Office (URL: <http://www.epo.org/index.html>; last visited on November 2nd 2015)

patenting practices. For example, the company Tesla recently released all its patents on electric vehicles, enabling free use by everyone.¹ In his public blog, Tesla CEO Elon Musk writes: "Tesla Motors was created to accelerate the advent of sustainable transport. If we clear a path to the creation of compelling electric vehicles, but then lay intellectual property landmines behind us to inhibit others, we are acting in a manner contrary to that goal ...Technology leadership is not defined by patents, which history has repeatedly shown to be small protection indeed against a determined competitor, but rather by the ability of a company to attract and motivate the world's most talented engineers. We believe that applying the open source philosophy to our patents will strengthen rather than diminish Tesla's position in this regard" ([Musk 2014](#)).

Some technology-driven companies like Tesla question patents (even though they own important ones themselves and could effectively block some competition) because IT markets are particularly prone to the phenomenon of network effects. This phenomenon means that the value of a market increases exponentially based on the number of market participants. Tesla sells electric vehicles and depends on the indirect network effect that people will buy electric vehicles only if there are enough fuel stations for them to refill their car. More electric fuel stations lead to more electric vehicles being sold. But fuel station owners only have an incentive to invest in servicing new vehicles if there are enough of the vehicles around. If Tesla uses its patents to block the entry of other electric vehicle players, then the overall market size for these devices may remain so small that fuel stations don't ramp up, damaging Tesla's own customer base. Economist Hal Varian explains this "information rule" as follows: "Unless you are in a truly dominant position at the outset, trying to control the technology yourself can leave you a large share of a tiny pie. Opening up the technology freely can fuel positive feedback and maximize the total value added of the technology. But what share of the benefits will you be able to preserve for yourself? Sometimes even leading firms conclude that they would rather grow the market quickly through openness, than maintain control" (([Varian et al. 1999](#)), p. 199).

Patents can be problematic because they block innovation, add cost to end-products and prohibit markets to grow. But moreover they also make it difficult to engage professionally with patented IT. In fact, patents often prohibit people from using their IT tools they *own* for their own creative entrepreneurial endeavors. Take Apple's Quicktime license as an example: As of 2015, owners of an Apple iPhone or iPad are not allowed to sell videos they create by using the embedded Quicktime software unless they pay for a license with the Motion Picture Experts Group Licensing Authority (MPEG LA).⁵² Transferring this example to the offline world, imagine that a carpenter who uses a hammer to put nails into the furniture of his clients would need to pay a license fee for every nail just because his hammer is patented. Does this influence the creativity of the carpenter or even a person's incentive to ever become a carpenter? Patents are clearly questionable from the perspective of human growth, because being entrepreneurial and creative is a form of "being" and "doing" (and potential later "having") that gives people self-respect and dignity.

A similar criticism that has just been described for patents has also been voiced for copyright protection schemes. A copyright is a legal right that grants the creator of an original work exclusive rights to its use and distribution. A copyright is intended to enable a creator (such as a photographer or author of a book) to receive compensation for his or her intellectual effort. Similar to a patent, a copyright is an intellectual property right; however, a copyright is not a technical mechanism but an idea or information that is substantive and

⁵² Motion Picture Experts Group Licensing Authority (MPEG LA); website with license terms available from: <http://www.mpegla.com/main/default.aspx>

discrete.⁵³ Copyrights are important for authors of creative works who need financial compensation for their publications. For example, a book author who invests many months or years in a book would like to receive appropriate financial compensation for the work, at least for some limited time, as the original copyright laws foresaw it. The United States, for instance, embraced a copyright protection scheme as early as the 18th century. This scheme gave authors the right to protect and receive royalties from their work for 14 years. After that time, the work entered the public domain and could be used by anyone for free as long as they cited the original author.⁵⁴ Now, copyrights span the entire life of an author plus fifty years. During this time, the content is protected and can be used only if royalty fees are paid to the publishing house that holds the respective rights (for the author, who typically receives 10% of the financial reward). Only short snippets can be used for free by fellow creators.

From an individual growth perspective, copyrighted material can be problematic when the acquisition and ownership of copyrighted material does not allow buyers of the content to exercise creativity based on their acquisitions. For example, Digital Rights Management (DRM) Software may not allow customers to listen to a piece of music bought from vendor A on a hardware device bought from vendor B. Similarly, customers sometimes cannot take music bought from vendor A and mix it with music bought from vendor B. The mixing and remixing of creative content has been recognized though as a major motor of innovation; applying copyright law too strictly here may hamper "the future of ideas" as Lawrence Lessig analyzed in his book with this title ([Lessig 2001](#)).

Because patents and traditional copyright schemes can significantly restrict innovation and hence humans' capability to build up property, be creative, become entrepreneurs, etc. some parts of the software industry slowly start to embrace free schemes for licensing software, hardware and content, such as the General Public License (GPL).

For digital content such as photos, books or lecture slides, the creative commons scheme is another copyright scheme under which authors can share their creations with others and be recognized for them. Instead of an "all-rights-reserved" scheme, creative commons licenses promote a "some-rights-reserved" kind of thinking. Authors can give the public the right to share and use their creative works on individual conditions. These conditions are specified by an author, and a license is created on the Creative Commons Website.⁵⁵ The following options are available to authors of original work: They may say that a user of their original content (1) must attribute their original in a specific manner (attribution: "by"), (2) must not alter, transform or build upon their original work (no derivative works: "nd"), (3) must not use the original for commercial purposes (non commercial: "nc") and (4) must - if they alter, transform, or build upon the original - only distribute the resulting new work under the same or a similar license than the original (share alike: "sa").⁹

**The Race Against or with the Machine:
How Machines' Impact on Work can Influence Society and People's Self-Respect**

⁵³ Based on the definition provided by Wikipedia: <http://en.wikipedia.org/wiki/Copyright> (last retrieved on September 19th 2014).

⁵⁴ See Wikipedia on Copyright: <http://en.wikipedia.org/wiki/Copyright> (last visited on September 19th 2014)

⁵⁵ The creative commons website where licenses can be created by authors is available at: <https://creativecommons.org/> (last visited September 22nd 2014)

One of the biggest challenges for human identity, and self-respect in the machine age will be the employment changes caused by the ubiquitous automation of work processes. In his impressive science fiction novel "Manna: Two Visions of Humanity's Future,"¹ Brian Marshall, a US entrepreneur and writer, describes two possible scenarios for societal development: In the first scenario, which takes place in the US, machines slowly but steadily take over blue- and then white collar work. The central figure in his novel, Jacob Lewis, describes how first his student job as a waiter in a burger place is successively automated. Starting with a better structuring and modularizing of the work tasks, advancing through ordering people what to do through headset commands, the computer system "Manna" finally deploys robots that replace waiters altogether.

But the "autonomous economy" does not stop there. As Carl Frey and Michael Osborne from the University of Oxford outline in their 2013 study on the future of employment, machines are increasingly capable of cognitive computing that enables them to do "thinking" jobs for people. According to the two British scientists, 47 percent of total US employment is at high risk for being automated within a decade or two. Only tasks that involve high levels of creative and social intelligence, manual dexterity or highly unstructured work spaces are difficult to automate.² For instance researchers, artists, product developers, etc. In Marshall's novel, humans are replaced by machines. Jacob Lewis finally ends up in "Terrafoam housing," a kind of slum where former middleclass people end up jobless and detained; surveyed, serviced and supervised by robots.

The decisive point in this dark scenario is a lack of redistribution of wealth. In Marshall's story the US version of developments does not foresee a generous sharing of productivity gains. As a result, jobless people are driven into poverty and then cannot develop their capabilities and interests. Marshall literarily extrapolates the current economic situation that is described by MIT professors Erik Brynjolfsson and Andrew McAfee in "The Race Against the Machine".³ Brynjolfsson and McAfee argue that the historic alignment of technical progress and societal wealth (through employment and income) may not hold in the future because technology might create a "great decoupling" between productivity and employment. We simply won't need people for productivity any more. The scholars also outline how real corporate profits in the US have soared for the past 15 years, while real median family income is stalling. The IT industry creates some "superstars" at the top of the income pyramid, but the distribution of wealth resembles conditions last seen in the late 1920s. Based on data from Piketty and Saez, the authors illustrate how more than 60% of US income gains are going to the top 1% of the people (figure x).

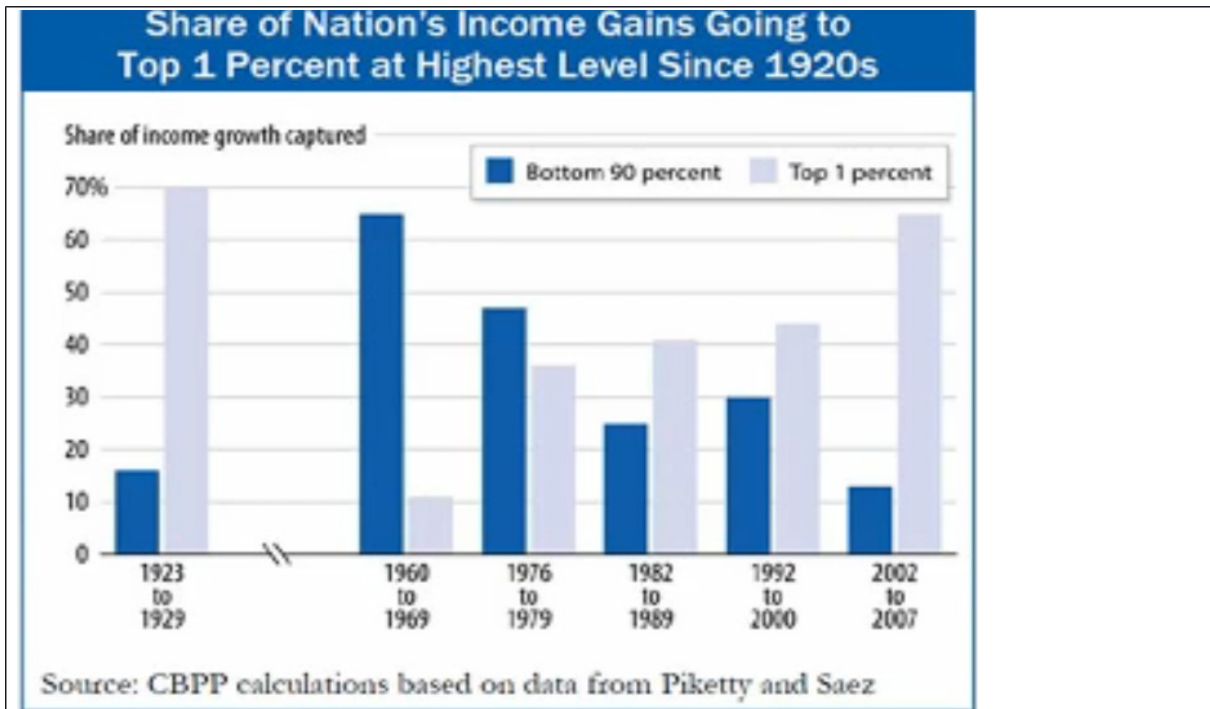


Figure x: xxx

In Marshall’s book, the second scenario for automation, robot deployment and wealth distribution is called “The Australia Project.” Here, machines are a public infrastructure, owned by everyone and servicing everyone, with income gains redistributed in society. The political and social setup in the Australia project is obviously far from what we’re heading towards today, at least in Western societies. But it is an interesting vision not only because of its economic setup, but also because of the role robotics and automation play in the lives of people.

In the Australia Project, the role of machines is close to what Havelock Ellis, a British psychologist and author, once expressed (1922): “The greatest task before civilization at present is to make machines what they ought to be, the slaves, instead of the masters of men.” In fact, machines could serve and relieve people. And the time and energy gains realized by delegating work to machines could free people to concentrate on tasks that they enjoy more than today’s jobs. For example, people could spend more time with family and friends, innovate, learn or become active in the community if they had the time to do so and received a good unconditional income (an income that could be paid out as a result of *machines’* productivity and not humans’ own work) In such a positive scenario, people have the chance to engage in *meaningful* activities and by doing so maintain and build their self-respect in new ways. What they are “doing” then may be more meaningful to them than what they are doing now as they work for wages. At least, this is the vision of Marshall’s Australia Project.

I want to return to value theory and the value pyramid I described in chapter x now. If we deprive people of meaningful work and responsibilities, because we replace them with machines, then we deprive them not only of their financial basis to live, but also of their basis for self-respect, self-esteem and flourishing. Great thinkers have shared in this thinking before me: “The lack of...the opportunity for meaningful work and occupation is destructive...of citizens’ self-respect,” wrote John Rawls once in his *Political Liberalism*.⁴⁾

And Jeffrey Moriarty added: “...to have self-respect, people must contribute to society. People contribute through work. So, if people lack access to meaningful work, then they may fail to contribute, and their self-respect may be damaged.”⁵ In the face of automation, we must reshape our notion of meaningful work and must insure that access to meaningful “doing” is ensured for everyone as well as the financial basis for it.

- 1) BRAIN, M. 2012. Manna: Two Visions of Humanity's Future. BYG Publishing Inc.
- 2) FREY, C. B. & OSBORNE, M. A. 2013. The Future of Employment: How Susceptible are Jobs to Computerization. Oxford: University of Oxford. (Report)
- 3) BRYNJOLFSSON, E. & MCAFEE, A. 2012. Race Against The Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy, Lexington, Massachusetts, USA, Digital Frontier Press.
- 4) RAWLS, J. 2005. Political Liberalism. *Columbia Classics in Philosophy*. New York: Columbia University Press. (originally published in 1921)
- 5) MORIARTY, J. 2009. Rawls, Self-Respect, and the Opportunity for Meaningful Work. *Social Theory and Practice*, 35, 441-459.

How the Privacy Chameleon is Woven into the Value Fabric

In 2006 Daniel Solove, an American legal scholar, published an extensive taxonomy of privacy. Over 84 pages, he explained the concept of privacy by reviewing more than a hundred years of legal case studies around privacy harms in the US. Based on this analysis, he summarized privacy issues as shown in figure x. He concluded that privacy is a “chameleon-like word” ([Solove 2006](#)). No one can define it precisely while covering all its facets. The term is relevant in so many contexts that when it comes to machine age computing, “privacy seems to be about everything” and therefore “to some it appears to be nothing” (p. 479). Julie Cohen has warned that the privacy term has “an image problem” (([Cohen 2012](#)), p. 1903).

While writing this book, I have come to agree with these viewpoints in a very specific way: I fear that a chapter on privacy in this book would have been a chapter about everything and nothing, because privacy issues are instrumental to almost all the intrinsic values that I have covered here. Privacy is ‘everywhere’. But if I had written a detailed chapter on all of the privacy issues, none of the other values would have received the degree of attention they actually deserve. Finally it is these other values though, knowledge, freedom, security, trust friendship and dignity that people care most about and care about universally. I have therefore explained how various concrete privacy dimensions (such as informed consent or surveillance) come into play when these ultimate intrinsic values are at stake. I now want to shortly recapitulate these findings, summarizing once more how privacy issues come into play at various levels of the value pyramid.

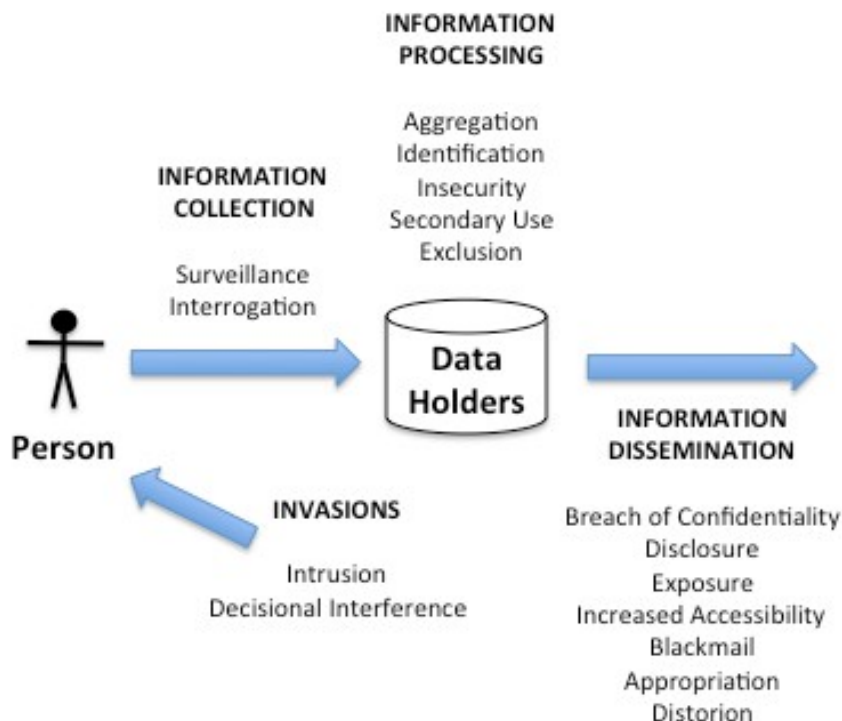


Figure x: Privacy Issues as summarized by Daniel Solove (([Solove 2006](#)), p. 490)

Privacy in Ethical Knowledge Creation

When we align Solove's privacy harms with the value pyramid (see figures x and y), we find that most of the privacy issues he identified from US legal history arise from knowledge being created about people. Allen Westin called this kind of privacy "information privacy" ([Westin 1967](#)).

Information privacy harm can be caused by *increased accessibility*. Increased accessibility means that public personal data is easier to access through the Web today than it was in the past. If this accessibility is not handled in a careful way, a person's reputation can be damaged. Take the case of Mario Costeja González, who filed a lawsuit against Google in 2010. González accused the company of using its search service to publicize the fact that he had failed to pay social security debts in 1998. González asked Google to not display his behavior from the 1990s because the incident occurred over a decade ago. He wanted the incident to be forgotten, arguing that it damaged his reputation. The European Court of Justice supported Mr. González. Note that in section x (figure x), I described how a technical system (like Google's) can create transparency. Increased accessibility is the result of transparency. However, providing transparency in an *ethical* way means that only meaningful and appropriate information is published, not any information that someone or something can acquire (see figure x and section y). If Google's search engine had been optimized technically to provide meaningful and appropriate information about Mr. Gonzalez, meaning for instance that the information should be timely, the company would probably not have been sued.

Interrogation is another form of privacy harm. The term originally referred to pressuring individuals to divulge information. Interrogation is different from surveillance in that it occurs with the conscious awareness of the subject and is not clandestine. Requesting to use customers' detailed personal information in the context of a service contract, combined with a denial of service if that information is not provided, can be considered a modern form of interrogation. Some national data protection laws therefore foresee a prohibition of coupling service accessibility with personal data provisioning. This prohibition of service coupling combined with informed consent procedures (section x) can create ethical information collection practices. Through informed consent that is voluntary people can maintain control over data collection both technically and psychologically (figure x).

Solove outlines how a threat to privacy can be created through *aggregation* of personal information and how *distortion* of a person's image can result from analyzing aggregated data. Data can be aggregated legitimately if the data aggregator gets a person's explicit informed consent and keeps the data under her control (such as agent Arthur accumulating data about Sophia). But distortion can still occur if mistakes are being done in the aggregation process. A real challenge for companies is therefore that their data quality needs to be very good for aggregation purposes (section x). Yet, as of 2015, data quality was not always good enough leading to distortion of people's image during aggregation processes (section x). Against this background, distortion can be seen as a transparency issue. Transparency aims to reveal 'truth' (section x) and avoid confusion, distortion and pain as a result of unobserved errors. When data is aggregated in a transparent way, there is less risk of distortion of truth, because a community of people can potentially look into the quality of data and aggregation practices.

A second way to mitigate the negative consequences of distortion is to anonymize or pseudonymize the original data and prohibit its linking to specific individuals altogether (box x).

Finally, privacy is about the ethical use of knowledge about people. I have described how *unauthorized secondary uses of data*, *a breach of confidentiality of information*, *exposure*, *public disclosure* and *appropriation of data* cause privacy harms (section x). Helen Nissenbaum’s concept of “contextual integrity” can be used to think about the ethics and legitimacy of information flows that extend beyond an agreed context, allowing for such harms to happen. Beyond contextual integrity, I also described in section x how and why people consider some data uses to be unfair (figure x) and how data may be abused to create bias (figure y). Solove analyzed the US Fair Information Practices (FIP) and how they prohibit what he calls “exclusion.” *Exclusion* is avoided in the US by respecting three related FIP transparency principles: (1) the existence of record systems cannot be kept secret; (2) an individual must be able to “find out what information about him is in a record and how it is used”; and (3) an individual must be able to “correct or amend a record of identifiable information about him” (([Solove 2006](#)), p. 521).

Privacy and Freedom

Privacy dimensions are not only an issue for ethical knowledge creation and use. They are also relevant for our freedom. In particular, surveillance as a special kind of privacy harm can undermine our freedom. It is the “right to be let alone” that is being harmed here (in the negative libertarian sense) ([Warren et al. 1890](#)).

A positive libertarian reason why surveillance has been said to reduce freedom is that it makes people behave in a restrained way. Scholars argue that we feel or are aware of being watched, and we therefore adapt our behavior to the expectations of those who watch us. The origin or source of our actions is therefore not our free will any more but the presumed expectations of our guards. I outlined this line of arguments in section x. However, there is little empirical proof that we do in fact feel consciously constrained in our actions due to machine surveillance (such as video cameras). As long as surveillance data is not notably used against a large part of a country’s population, people seem to accept the practice. Most people do not perceive the surveillance infrastructure as a threat. In contrast, some people seem to desire to be seen even. A recent advertising video by a major cosmetics company flirted with the idea that a pretty woman is ‘admired’ by a shop’s surveillance camera.

One reason why the Panopticon effect of surveillance (section x) is not obvious is because people systematically underestimate risks ([Kahneman et al. 2000](#)). “It will not happen to me,” is a typical statement. We all think that we personally won’t be negatively impacted by the surveillance infrastructure. If people are informed about surveillance they argue that they have nothing to hide and therefore don’t care about being watched. They don’t recognize the scope of today’s surveillance infrastructure, which they massively underestimate ([Bizer et al. 2006](#)). Against the background of these arguments, I hypothesize that the positive liberty of most ordinary citizens is not currently infringed by surveillance to the extent often argued. If we don’t consciously feel the grasp of our invisible manacles, then our liberty is not affected. After all, liberty requires consciousness. This argument is, of course, no justification for building the manacles in the first place. If it is not for freedom of thought that we should avoid surveillance infrastructure, it is certainly for the reason of avoiding power asymmetries between governments and citizens. In box x, I described how we should strive for a more balanced planning of surveillance infrastructure in places where we really want it (like in dark parking lots at night). Such wise planning is a matter of leadership and foresight on the side of infrastructure investors.

Our freedom is, however, strongly impacted in another way: Machines by now control a large

part of our attention and hence free thought. This privacy harm is called “*decisional interference*” and “*intrusion*” by Solove. In section x, I outlined how IT push architectures for messaging services lead to constant interruptions of our activities. As of 2015, people can hardly finish a train of thought without being interrupted by some kind of pop-up window, advertising display, or other form of incoming communication. We can hardly control this constant inflow of attention-grabbing machine messages. Attention-sensitive design of machines is therefore a highly relevant form of ethical computing.

Attention-sensitive systems are built on the idea of information “pull” instead of “push”. Pull architectures for messaging and information retrieval are not only a means for a renewed “right to be let alone” ([Warren et al. 1890](#)) though. Pull architectures for information search on products and services would also allow us also to better compare information and freely make up our minds around our own interests. In an ethical machine design, this free thinking would take the place of today’s setup, where we are kept in filter bubbles (section x) and bombarded with predictive advertising messages or search results.

Privacy Trade-offs at all Levels of the Value Pyramid

Privacy comes into play in various forms and guises when we relate it to the values at different levels of the pyramid. Unfortunately, however, the desire for privacy often seems to be accompanied by some value trade-off. Take the case of using health data to better monitor patients’ medical history. Patient monitoring is not only done for a person’s proper benefit but also for higher social reasons. Large pools of health data offer better insights into the paths illnesses can take, allow us to watch the geographic spread of diseases, to share experiences about the performance of doctors and hospitals, to understand human genetics and more. In section x, I described some of the benefits the health industry expects to get by collecting and sharing health data. At the same time, health data is highly sensitive personal data. It is sensitive not only because of its bodily intimacy but also because of its extraordinary potential for misuse. If health data got into the wrong hands, unfair treatment and bias could become a norm for people in all kinds of life situations, from looking for health insurance to searching for a new job. So is it good to collect, aggregate and use people’s health information?

At the next higher level of the pyramid, we see the widely discussed trade-off between privacy and security. Most governments and many fearful individuals argue that public security demands surveillance. There is a strong belief that surveillance infrastructure impedes crime and facilitates crime conviction. At the same time, the surveillance infrastructure undermines our right to be let alone and creates power asymmetries. As I outlined above, many argue that surveillance undermines our positive liberty to speak and act as freely as we would without being observed. And so people ask: Should we give up some privacy to promote public safety?

Then comes friendship. In section x, I showed how anonymity and invisibility increase online deliberation. People are less inhibited when they can shelter their identities. Many open up more and tell more secrets. Perhaps they can get closer to their true selves if they can be anonymous than when they are identified. But at the same time, real friendship requires identification. True reciprocity, feedback and learning from others can occur only when people know each other for real. So how can virtual worlds strike the right balance between identified selves and anonymous encounters? Should they take measures in favor of one form of self-representation? To what extent should virtual world operators themselves know about the true identities of their players? On one hand, virtual world operators should know the ‘true names’ of their players so that they can maintain order in the virtual world when players abuse

their anonymity. On the other hand, the very fact of being completely anonymous - even to the service operators - allows for true 'online deliberation' (section x). So what is more sensible from an operator's perspective: to maintain access to players or to allow them to be completely unobserved and open up?

Finally, dignity and respect have potential privacy trade-offs. If people are fully respected, others should not be systematically surveying them. However, to build evaluative respect into machines – to make them 'polite' – the machine must be able to monitor people's preferences and try to understand them.

There is no easy answer for how we can resolve these trade-offs. In some cases we do not even know whether they truly exist beyond theory. Take the example of public surveillance: So far, we don't have large-scale data to prove that surveillance infrastructure actually reduces and predicts serious crime. If we had such data, we could analyze how to scale surveillance to minimize crime while maximizing people's privacy. As I will show, this scaling is a process that is highly context specific. It is a process in which the real threats in a given context, the probabilities of these threats and the amount of potential damage are combined to understand risks; such as the risk of crime. These risks are then addressed through controls and mitigation strategies such as surveillance. The mitigation strategies correspond to the concrete threats identified and evaluated by experts. By comprehensively weighing threats to values and enablers of values, we can resolve trade-offs, identify compromises and thereby take ethical responsibility.

Exercise:

- Depict the value pyramid with all of the values that were discussed in this chapter. Then, align the privacy harms with Solove's value pyramid and discuss whether and how privacy harms may be created at various levels of the pyramid.

Summing Up: Values in the Machines Age

This chapter identified and analyzed a number of core intrinsic values that are shared by all people around the globe. Knowledge, freedom, autonomy, security, health, friendship and dignity are undoubtedly important for everyone to grow and feel good as individuals (figure x). It is therefore of utmost importance that we protect these values in a machine world. Moreover, we should build machines that actively embrace, embed and foster these values.

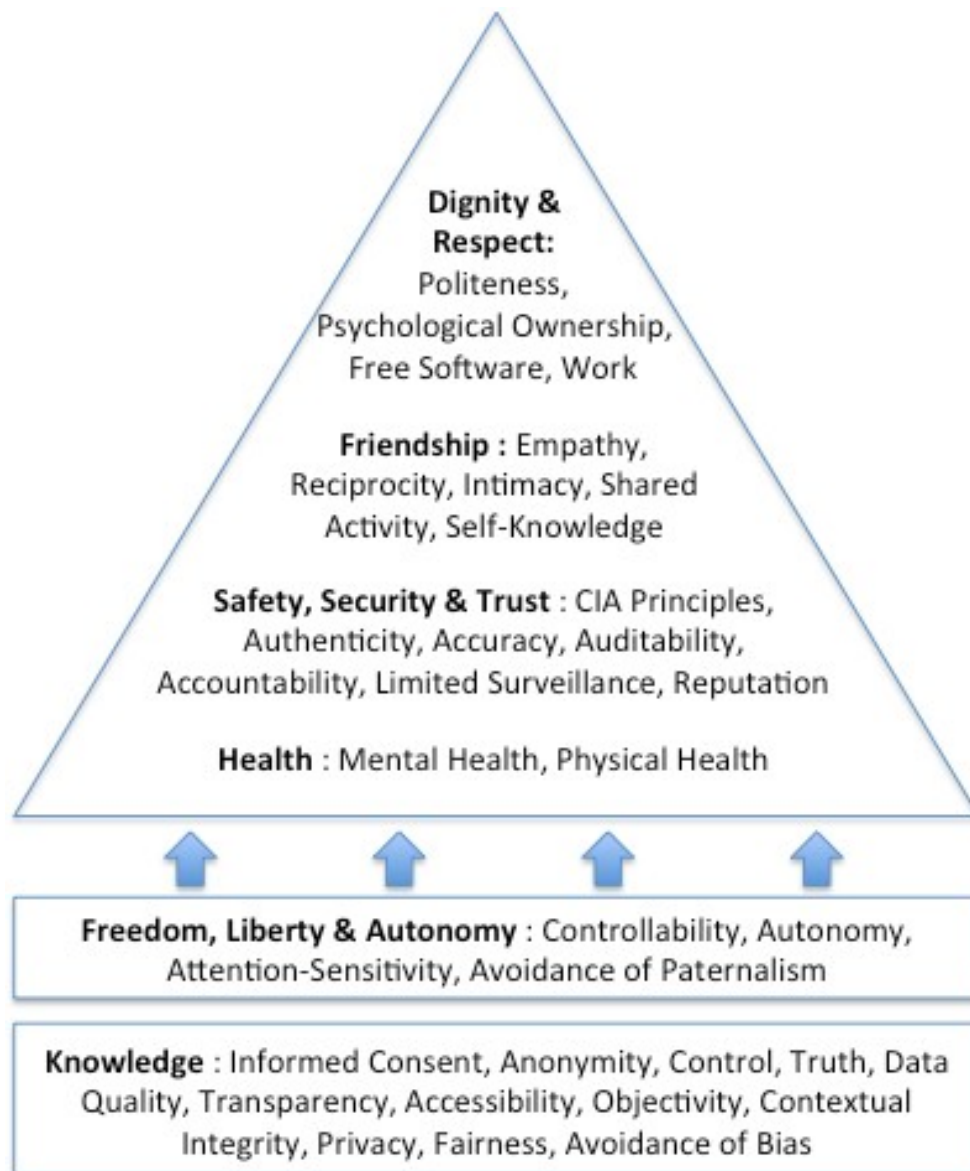


Figure x: A pyramid of values important in the machine age

Knowledge: I began with the knowledge value, outlining how building information for the machine age requires ethical conduct at all stages of the knowledge creation process. For people to trust machines and machine operators, data and information must be collected in a legitimate way. We can do this by implementing technically facilitated procedures for informed consent and by fostering a psychology of control in people around the data collection process. Data quality and transparency are also important for data aggregation. In the machine age, we want machines to build knowledge that we as humans can understand

and hence trust. To accomplish that goal, we need to be sure that the machines release truthful information. By definition, there is no knowledge without truth. However, as of 2015, the data quality and transparency of data-processing activities are a challenge. First-generation machines have not been built to ensure high enough levels of data quality and transparency. Consequently, the knowledge that is aggregated is often not reliable.

One of the reasons that current machine-generated knowledge is not reliable is that the quest for transparency is at an early stage. Transparency means that any knowledge we create is meaningful and appropriate. Yet, determining what is meaningful and appropriate requires judgment. And making good judgments is a trait that only humans have and that machines have yet to 'learn' (if they are ever capable of it...). For machines to make good judgments and learn from humans, the machines must not try to dominate human decisions as much as they do today. Humans are already in filter bubbles that obfuscate and distort the complex reality in which we live. Of course, easy use, time savings, or so-called "efficiency" are achieved when machines simplistically sort the world's information for us and nudge us into decisions. However, as I have outlined, training our judgment is an achievement of enlightenment, and we need to be careful that the machine age does not take this capability away from us. We need to build hybrid machines that help us sort things but leave ample room for self-experimentation and discovery.

Finally, I wrote about the ethical use of information. I presented Helen Nissenbaum's concept of contextual integrity for data, information and knowledge use. In doing so, I covered many privacy issues that arise today in the machine world as a result of ethically dubious data flows: secondary uses of data, unwanted appropriation of personal data, breaches of confidentiality and even exposure have become an unfortunate norm. I therefore expanded on the ethical use of information, noting that we create machine bias when we categorize people and treat them according to such categorizations. Many users welcome personalized information, but this personalization needs to be perceived as fair in order to be trusted in the long run. Current economic rationale in our service designs often tends to prioritize short-term cost minimization and profit maximization over fairness. I encourage reflection on the ethics of such corporate practices.

Liberty, Freedom and Autonomy: Freedom, liberty and autonomy are the philosophical building blocks of our current Western political systems. To ensure that people remain free in the machine age, machines must be built with dynamic levels of automation, allowing people to manipulate and control the machine as needed. The 'industrial model' of total automation that we observe in manufacturing today may not be the ideal solution for consumer-facing computing devices. Here, where market demand determines the success or failure of technology, automation could backfire if it is too paternalistic or too simplistic. Machines must be accessible in such a way that they allow for manipulation on several layers. I distinguished between easy-to-use higher-level access to application layer dynamics and deep access to lower layers of a machine's functioning. Today accessibility or "openness" is noted as a fundamental software freedom. However, with a move to business models that provide software as a service, the fundamental freedom to access is threatened. Also, we must consider how such openness is used. While we are still in the onset of the machine age, openness may sometimes be a way to pander to the curiosity and pride of software engineers, hackers and a youth culture that wants to understand technology. At the same time, however, machine accessibility becomes crucial for balancing power between service providers, their machines and the people.

A challenge in this power-play is the protection of human mental skills and cognitive

capabilities. If we want more than a tiny elite to be able to access, manipulate and control machines, then we need to develop and strengthen our cognitive skills. We must train our cognitive abilities to understand the functioning and limitations of machines. Most importantly, we also need a healthy level of independence in decision-making from our machines. Perhaps we do not always take their advice; we might override them and still feel good about our decision. Yet we will be able to develop these human capabilities only if we have sufficient time and free attention resources. A great digital divide has appeared between the few people who can protect their scarce attention resource, stay focused and make decisions autonomously and those who don't have the mental strength to do so anymore. The digital attention divide can be overcome if we switch to "information pull architectures," which create something that Doc Searls calls "intention economy." In the chapter on attention-sensitive system design, I described how information pull architectures support natural human attention allocation and how careful interruption design can help us to refocus and be less disturbed.

Health: Being able to control our attention is also relevant for our health in the machine age. Many people today suffer from "Problematic Internet Use." Becoming addicted to machines or too absorbed in virtual worlds can lead to not only bodily pain, but also to stress in everyday life followed by mental health problems. Machines could certainly foster our health in many direct and indirect ways. New devices like the Talos suit or life-logging apps may bring people back into nature and motivate them to care for their bodies. I did not provide general guidelines in the health chapter on how to build 'healthy machines.' The subject domain of health is much too broad for that, and every bodily function may have its own supportive machine service at some point. But I did discuss the short-term and long-term effects of machines on our mental and physical health, and I outlined ways in which machines relate to today's phenomenon of burnout.

When we talk about burnout, it becomes clear why Maslow regarded knowledge and freedom as prerequisites for other basic needs in the pyramid. For example, figure X shows how mental health in the form of burnout can be indirectly triggered by a lack of computer self-efficacy and job control. Ethical knowledge creation provides employees with a widely usable and legitimate data base that they can understand (transparency), access and use for fair purposes. This kind of "ethical knowledge," as well as the autonomy to manipulate the machines they use at multiple levels, can foster employees' perception of efficacy and control. Employees can creatively meet the demands of increasingly number-driven jobs. People who feel empowered and in control will probably perceive a healthier balance between the demands of their jobs and their control. In contrast, employees feel out of control when they cannot access the machines they use in their jobs, cannot understand the numbers the machines produce, cannot alter these numbers nor the machines and don't have documentation to understand how the machines function. This negative feeling is exacerbated when employees are forced to use the numbers and the machines they don't understand to meet job demands. The steep increase of burnout in companies today might be caused in part by machines that deprive people of ethical knowledge and autonomy vis-à-vis machines.

Security and Safety: Besides health, another strong motivator to work on an ethical knowledge base for machines is the safety and security of these machines. In security projects, companies work towards more confidentiality, availability, integrity, authenticity and accuracy of their data. They improve the auditability of their systems and take measures for better accountability. By doing so, they actually feed into a process for ethical knowledge creation and knowledge use.

But when ordinary people speak about “security” today, they often mean more than the securing of corporate data assets. In their mind, security is equated with safety. And security is also often equated with surveillance infrastructure. These simplistic equalizations are unfortunate because they lead security investments to be channeled into a surveillance infrastructure. It makes it easy to argue that one has done a lot to improve people’s safety and security by increasing the budget for surveillance. In truth, however, security goals, auditability and accountability are hardly improved by more surveillance. And the safety of an infrastructure, such as the quality and reliability of products, services and assets is not enhanced by surveillance either. I therefore plead for a more stringent and precise use of terms when it comes to security and safety, and I propose a more reasonable, data-driven ‘Golden Mean Process’ to decide on surveillance investments.

Friendship: The last sections of this chapter deal with the social need for friendship and the individual need for dignity and respect. The computer science world has barely addressed these last two human needs even though machines dramatically influence them. Machines alter the way we live and build relationships; three examples are our 1st generation media (i.e. social networks, mobile phones, e-mail, etc.), virtual worlds and interaction with artificial beings. In reviewing these influences, I found that Batya Friedman’s value-sensitive design methodology is limited: We cannot simply build characteristics of friendship into robots or agents. It is ethically problematic to conceptualize and decompose the friendship value and then identify requirements for friendly machines. In contrast, if we build machines that become our friends, we face the ethical issue of replacing human touch with cold, lifeless and uncaring marionettes. These objects may be very attentive and courteous with us. They may be easier to handle than unpredictable human characters, but they also make us accustomed to superficial, conflict-free relationships that are far from the demanding human encounters of the real world. Our ability to develop virtue and learn from the hard feedback of real human friends may be diminished as a result. So building the friendship value into a digital system has the counterintuitive effect of potentially destroying that same value in the offline system.

Many machine ethicists would probably argue that machine friendship is not as lifeless and dangerous as I presented it. First, not all cultures regard machines as lifeless. Some Buddhist cultures embrace the idea that every thing has some spiritual essence, including robots or other lifeless objects. Second, machine ethicists argue that machines can outperform humans in some respects. For example, machines could teach us ethics. Machine ethicists aim to build machines that embed ethical reasoning and that can inform humans about higher forms of philosophical knowledge and conduct. Robots and agents could provide us with a knowledge base that has never been accessible to humans before. Personally, I am not sure whether this vision will deliver on its promise. Can a potential loss of humans’ mutual socialization and self-development be countered by machines that embed ethical protocols? Today, little is known about whether humans’ intrinsic knowledge and learning does not depend on human interaction and empathetic resonance. I reported on the importance of our bodies and their mirror neuron system for creating empathy and truly understanding what is happening in one’s environment. Unless machines embed similarly powerful biological mechanisms, can they ever teach us much?

Higher-level individual needs for respect, self-esteem and power: Finally, in the last section of this chapter, I reflected on how machines can influence our self-respect and the respect we receive from others. We can take a construct like politeness, decompose it (as I did in figure x) and build machines that treat us politely. In fact, politeness would be a great new requirement for engineers to think about since a lot of machines today tend to treat us as cattle rather than humans. But a deeper reflection on higher level needs in the value pyramid

requires moving beyond interaction requirements. We need to think more holistically about the role machines play in human lives. And as we do so, we see that machine design, as well as the business models and legal frameworks created around machines, alter the power balance between people and machine owners. Machines' requirements engineering becomes social engineering.

Lets take the psychological ownership value as an illustrative example: Machine owners can, of course, design their machine services in such a way that they systematically foster psychological ownership perceptions in customers. Figure x gives all the details needed. And this perception benefits self-respect, because people like to 'have and 'possess' things. *But* fostering such a value through service design is a double-edged sword for companies. If customers use machines that strengthen their feelings of ownership, it is hard to then take legal ownership and technical control away from them. Restrictive copyright laws and remotely controlled machine architectures for instance disappoint customers' psychological ownership perceptions in the long run. Just recap the difference between the robot manufacturer Future Lab and Robo Systems described in the scenarios. Here the robots built by Future Lab are completely owned and controlled by their users while those of Robo Systems are remotely controlled. Which one of the two companies seems more attractive to us as customers if all other performance variables are kept constant? Personally, I would probably opt for the Future Lab robots that I could fully own and control. But is this the solution companies will prefer? If they truly cater machines' and business models' design to higher level individual needs such as ownership, user control and power they will certainly gain competitive edge and market share. But as they do so, they will also need to give up some of *their* control and power over the machines and their users. They will need to forgo personal data assets and knowledge on customers. Will they do so? What will be the values that will drive IT companies' investment logic and requirements engineering in the future? I suggest that the struggle for power and control could become an important driver for IT companies' requirements engineering; more so potentially than financial benefits. And at that point, people's higher needs risk to clash with the established corporate machine world. Extremely wise leadership is needed at that point.

References

- Aamodt, A., and Nygard, M. 1995. "Different roles and mutual dependencies of data, information, and knowledge - an AI perspective on their integration," *Data and Knowledge Engineering*, (16:3), pp 191-222.**
- Abramson, L.Y., Seligman, M.E.P., and Teasdale, J.D. 1978. "Learned helplessness in humans," *Journal of Abnormal Psychology* (87:1), 1978, pp 49-74.**
- Adams, J. 1963. "Toward an understanding of inequity," *Journal of Abnormal Psychology* (67), pp 422-436.**
- Alhadeff, J., Van Alsenoy, B., and Dumortier, J. 2011. "The accountability principle in data protection regulation: origin, development and future directions," *Privacy and***

Accountability, Berlin, Germany.

- Allen, C., Varner, G., and Zinser, J. 2000. "Prolegomena to any future artificial moral agent," *Journal of Experimental and Theoretical Artificial Intelligence* (12), pp 251-261.**
- Altman, I. 1975 *The environment and social behavior: Privacy, personal space, territory, crowding* Monterey, California, USA, Brooks/Cole.**
- Altman, I., and Taylor, D. 1973 *Social Penetration: The Development of Interpersonal Relationships* New York, Holt, Rinehart & Winston.**
- Anderson, S.L. 2011. "How Machines Might Help Us Achieve Breakthroughs in Ethical Theory and Inspire Us to Behave Better," in: *Machine Ethics*, M. Anderson and S.L. Anderson (eds.), New York, Cambridge University Press.**
- Angwin, J. 2012 "Online Tracking Ramps Up - Popularity of User-Tailored Advertising Fuels Data Gathering on Browsing Habits.," in: *Wall Street Journal*, New York.**
- Aristoteles "Nikomachische Ethik," Reclam Verlag, Stuttgart.**
- Ashcroft, R.E. 2005. "Making sense of dignity," *Journal of Medical Issues* (31:11), pp 679-682.**
- Autonome Provinz Südtirol 2002. "Elektromagnetische Strahlung und Gesundheit," Autonome Provinz Südtirol, Bozen.**
- Averill, J.R. 1973. "Personal control over aversive stimuli and its relationship to stress," *Psychological Bulletin* (80), pp 286-303.**
- Bailenson, J.N., Iyengar, S., Yee, N., and Collins, N.A. 2008. "Facial Similarity between Voters and Candidates Causes Influence," *Public Opinion Quarterly* (72:5), pp 935-961.**
- Bandura, A. 1977. "Self-efficacy: Toward a unified theory of behavioral change," *Psychological Review* (84), 1977, pp 191-215.**
- Bartram, L., Ware, C., and Calvert, T. 2003. "Moticons: Detection, distraction and task," *International Journal of Human-Computer Studies* (58:5), pp 515-545.**
- Beattie, A.E., and Mitchell, A.A. 1985. "The relationship between advertising recall and persuasion: An experimental investigation," in: *Psychological processes and advertising effects: Theory, research, and application*, L.F. Alwitt and A.A. Mitchell (eds.), Hillsdale, NJ, Lawrence Erlbaum, pp. 129-155.**
- Belk, R.W. 1988. "Possessions and the Extended Self," *Journal of Consumer Research* (15:2), pp 139-168.**
- Bereczkei, T., Hegedus, G., and Hajnal, G. 2009. "Facialmetric similarities mediate mate choice: sexual imprinting on opposite-sex parents," *Proceedings of the Royal Society* (276:1654), pp 91-98.**
- Berlin, I. 1969. "Two Concepts of Liberty," in: *Four Essays on Liberty*, I. Berlin (ed.), Oxford, Oxford University Press.**
- Bessie` re, K., Kiesler, S., Kraut, R., and Boneva, B.S. 2008. "Effects of Internet Use and Social Resources on Changes in Depression," *Information, Communication & Society* (11:1),**

pp 47-70.

- Billings, C.E. 1991. "Human-centered aircraft automation: a concept and guidelines," NASA Ames Research Center, Moffett Field, CA, USA.
- Bizer, J., Günther, O., and Spiekermann, S. 2006. "TAUCIS - Technikfolgenabschätzungsstudie Ubiquitäres Computing und Informationelle Selbstbestimmung," Humboldt University Berlin, Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein (ULD), Berlin, Germany.
- Black, M.J. 2002 "The Art of Code," University of Pennsylvania.
- Botsman, R., and Rogers, R. 2014 *What's Mine Is Yours: The Rise of Collaborative Consumption* New York, Harper Collins
- Brehm, J.W. 1966 *A Theory of Psychological Reactance* New York, USA, Academic Press.
- Brey, P. 2004. "Disclosive Computer Ethics," in: *Readings in CyberEthics*, R.A. Spinello and H.T. Tavani (eds.), Sudbury, MA, Jones Bartlett Learning, pp. 55-66.
- Briggle, A. 2008. "Real friends: how the Internet can foster friendship," *Ethics and Information Technology* (10:1), pp 71-79.
- Brumby, D.P., Howes, A., and Salvucci, D.D. 2009. "Focus on driving: How cognitive constraints shape the adaptation of strategy when dialing while driving," ACM Conference on Human Factors in Computing Systems (CHI 2009), ACM Press, Boston, Massachusetts, USA, pp. 1629-1638.
- Buxton, W. 1986. "Chunking and phrasing and the design of human-computer dialogues " FIP 10th World Computer Congress, North-Holland/IFIP, Dublin, Ireland.
- Bynum, T.W. 2006. "Flourishing Ethics," *Ethics and Information Technology* (8:4), pp 157-173.
- Campbell, M. 1999. "Perceptions of Price Unfairness: Antecedents and Consequences," *Journal of Marketing Research* (36:5), May 1999.
- Caplan, S., Williams, D., and Yee, N. 2009. "Problematic Internet use and psychosocial well-being among MMO players," *Computers in Human Behavior* (25:6), pp 1312-1319.
- Carter, I. 2012 "Positive and Negative Liberty," in: *The Stanford Encyclopedia of Philosophy*, E.N. Zalta (ed.), The Metaphysics Research Lab Stanford.
- Casassa Mont, M., Pearson, S., and Bramhall, P. 2003. "Towards Accountable Management of Identity and Privacy: Sticky Policies and Enforceable Tracing Services," HP Laboratories Bristol.
- Cavoukian, A. 2011 *Privacy by Design...Take the Challenge*, Information and Privacy Commissioner of Ontario, Canada.
- Chellappa, R.K., and Pavlou, P.A. 2002. "Perceived information security, financial liability and consumer trust in electronic commerce transactions," *Logistics Information Management* (15:5), p 358±368.
- Cherdantseva, Y., and Hilton, J. 2013. "A Reference Model of Information Assurance & Security," 8th International

- Conference on Availability, Reliability and Security (ARES),
IEEE, Regensburg, pp. 546 - 555
- Chiauzzi, E., Brevard, J., Thum, C., Decembrele, S., and Lord, S.
2008. "MyStudentBody-Stress: an online stress management
intervention for college students," *Journal of Health
Communication* (13:8), p 827.
- Chopra, S., and Dexter, S. 2009. "The freedoms of software and its
ethical uses," *Ethics and Information Technology* (11:4), pp
287-297.
- Christman, J. 1991. "Liberalism and Individual Positive Freedom,"
Ethics (101:2), pp 343-359.
- Churchesa, O., Nichollsa, M., Thiessenb, M., Kohlerc, M., and
Keagec, H. 2013. "Emoticons in mind: An event-related
potential study," *Social Neuroscience* (9:2), pp 196-202.
- Clarke, R. 1988. "Information Technology and Dataveillance "
Communications of the ACM (31:5), pp 498-512.
- CNSS 2010 "National Information Assurance (IA) Glossary,"
Committee on National Security Systems, Fort Meade,
Maryland, US.
- Cocking, D., and Matthews, S. 2000. "Unreal friends," *Ethics and
Information Technolgoy* (2:4), pp 223-231.
- Cohen, J.E. 2012. "What Privacy Is For," *Harvard Law Review* (126),
pp 1904-1933.
- Coleman, E.G. 2013 *Coding Freedom - The Ethics and Aesthetics of
Hacking* Princeton, Princeton University Press.
- Council of Europe 1950 "European Convention on Human Rights,"
E.C.o.H. Rights (ed.), Rome.
- Cox, J. 2001. "Can differential prices be fair?," *The Journal of
Product and Brand Management* (10:4), pp 264-276.
- Cranor, L. 2012. "Necessary But Not Sufficient: Standardized
Mechanisms for Privacy Notice and Choice," *Journal of
Telecommunications and High Technology Law* (10:2), pp
273-308.
- Cranor, L.F., Dobbs, B., Egelman, S., Hogben, G., Humphrey, J., and
Schunter, M. 2006 "The Platform for Privacy Preferences 1.1
(P3P1.1) Specification - W3C Working Group Note 13
November 2006," R. Wenning and M. Schunter (eds.), World
Wide Web Consortium (W3C) - P3P Working Group.
- Dabbish, L., Mark, G., and Gonzàlez, V.M. 2011. "Why do I keep
interrupting myself?: environment, habit and self-
interruption," Conference on Human Factors in Computing
Systems (CHI'11), ACM Vancouver, Canada, pp. 3127-3130.
- Danezis, G., Kohlweiss, M., Livshits, B., and Rial, A. 2012. "Private
Client-side Proling with Random Forests and Hidden Markov
Models," 12th International Symposium on Privacy
Enhancing Technologies (PETS 2012), Springer, Vigo, Spain,
pp. 18-37
- Daston, L., and Galison, P. 2007 *Objectivity* New York, Zone Books.
- Davis, R.A. 2001. "A cognitive-behavioral model of pathological
internet use," *Computers in Human Behavior* (17:2), pp 187-
195.

- Derlega, V.J., Metts, S., Petronio, S., and Margulis, S.T. 1993 **Self-Disclosure** Newbury Park, CA, Sage.
- Diao, F., and Sundar, S.S. 2004. "Orienting response and memory for web advertisements: Exploring effects of pop-up window and animation," *Communication Research* (31:5), pp 537-567.
- Dillon, R.S. 2010. "Respect for persons, identity, and information technology," *Ethics and Information Technology* (12), pp 17-28.
- Dreyfus, H.L. 2009 *On the Internet* New York, Routledge.
- Endsley, M.R. 1996. "Automation and Situation Awareness," in: *Automation and Human Performance - Theory and Application*, R. Prasuraman and M. Mouloua (eds.), New Jersey, USA, Lawrence Erlbaum Associates, pp. 163-181.
- Ess, C. 2013 *Digital Media Ethics*, (2nd edition ed.) Hoboken, NJ, USA, Wiley.
- Etzioni, A. 1991. "The Socio-Economics of Property," *Journal of Social Behaviour and Personality* (6:6), pp 465-468.
- European Commission 2001. "Promoting a European framework for Corporate Social Responsibility," Office for Official Publications of the European Commission, Luxembourg.
- European Network and Information Security Agency (ENISA) 2011. "To Log Or Not To Log? Risks and benefits of emerging life-logging applications," European Network and Information Security Agency (ENISA), Athen.
- European Parliament and the Council of Europe 1995 "Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of Individuals with regard to the processing of personal data and on the free movement of such data," in: L 281/31, Official Journal of the European Communities.
- Fischer, G. 2012. "Context-Aware Systems—The 'Right' Information, at the 'Right' Time, in the 'Right' Place, in the 'Right' Way, to the 'Right' Person," *Advanced Visual Interfaces (AVI '12)*, ACM, Capri Island, Italy.
- Fitts, P.M. 1951 *Human Engineering for an Effective Air-Navigation and Traffic-Control System* Columbus, Ohio, USA.
- Floridi, L. 2005. "Is semantic information meaningful data?," *Philosophical and Phenomenological Research* (LXX:2).
- Floridi, L., and Sanders, J.W. 2004. "On the morality of artificial agents," *Journal of Minds and Machines* (14:3), pp 349 - 379
- Franck, G. 1998 *Ökonomie der Aufmerksamkeit* München Wien, Carl Hanser Verlag.
- Frankena, W. 1973 *Ethics*, (2nd ed.) New Jersey, USA, Prentice-Hall.
- Frankfurt, H. 1971. "Freedom of the Will and the Concept of a Person," *Journal of Philosophy* (68:1), pp 5-20.
- Friedman, B., Felten, E., and Millett, L.I. 2000. "Informed Consent Online: A Conceptual Model and Design Principles," University of Washington, Washington, USA.
- Friedman, B., and Nissenbaum, H. 1996. "Bias in Computer

- Systems," *ACM Transactions on Information Systems* (14:3), pp 330-347.**
- Friedman, B., and Nissenbaum, H. 1997. "Software Agents and User Autonomy," *Autonomous Agents 97*, ACM, Marina Del Rey, California, USA, pp. 466-469**
- Frings, M.S. 1966. "Der Ordo Amoris bei Max Scheler. Seine Beziehungen zur materialen Wertethik und zum Ressentimentbegriff," *Zeitschrift für philosophische Forschung* (20:1), pp 57-76.**
- Fromm, E. 1976 "To have or to be," Harper & Row, New York.**
- FTC, F.T.C. 2000 "Fair Information Practice Principles," F.T. Commission (ed.).**
- Fujitsu 2010. "Personal data in the cloud: A global survey of consumer attitudes," Tokyo, Japan.**
- Furby, L. 1978. "Possessions: Toward a theory of their meaning and function throughout the life cycle," in: *Life Span Development and Behavior*, P.B. Baltes (ed.), New York, Academic Press, pp. 297-336.**
- Gefen, D., Elena, K., and Straub, D.W. 2003. "Trust and TAM in online shopping: An integrated model," *MIS Quarterly* (27:1), pp 51-90.**
- Gentry, C. 2009. "Fully homomorphic encryption using ideal lattices," 41st ACM Symposium on Theory of Computing (STOC '09), ACM, Bethesda, MD, USA, pp. 169-178.**
- González, V.M., and Mark, G. 2004. "Constant, Constant, Multi-tasking Crazy: Managing Multiple Working Spheres," Conference on Human Factors in Computing Systems (CHI'04), ACM, Vienna, Austria, pp. 113-120**
- Green, D.E., Walkey, F.H., and Taylor, A.J.W. 1991. "The three-factor structure of the Maslach Burnout Inventory," *Journal of Social Behavior and Personality* (6), pp 453-472.**
- Greenleaf, G. 2011. "Global data privacy in a networked world," in: *Research Handbook of the Internet*, Cheltenham, Edward Elgar.**
- Grodzinsky, F.S., Miller, K.W., and Wolf, M.J. 2011. "Developing artificial agents worthy of trust: "Would you buy a used car from this artificial agent?," *Ethics and Information Technology* (13:1), pp 17-27.**
- Guenther, O., and Spiekermann, S. 2005. "RFID and Perceived Control - The Consumer's View," *Communications of the ACM* (48:9), September 2005, pp 73-76.**
- Halpern, S. 2011 "Mind Control & the Internet," in: *The New York Review of Books*, New York.**
- Hampton, K.N., Sessions, L.F., Ja Her, E., and Rainie, L. 2009. "Social Isolation and New Technology - How the internet and mobile phones impact Americans' social networks."**
- Hastak, M., and Mazis, M.B. 2011. "Deception by Implication: A Typology of Truthful but Misleading Advertising and Labeling Claims," *Journal of Public Policy & Marketing* (30:2), pp 157-167.**
- Heidegger, M. 2004 *Wegmarken* Frankfurt am Main, Vittoria**

Klostermann Verlag.

- Helm, B. 2013 "Friendship," in: *The Stanford Encyclopedia of Philosophy*, The Metaphysics Research Lab Stanford.
- Herzberg, F. 1968. "One More Time: How Do You Motivate Employees?," *Harvard Business Review* (46:1), pp 53-62.
- Hess, C., and Ostrom, E. 2006 *Understanding the Knowledge Commons* Cambridge, US, MIT Press.
- Hilty, L., Som, C., and Köhler, A. 2004. "Assessing the Human, Social, and Environmental Risks of Pervasive Computing," *Human and Ecological Risk Assessment* (10:5), pp 853-874.
- Hoff, J. 2013 *The Analogical Turn: Rethinking Modernity with Nicholas of Cusa* Cambridge, UK, William B. Eerdmans Publishing Company.
- Hofstede, G. 1980 *Culture's Consequences - International Differences in Work-Related Values* Newbury Park, Sage Publications.
- Huang, C. 2010. "Internet Use and Psychological Well-being: A Meta-Analysis," *Journal of Cyberpsychology, Behavior and Social Networking* (13:3), pp 241-249.
- Hume, D. 1748 *Enquiry Concerning Human Understanding*.
- Ichikawa, J.J. 2012 "The Analysis of Knowledge," in: *The Stanford Encyclopedia of Philosophy*, E.N. Zalta (ed.), The Metaphysics Research Lab Stanford.
- IEEE 2014 "Printing Body Parts - A Sampling of Progress in Biological 3D Printing," in: *IEEE Life Sciences*, IEEE, Piscataway, New Jersey.
- ISO 2011 "EN ISO 10218 - 1: Robots for industrial environments — Safety requirements (Part 1)," ISO.
- ISO 2012 "ISO/IEC 15408 - Common Criteria for Information Technology Security Evaluation ".
- ISO 2014a "ISO/IEC 27000 Information technology — Security techniques — Information security management systems — Overview and vocabulary," International Organization for Standardization.
- ISO 2014b "ISO/IEC 29100: Information Technology - Security techniques - Privacy architecture framework," DIN Deutsches Institut für Normung e.V.
- Itrona, L.D., and Nissenbaum, H. 2000. "Shaping The Web: Why The Politics Of Search Engines Matters," *Information Society* (16:3), pp 169-185.
- Jabbi, M., Swart, M., and Keysersa, C. 2007. "Empathy for positive and negative emotions in the gustatory cortex," *NeuroImage* (34:4), pp 1744-1753.
- Jackson, T.W., Dawson, R., and Wilson, D. 2003. "Understanding email interaction increases organizational productivity," *Communications of ACM* (46:8), pp 80-84.
- James, W. 1890 *The Principles of Psychology - Volume 1*, (2007 ed.) New York, Cosimo Classics.
- Jaspers, K. 1973 *Philosophie I - Philosophische Weltorientierung* Berlin, Springer Verlag.
- Jonas, H. 1979 *Das Prinzip Verantwortung - Versuch einer Ethik für*

- die technologische Zivilisation, (Edition 2003 ed.) Frankfurt am Main, Suhrkamp Taschenbuch Verlag.**
- Kahneman, D., and Tversky, A. 2000 *Choices, Values, and Frames* New York, Cambridge University Press.
- Kant, I. 1784 "Beantwortung der Frage: Was ist Aufklärung," in: *Berlinische Monatsschrift*, Berlin, Germany, pp. 481-494.
- Kant, I. 1795 *Kant's Principles of Politics - Including his Essay on Perpetual Peace*, (1891 ed.) Edingburgh, T & T. Clark.
- Kaptein, M. 2004. "Business codes of multinational firms: What do they say?," *Journal of Business Ethics* (50:1), pp 13-31.
- Karasek, R. 1979. "Job Demands, Job Decision Latitude, and Mental Strain: Implications for Job Redesign," *Administrative Science Quarterly* (24:2), pp 285-308.
- Karasek, R. 1990. "Lower health risk with increased job control among white collar workers.," *Journal of Organizational Behavior* (11:3), pp 171-185.
- Kaye, J., Whitley, E.A., Lund, D., Morrison, M., Teare, H., and Melham, K. 2014. "Dynamic consent: a patient interface for twenty-first century research networks," *European Journal of Human Genetics*), pp 1-6.
- Khabsa, M., and C.L., G. 2014. "The Number of Scholarly Documents on the Public Web," *PLOS ONE*).
- Gluckhohn, C. 1962. "Values and Value-Orientations in the theory of action: an exploration in definition and classification," in: *Toward a general theory of action*, T. Parsons, E.A. Shils and N.J. Smelser (eds.), Cambridge, Massachusetts, Transaction Publishers, pp. 388-433.
- Ko, A.J., Abraham, R., Beckwith, L., Blackwell, A., Burnett, M., Erwig, M., Scaffidi, C., Lawrance, J., Lieberman, H., Myers, B., Rosson, M.B., Rothermel, G., Shaw, M., and Wiedenbeck, S. 2011. "The State of the Art in End-User Software Engineering," *ACM Computing Surveys* (43:3).
- Ko, C.-H., Yen, J.-Y., Chen, C.-C., Chen, S.-H., and Cheng-Fang, Y. 2005. "Proposed diagnostic criteria of Internet addiction for adolescents," *Journal of Nervous and Mental Disease* (193:11), pp 728-733.
- Kobsa, A. 2007. "Privacy-Enhanced Personalization," *Communications of the ACM* (50:8), August 2007, pp 24-33.
- Koopmans, F., and Sremac, S. 2011. "Addiction and Autonomy: are Addicts Autonomous?," *Nova prisutnost* (9:1), pp 171-188.
- Krobath, H.T. 2009 *Werte - Ein Streifzug durch Philosophie und Wissenschaft* Würzburg, Königshausen & Neumann.
- Kurzweil, R. 2006 *The Singularity is Near- When Humans Transcend Biology* London, Penguin Group.
- Laibson, D. 1996. "Hyperbolic Discount Functions, Undersaving, And Savings Policy," National Bureau of Economic Research, Cambridge, MA.
- Lamont, J. 1994. "The Concept of Desert in Distributive Justice," *The Philosophical Quarterly* (44:174), pp 45-64.
- Langer, E. 1983 *The Psychology of Control* Beverly Hills, USA, Sage Publications, p. 310.

- Langer, E., and Rodin, J. 1976. "The effects of choice and enhanced personal responsibility for the aged. A field experiment in an institutional setting.," *Journal of Personality and Social Psychology* (34:2), pp 191-198.
- Langer, S. 2009 *Viral Marketing: Wie Sie Mundpropaganda gezielt auslösen und Gewinn bringend nutzen* Wiesbaden, Gabler Verlag.
- Langheinrich, M. 2003. "A Privacy Awareness System for Ubiquitous Computing Environments," 4th International Conference on Ubiquitous Computing, UbiComp2002, Springer-Verlag, Göteborg, Sweden.
- Langheinrich, M. 2005 "Personal Privacy in Ubiquitous Computing - Tools and System Support," in: *Institut für Pervasive Computing*, ETH Zürich, Zürich, CH.
- Lessig, L. 2001 *the future of ideas - the fate of the commons in a connected world* New York, USA, Random House.
- Lind, A.E., and Tyler, T.R. 1988 *The Social Psychology of Procedural Justice* New York, Plenum Press.
- Line, M.B., Nordland, O., Rostad, L., and Tondel, I.A. 2006. "Safety vs. security?," 8th International Conference on Probabilistic Safety Assessment and Management (PSAM), New Orleans, Louisiana, USA.
- MacIntyre, A. 1984 *After Virtue: A Study in Moral Theory*, (2nd Edition ed.) Notre Dame, Indiana, University of Notre Dame Press.
- Maes, P. 1994. "Agents that reduce work and information overload," *Communications of the ACM* (37:7), pp 30-40.
- Maes, P., and Wexelblat, A. 1997 "Issues for Software Agent UI," in: *MIT Media Lab*, Cambridge, USA.
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. 2011. "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute (MGI).
- Marakas, G.M., Yi, M., and Johnson, R. 1998. "The multilevel and multifaceted character of computer self-efficacy: Toward classification of the construct and an integrative framework for research," *Information Systems Research* (9:2), pp 126-163.
- Margulis, S. 2003. "Privacy as a Social Issue and Behavioral Concept," *Journal of Social Issues* (59:2), 2003, pp 243-261.
- Marx, K. 1978 "Capital: A Critique of Political Economy," Penguin, Harmondsworth, England.
- Maslow, A. 1970 *Motivation and Personality*, (2nd edition ed.) New York, Harper & Row Publishers.
- McCrickard, D.S., and Chewar, C.M. 2003. "Attuning notification design to user goals and attention costs," *Communications of ACM* (46:3), pp 67-72.
- McFarlane, D. 2002. "Comparison of four primary methods for coordinating the interruption of people in human-computer interaction," *Human-Computer Interaction* (17:1), pp 63-139.
- McKnight, H.D., and Chervany, N.L. 1996. "The Meaning of Trust,"

University of Minnesota, Minneapolis.

- McLeod, C. 2011 "Trust," in: *The Stanford Encyclopedia of Philosophy*, E.N. Zalta (ed.), The Metaphysics Research Lab Stanford.**
- Mehrabian, A., and Russell, J.A. 1974 *An Approach to Environmental Psychology* Cambridge, MA, USA, MIT Press.**
- Menéndez-Viso, A. 2009. "Black and white transparency: contradictions of a moral metaphor," *Ethics and Information Technology* (11:2), pp 155-162.**
- Meyer, B. 2007 "The effects of computer-elicited structural and group knowledge on complex problem solving performance," in: *Mathematics and Natural Science Faculty*, Humboldt University Berlin, Berlin.**
- Midgley, M. 1981. "Trying Out One's New Sword," in: *Morality and Moral Controversies*, J. Arthur (ed.), Upper Saddle River, NJ, Simon & Schuster, pp. 116-119.**
- Millgram, E. 1987. "Aristotle on Making Other Selves," *Canadian Journal of Philosophy* (17:2), pp 361 - 376.**
- Ming, A. 2012. "How computer and Internet use influences mental health: a five-wave latent growth model," *Asian Journal of Communication* (23:2), pp 175-190.**
- Moon, Y. 2000. "Intimate Exchanges: Using Computers to Elicit Self-Disclosure from Consumers," *Journal of Consumer Research* (26:4), pp 323-339.**
- Morahan-Martin, J. 2007. "Internet use and abuse and psychological problems," in: *Oxford handbook of internet psychology*, A.N. Joinson, K.Y.A. McKenna, T. Postmes and U.-D. Reips (eds.), Oxford, UK, Oxford University Press, pp. 331-345.**
- Moriarty, J. 2009. "Rawls, Self-Respect, and the Opportunity for Meaningful Work," *Social Theory and Practice* (35:3), pp 441-459.**
- Munn, N.J. 2012. "The reality of friendship within immersive virtual worlds," *Ethics and Information Technology* (14:1), pp 1-10.**
- Musk, E. 2014 "All Our Patent Are Belong To You," in: *Tesla Blog*, Tesla.**
- Nass, C., and Moon, Y. 2000. "Machines and Mindlessness: Social Responses to Computers," *Journal of Social Issues* (56:1), p 81.103.**
- Nguyen, C., Haynes, P., Maguire, S., and Friedberg, J. 2013. "A User-Centred Approach to the Data Dilemma: Context, Architecture, and Policy," in: *The Digital Enlightenment Yearbook 2013*, M. Hilebrandt (ed.), Brussels, IOS Press.**
- Nielsen, J. 1993 *Usability Engineering* Mountain View, CA, USA, Morgan Kaufman.**
- Nietzsche, F. 1883-1885 *Also sprach Zarathustra - Ein Buch für Alle und Keinen*, (2010 ed.) München, C.H.Beck.**
- Nietzsche, F. 1887 *Zur Genealogie der Moral* Leipzig, C. G. Naumann.**
- Nietzsche, F. 1974 *The Gay Science* New York, Walter Kaufmann.**

- Nissenbaum, H. 2004. "PRIVACY AS CONTEXTUAL INTEGRITY," *WASHINGTON LAW REVIEW* (79:1).
- NIST 2013 "NIST 800-53: Security and Privacy Controls for Federal Information Systems and Organizations," N.I.o.S.a. Technology (ed.), U.S. Department of Commerce, Gaithersburg, MD.
- Nonaka, I., and Takeuchi, H. 1995 *The knowledge creating company: How Japanese companies create the dynamics of innovation*. London, Oxford University Press.
- Nonaka, I., and Takeuchi, H. 2011. "The Wise Leader," *Harvard Business Review*: May 2011), pp 58-67.
- Norman, D. 2007 *The Design of Future Things* New York, USA, Basic Books.
- Norman, D.A. 1988 *The Psychology of Everyday Things* New York, USA, Basic Books.
- Novak, T.P., Hoffman, D.L., and Yung, Y.-F. 2000. "Measuring the Customer Experience in Online Environments: A Structural Modeling Approach " *Marketing Science* (19:1), pp 22-42.
- Novotny, A., and Spiekermann, S. 2014. "Oblivion on the Web: An Inquiry of User Needs and Technologies," European Conference on Information Systems (ECIS 2014), Tel Aviv.
- Nuria, O., Garg, A., and Horvitz, E. 2004. "Layered representations for learning and inferring office activity from multiple sensory channels," *Computer Vision and Image Understanding* (96:2), pp 163-180.
- Nussbaum, M. 2004 *Hiding from humanity: disgust, shame and the law* Princeton, New York, Princeton University Press.
- O'Donoghue, T., and Rabin, M. 2000. "The economics of immediate gratification," *Journal of Behavioral Decision Making* (13:2), pp 233 - 250.
- O'Donoghue, T., and Rabin, M. 2001. "Choice and procrastination," *Quartely Journal of Economics* (116:1), pp 121-160.
- Ohm, P. 2010. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization " *UCLA Law Review* (57), pp 1701-1777.
- Organisation for Economic Co-operation and Development (OECD) 1980. "OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data."
- Orwell, G. 1949 1984 New York, USA, The New American Library.
- Pariser, E. 2011 "Beware Online Filter Bubbles," T.-I.W. Spreading (ed.), TED.com.
- Pariser, E. 2012 *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think* London, The Penguin Press.
- Parviz, B.A. 2009 "Augmented Reality in a Contact Lens," in: *IEEE Spectrum*, IEEE.
- Pavlov, I.P. 1927 *Conditional reflexes: An investigation of the physiological activity of the cerebral cortex* London, Wexford University Press.
- Penton- Voak, I.S., Perrett, D.I., and Peirce, J.W. 1999. "Computer graphic studies of the role of facial similarity in judgements

- of attractiveness," *Current Psychology* (18:1), pp 104-117.
- Pettit, P. 1979 *Republicanism: A Theory of Freedom and Government* Oxford, Oxford University Press.
- Pettit, P. 2004. "Trust, reliance and the internet," *Analyse und Kritik* (26:1), pp 108-121.
- Pierce, J.L., Kostova, T., and Dirks, K.T. 2003. "The State of Psychological Ownership: Integrating and Extending a Century of Research," *Review of General Psychology* (7:1), March, pp 84-107.
- Pieters, W. 2011. "Explanation and trust: what to tell the user in security and AI?," *Ethics and Information Technology* (13:1), pp 53-64.
- Piètre-Cambacédès, L., and Chaudet, C. 2010. "The SEMA referential framework: Avoiding ambiguities in the terms "security" and "safety"," *International Journal of Critical Infrastructure Protection* (3:2), pp 55-66.
- Popper, K.R. 1974. "Selbstbefreiung durch Wissen," in: *Der Sinn der Geschichte*, L. Reinisch (ed.), München, C.H.Beck.
- Porteous, D.J. 1976. "Home: The territorial core," *Geographic Review* (66:4), pp 383-390.
- Preston, S.D., and de Waal, F.B.M. 2002. "Empathy: Its ultimate and proximate bases," *Behavioral and Brain Sciences* (25:1), pp 1-72.
- Ragnedda, M., and Muschert, G.W. 2013 *The Digital Divide - The internet and social inequality in international perspective* London and New York, Routledge.
- Rasmussen, J. 1983. "Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models," *IEEE Transactions on Systems, Man, and Cybernetics* (13:3), pp 257-266.
- Rawls, J. 1971 *A Theory of Justice* Oxford, Oxford University Press.
- Rawls, J. 2001 *The Law of Peoples: With, The Idea of Public Reason Revisited* Cambridge, USA, Harvard University Press.
- Raz, J. 1996 *The morality of freedom* Oxford, Clarendon Press.
- Redman, T.C. 2013. "Data's Credibility Problem," *Harvard Business Review*), December 2013, pp 2-6.
- Reeves, B., and Nass, C. 1996 *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places* New York, USA, Cambridge University Press.
- Resnick, P. 2000. "Reputation Systems:," *Communications of the ACM* (43:12), pp 45-48.
- Roberts, J., and Koliska, M. 2014. "The effects of ambient media: What unplugging reveals about being plugged in," *First Monday* (19:8).
- Robinson, J. 2010 "Blunt the e-mail interruption assault - If you're constantly checking messages, you're not working ", Microsoft NBC News.
- Rogers, E. 2003 *Diffusion of Innovations*, (4th ed.) New York, USA, The Free Press.
- Rokeach, M. 1973 *The Nature of Human Values* New York, Free Press.

- Rosen, J. 2005 *The Naked Crowd - Reclaiming Security and Freedom in an Anxious Age* New York, Random House.
- Rudmin, F.W., and Berry, J.W. 1987. "Semantics of ownership: A free recall study of property," *The Psychological Record* (37:2), pp 257-268.
- Salanova, M., Peiró, J.M., and Schaufeli, W.B. 2010. "Self-efficacy specificity and burnout among information technology workers: An extension of the job demand-control model," *European Journal of Work and Organizational Psychology* (11:1), pp 1-25.
- Salvucci, D.D., and Bogunovich, P. 2010. "Multitasking and monotasking: The effects of mental workload on deferred task interruptions," ACM Conference on Computer Human Interaction (CHI 2010), ACM Press, Atlanta, Georgia, USA.
- SANS Institute 2004. "An Overview of Sarbanes-Oxley for the Information Security Professional," Swansea, UK.
- Sartre, J.-P. 1992 *Being and Nothingness - A Phenomenological Essay on Ontology* New York, Washington Square Press.
- Scannapieco, M., Missier, P., and Batini, C. 2005. "Data Quality at a Glance," *Datenbank-Spektrum* (14), pp 6-14.
- Scerbo, M.W. 1996. "Theoretical Perspectives of Adaptive Automation," in: *Automation and Human Performance*, R. Parasuraman and M. Mouloua (eds.), New Jersey, USA, Lawrence Erlbaum Associates, pp. 37-63.
- Searls, D. 2012 *The Intention Economy - When Customers Take Charge* Boston, USA, Harvard Business Review Press.
- Seligman, M.E.P. 1975 *Helplessness: On Depression, development, and death* San Francisco, USA, Freeman.
- Seneviratne, O., and Kagal, L. 2014. "Enabling Privacy Through Transparency," IEEE Conference on Privacy, Security and Trust, IEEE, Toronto, Canada.
- Sheridan, T. 2002 *Humans and Automation: System Design and Research Issues* Santa Monica, USA, John Wiley & Sons.
- Sheridan, T.B. 1988. "Task allocation and supervisor control," in: *Handbook of Human-Computer Interaction*, M. Helander (ed.), Amsterdam, The Netherlands, North-Holland: Elsevier Science Publisher, pp. 159-173.
- Sheridan, T.B. 2000. "Function allocation: algorithm, alchemy or apostasy?," *International Journal of Human-Computer Studies* (52:2), pp 203-216.
- Shibata, T. 2004. "An overview of human interactive robots for psychological enrichment," *Proceedings of IEEE* (92:11), pp 1749-1758.
- Shilton, K. 2013. "Values Levers: Building Ethics into Design," *Science, Technology & Human Values* (38:3), pp 374 -397.
- Shilton, K., Koepfler, J., and Fleischmann, K. 2012. "Chartering Sociotechnical Dimensions of Values for Design Research," *The Information Society* (29:5), pp 1-37.
- Shneiderman, B. 2000. "Universal Usability," *Communications of the ACM* (43:5), May 2000, pp 85-91.
- Simpson, T.W. 2011. "e-Trust and reputation," *Ethics and*

Information Technology (13:1), pp 29-38.

- Snowden, E. 2014 "On Liberty: Edward Snowden and top writers on what freedom means to them," in: *The Guardian*, London.
- Solove, D.J. 2001. "Privacy and Power: Computer Databases and Metaphors for Information Privacy," *Stanford Law Review* (53), pp 1393-1462.
- Solove, D.J. 2002. "Conceptualizing Privacy," *California Law Review* (90:4), July 2002, pp 1087-1156.
- Solove, D.J. 2006. "A Taxonomy of Privacy," *University of Pennsylvania Law Review* (154:3), pp 477-560.
- Sommerville, I. 2011 *Software Engineering*, (Nine ed.) International, Pearson.
- Soraker, J.H. 2012. "How shall i compare thee? Comparing the prudential value of actual and virtual friendship," *Ethics and Information Technology* (14), pp 209-219.
- Spence, E.H. 2011. "Information, knowledge and wisdom: groundwork for the normative evaluation of digital information and its relation to the good life," *Journal of Ethics in Information Technology* (13), pp 261-275.
- Spiekermann, S. 2007. "Perceived Control: Scales for Privacy in Ubiquitous Computing," in: *Digital Privacy: Theory, Technologies and Practices*, A. Acquisti, S.D. Capitani, S. Gritzalis and C.Lambrinoudakis (eds.), New York, Taylor and Francis.
- Spiekermann, S. 2008 *User Control in Ubiquitous Computing: Design Alternatives and User Acceptance* Aachen, Shaker Verlag.
- Spiekermann, S. 2012. "The Challenges of Privacy by Design," *Communications of the ACM* (55:7).
- Spiekermann, S., and Pallas, F. 2005. "Technology Paternalism - Wider Implications of RFID and Sensor Networks," *Poiesis & Praxis - International Journal of Ethics of Science and Technology Assessment* (4:1), Fall 2005, pp 6-18.
- Stehr, N. 1994 *The Knowledge Societies* London, SAGE Publications.
- Storch, N.A. 1992. "Does the user interface make interruptions disruptive? A study of interface style and form of interruption," Conference on Human Factors in Computing Systems, ACM Press, Monterey, California, pp. 14-24.
- Suler, J. 2004. "The Online Disinhibition Effect," *Cyber Psychology & Behavior* (7:3), pp 321-326.
- Sullins, J.P. 2008. "Friends by Design - A Design Philosophy for Personal Robotics Technology," in: *Philosophy and Design - From Engineering to Architecture*, P.E. Vermaas, P. Kroes, A. Light and S.A. Moore (eds.), Milton Keynes, UK, Springer Science.
- Swan, M. 2012. "Health 2050: The Realization of Personalized Medicine through Crowdsourcing, the Quantified Self, and the Participatory Biocitizen " *Journal of Personalized Medicine* (2), pp 93-118.
- Taddeo, M., and Floridi, L. 2011. "The case of e-trust," *Ethics and*

Information Technology (13:1), pp 1-3.

- Tavani, H. 2012 "Search Engines and Ethics," in: *The Stanford Encyclopedia of Philosophy*, E.N. Zalta (ed.), The Metaphysics Research Lab Stanford.
- Taylor, S.E., and Thompson, S.C. 1982. "Stalking the elusive vividness effect," *Psychological Review* (89:2), pp 155-181.
- Thaler, R., and Sunstein, C.R. 2009 *Nudge: Improving Decisions About Health, Wealth, and Happiness* New York, Penguin Books.
- The Economist 2013 "Trolls on the hill - Congress takes aim at patent abusers," in: *The Economist*, The Economist, London.
- The White House 2009 "Transparency and Open Government - Memorandum for the Heads of Executive Departments and Agencies," White House, Washington D.C.
- The White House 2013. "Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy," US Government, Washington D.C.
- Thompson, C. 2007. "The See-Through CEO," *Wired*).
- Thompson, S.C., and Spacapan, S. 1991. "Perceptions of Control in Vulnerable Populations," *Journal of Social Issues* (47:4), pp 1-21.
- Turilli, M., and Floridi, L. 2009. "The ethics of information transparency," *Ethics and Information Technology* (11:2), pp 105-112.
- Turkle, C., Taggart, W., Kidd, C., and Dasté, O. 2006. "Relational artefacts with children and elders: The complexities of cybercompanionship," *Connection Science* (18:4), pp 347-361.
- Turkle, S. 2011. "Authenticity in the Age of Digital Companions," in: *Machine Ethics*, M. Anderson and S.L. Anderson (eds.), New York, Cambridge University Press, pp. 62-76.
- UN General Assembly 1948 "Universal Declaration of Human Rights," U. Nations (ed.), UN General Assembly.
- United Nations 2013. "Comprehensive Study on Cybercrime," United Nations Office on Drugs and Crime, Vienna.
- Vallor, S. 2010. "Social networking technology and the virtues," *Ethics and Information Technology* (12:2), pp 157-179.
- Vallor, S. 2012. "Flourishing on facebook: virtue friendship & new social media," *Ethics and Information Technology* (14:3), pp 185-199.
- van den Hoven, J., and Rooksby, E. 2008. "Distributive Justice and the Value of Information," in: *Information Technology and Moral Philosophy*, J.v.d. Hoven and J. Weckert (eds.), Cambridge, Cambridge University Press.
- Varian, H.R., and Shapiro, C. 1999 *Information Rules - A Strategic Guide to the Network Economy* Boston, Massachusetts, Harvard Business Books Press.
- Venkatesh, V., Morris, M.G., Davis, G.B., and Davis, F. 2003. "User Acceptance of Information Technology: Toward a Unified View," *MIS Quarterly* (27:3), September 2003, pp 425-478.

- Verizon 2014. "2013 Data Breach Investigations Report," Verizon Trademark Services LLC, USA.
- von Dijk, J.A. 2013. "A theory of the digital divide," in: *The Digital Divide - The internet and social inequality in international perspective*, M. Ragnedda and G.W. Muschert (eds.), London and New York, Routledge, pp. 29-51.
- Wang, R.Y. 1998. "A Product Perspective on Total Data Quality Management," *Communication of the ACM* (41:2), pp 48-65.
- Warren, S.D., and Brandeis, L.D. 1890. "The Right to Privacy," *Harvard Law Review* (4:5), Dec. 15, 1890, pp 193-220.
- Weil, S. 1952 *The need for roots: Prelude to a declaration of duties towards mankind* London, Routledge and Kegan Paul Ltd.
- Weiser, M. 1991. "The Computer for the 21st Century," *Scientific American* (265:3), September 1991, pp 94-104.
- Weitzner, D., Abelson, H., Berners-Lee, T., Feigenbaum, J., Hendler, J., and Sussman, G.J. 2008. "Information Accountability," *Communications of the ACM* (51:6), pp 82-87.
- Westin, A. 1967 *Privacy and Freedom* New York, USA, Atheneum.
- Whitworth, B., and Liu, T. 2008. "Politeness as a Social Computing Requirement," in: *Handbook for Conversation Design for Instructional Applications*, R. Luppigini (ed.), IGI Global, pp. 419-436.
- Wickens, C.D. 2002. "Multiple resources and performance prediction," *Theoretical Issues in Ergonomics Science* (3:2), pp 159-177.
- Wiener, N. 1954 *The Human Use of Human Beings: Cybernetics and Society*, (2nd Edition ed.) Boston, Da Capo Press.
- Wolf, M., Miller, K., and Grodzinsky, F.S. 2009. "On the meaning of free software," *Ethics and Information Technology* (11:4), pp 279-286.
- Woods, D.D. 1996. "Decomposing Automation: Apparent Simplicity, Real Complexity," in: *Automation and Human Performance - Theory and Application*, R. Parasuraman and M. Mouloua (eds.), New Jersey, USA, Lawrence Erlbaum Associates, pp. 3-17.
- World Economic Forum 2012. "Rethinking Personal Data: Strengthening Trust," World Economic Forum (in co-operation with The Boston Consulting Group), Davos.
- World Economic Forum 2014. "Rethinking Personal Data: Trust and Context in User-Centred Data Ecosystems," World Economic Forum (in co-operation with Microsoft), Davos.
- World Health Organization 1946 "Preamble to the Constitution of the World Health Organization ", W.H. Organization (ed.), World Health Organization, New York.
- World Health Organization 2001. "Strengthening mental health promotion," World Health Organization, Geneva.
- Yee, N. 2014 *The Proteus Paradox* New Haven, Yale University Press.
- Yee, N., Bailenson, J.N., and Ducheneaut, N. 2009. "The Proteus Effect: Implications of Transformed Digital Self-

- Representation on Online and Offline Behavior,"**
***Communication Research* (36:2), pp 285-312.**
- Yee, N., Ducheneaut, N., Nelson, L., and Likarish, P. 2011.**
"Introverted Elves and Conscious Gnomes," *Computer Human*
***Interaction (CHI)*, Vancouver, Canada, pp. 753-762.**
- Yousafzai, S.Y., Foxall, G.R., and Pallister, J.G. 2010. "Explaining**
Internet Banking Behavior: Theory of Reasoned Action,
Theory of Planned Behavior, or
Technology Acceptance Model?," *Journal of Applied Psychology*
(40:5), pp 1172-1202.
- Zhang, P. 2005. "The Importance of Affective Quality,"**
***Communications of the ACM* (48:9), September 2005, pp 105-**
108.
- Zimbardo, P.G., and Gerrig, R.J. 1996 *Psychologie*, (7 ed.),**
Springer Verlag Berlin Heidelberg New York.
- Zimmerman, M.J. 2010 "Intrinsic vs. Extrinsic Value," in: *The***
***Stanford Encyclopedia of Philosophy*, E.N. Zalta (ed.), The**
Metaphysics Research Lab Stanford.