Ethics and Information Technology (2022) 24: 35
https://doi.org/10.1007/s10676-022-09661-y

ORIGINAL PAPER

Technology and moral change: the transformation of truth and trust

John Danaher[1] · Henrik Skaug Sætra[2]

Accepted: 19 July 2022 / Published online: 20 August 2022

Abstract
Technologies can have profound effects on social moral systems. Is there any way to systematically investigate and anticipate these potential effects? This paper aims to contribute to this emerging field on inquiry through a case study method. It focuses on two core human values—truth and trust—describes their structural properties and conceptualisations, and then considers various mechanisms through which technology is changing and can change our perspective on those values. In brief, the paper argues that technology is transforming these values by changing the costs/benefits of accessing them; allowing us to substitute those values for other, closely-related ones; increasing their perceived scarcity/abundance; and disrupting traditional value-gatekeepers. This has implications for how we study other, technologically-mediated, value changes.

Keywords Technology · Moral change · ICT · AI · Robotics · Truth · Trust

Introduction

Technologies can have profound effects on social moral systems. Consider the stirrup. According to Lynn White Jr's classic study *Medieval Technology and Social Change*, the invention of the stirrup was the primary technological facilitator of the development of the feudal system (White Jr, 1962). The feudal system created a new social moral system in which mounted knights were seen to be the most valuable and respected contributors to aristocratic armies. They were to be supported and sustained by large estates that gave them the food and material resources they required. New norms of social hierarchy, chivalry and property rights emerged as a result.

How could the humble stirrup be responsible for all this? Prior to the invention of the stirrup, mounted warriors were not particularly effective fighters. They relied on their own strength to maintain their stability on top of a horse. They could occasionally throw a spear or slash a sword, but they could do little more. They were vulnerable to attack, rarely able to carry both a weapon and a shield. The stirrup changed all that by stabilising the warrior and enabling him to fuse himself and the horse into a single fighting unit. Suddenly mounted knights could provide the decisive difference between defeat and victory. White's thesis is, of course controversial, and some claim he overstates the role of the stirrup in founding the feudal system. But it seems safe to say that the stirrup did transform the power of mounted cavalry and this had knock-on implications for military decision-making and social power relations.[1]

If the invention of the stirrup could have such a profound and transformative effect on the medieval social value system, what might contemporary digital technologies (ICT, AI and robotics in particular) be doing to our current value systems? Is there any way to systematically investigate and anticipate these potential transformations? These are important questions and an emerging field of scholarship is dedicated to answering them (e.g. Danaher, 2021; Hoepster, 2022; Klenk et al., 2022; Nickel et al., 2021; van de Poel, 2021; van de Poel & Kudina, 2022; Swierstra et al., 2009; Verbeek, 2012, 2013). We aim to contribute to this emerging field through case studies: we identify two core human values, describe their structural properties and conceptualisations and then consider various mechanisms through which

✉ John Danaher
johndanaher1984@gmail.com; john.danaher@nuigalway.ie

Henrik Skaug Sætra
henrik.satra@hiof.no

[1] School of Law, NUI Galway, University Road, Galway, Ireland

[2] Department of Computer Science and Communication, Ostfold University College, Halden, Norway

[1] For a similar story of how a single technology resulted in significant social change, see Pelto (1973) on the impact of the snowmobile on the Arctic communities.

technology is changing and might in the future change our perspective on those values.

The paper is divided into three main parts. First, we consider the nature of social moral change and why it is important to study potential future changes to social morality. Second, we outline some basic mechanisms through which technology can effect moral change. Third, we discuss two core human values—truth and trust—and consider the various ways in which digital technologies, particularly AI and robotics, are and might be transforming how we think about them.

The bulk of this article is taken up with the case studies. You might wonder how we arrived at them. One answer is simply that we have studied them closely in our previous work (refs omitted). As a result, their selection is the product of our scholarly histories. But that is not the only reason. There are important connections between these values. For instance, they are both, in part, epistemic values, relating to how we acquire or forgo the need for knowledge. They both have instrumental and intrinsic value. And they are related to one another: truth is related to trust insofar as trust is often needed when we lack direct access to truth; and trust is related to truth insofar as some people argue that we can forgo trust when we have direct access to the truth. Some technologies promise to replace the need for trust with direct access to truth; contrariwise some technologies undermine access to the truth (and our capacity to form true beliefs) and thus increase the need for trust. Studying the interrelationship between these two values is, we will suggest, particularly illuminating when it comes to understanding how technology impacts social morality.

Why study future moral change?

Why should we care about the relationship between technology and moral change? To answer that question we need to consider exactly what it is that we are inquiring into.

First, what is social moral change? Put simply, it is change in people's moral beliefs and practices. According to most philosophical accounts, the study of morality has two main branches to it: the inquiry into what is good/bad and the inquiry into what is right/wrong (Ross, 1930). The first inquiry covers all the things we think are morally valuable—i.e. worth pursuing, promoting and cherishing—as well as the things we think are morally disvaluable—i.e. worth ignoring, undermining and minimising. Valuable things (might) include freedom, love, happiness, achievement, beauty, truth and so on. Disvaluable things might include slavery, hate, sadness, failure, ugliness, falsehood and so on. The second inquiry covers all the actions we think we are morally required to perform—i.e. our obligations and duties—as well as the actions we are permitted to perform and forbidden to perform. Changes in social morality are thus characterised by changes in what people believe to be good/bad or right/wrong. For example, where once upon a time most people thought that slavery was permissible (and also perhaps even good, all things considered), most people now accept that it is impermissible (and bad, all things considered). This is a classic example of a social moral transformation (Appiah, 2010).

The careful reader will note that the focus here is on changes in social morality and not changes to ideal morality. As Calhoun notes, the majority of moral philosophers think that their job is to inquire into the nature of ideal morality (Calhoun, 2015). Ideal morality consists of the things that are actually good or bad and right or wrong, irrespective of what people believe or do. Some, but not all, theories of ideal morality hold that what is moral (i.e. actually good and actually right) is invariant across time and space. It is not the kind of thing that can change. But social morality is very different. It consists in people's beliefs about what is good or bad and right or wrong. This can and does change, irrespective of what ideal morality might be. The focus in this paper is on changes in social morality.

One concern you might have about this inquiry is whether we can meaningfully distinguish between changes to *moral* beliefs and practices and changes to other kinds of belief and practice. There are two aspects to this concern. On the one hand, you might be concerned about drawing the line between an occasional or ephemeral change in behaviour or belief and a genuine change in social morality. People sometimes behave or believe differently in different contexts. For example, many of us, when under pressure, lie to our friends. When challenged about this, we might even provide some moral justification of our behaviour (according to some moral psychologists this is common—see Haidt 'Emotional tale and cognitive dog'). Later, on reflection, we might agree that we did something wrong. Surely the original temporary change in behaviour (and associated moral belief) cannot constitute a change in social morality? But, if not, where do we draw the line? When does a change become sufficiently sticky and sustained to constitute a change in social morality? On the other hand, you might be concerned about the number of people or institutions that have to get onboard with a change in behaviour and belief for it to become a moral change. After all, compliance with moral norms is never complete. Some people murder, rape and steal. Some even believe they are doing the right thing. Nevertheless, most of us, most of the time, recognise these as moral transgressions.

We have no easy answers to these concerns. Social scientists have long-noted that it is hard to draw the line between the moral and the non-moral. For instance, there is a rich literature on the distinction between moral and social norms, with several competing accounts and little agreement on what, exactly, distinguishes the two (refs. Bicchieri, Elster,

article about both). Do moral norms have a specific type of content, or a elicit a specific type of emotional/cognitive reaction? Is it a bit of both? It's hard to say. When it comes to changes in social morality, the easiest test would be to say that a belief or practice counts as moral when (enough) people use moral language to represent and describe it and moralised emotions (pride, guilt, shame etc.) to respond to it and evaluate it. If we say that education is 'good', giving money to charity is 'right', murder is 'wrong' (and so forth), and if we respond to people by praising, blaming or shaming them for engaging in those acts or pursuing those ends, then it seems clear that we think those states of affairs and actions belong to the moral realm. If we make systematic changes to how we apply moral language and moral emotions, then it is safe to say that this constitutes a change to social morality. Although there are some moral theories that allow for a purely subjective determination of good/bad and right/wrong (e.g. ethical egoism), and hence a purely subjective application of moral language and emotions, most moral theories assume that these things are matters of either objective fact or widespread social consensus, even though there may be dispute as to what counts as objective fact and how widespread the social consensus has to be.

It is likely, then, that the borderline between a moral and a non-moral change will always be a fuzzy one. Nevertheless, in what follows, whenever we claim that a change in social belief and practice constitutes a change in social morality, we assume that this requires (a) a systemic pattern of behavioural and cognitive changes (i.e. not merely accidental or temporary); (b) that is observed across a sufficiently wide population (and not just one or two individuals); and (c) that is commonly described using moralised language (the language or good/bad or right/wrong) and evaluated in terms of moral emotions. Some of the changes we describe below will be purely speculative (things that might occur in the future) and based on emerging trends; some will be less speculative and based on observable patterns in contemporary life. To support our claims that these constitute changes to social morality, we will appeal to philosophical and ethical commentary on these changes that suggest they satisfy properties (a)–(c) and, wherever possible, empirical, particularly psychological, studies of these changes that suggest they satisfy properties (a)–(c).

There are practical and moral reasons to want to study changes to social morality, so defined. Practically speaking, we care about the future. We make decisions now that will affect ourselves and others in the future. If we build a road through some idyllic countryside, we know that this can have consequences for social organisation for decades, perhaps even millennia into the future. It is important to anticipate these changes in order to work out the long-term costs and benefits of the project. Changes to our moral beliefs and practices are just another kind of change that can impact the long-term costs and benefits of our projects. Anticipating and planning for those changes is important if we are to get a reasonable picture of whether the project is worth it. This is why an increasing number of researchers think that responsible innovation and design must include some consideration of possible future value changes (Kudina & Verbeek, 2019; van de Poel 2019; Verbeek, 2012).

Morally speaking, the idea that future generations might have different moral beliefs and practices to our own can be both uplifting and concerning. If there are things we currently value that might be threatened by future changes to social morality, we may have reason to campaign against those changes. For example, many people value privacy but are concerned that ICT and AI is slowly corroding people's attachment to privacy (Debrabander, 2020; Hartzog, 2018). This gives them a moral reason to limit this corrosion and they acquire this reason, in part, because they are willing to anticipate and plan for possible changes to social morality. Similarly, a fundamental tenet of conservatism as a political ideology is the notion that human fallibility provides good reason to be wary of radical change, that there is a certain wisdom in the existing order of things that we cannot necessarily fully grasp but that we should nevertheless respect (Burke, 1790). Society, as Edmund Burke argued, is in fact a partnership between us, those that are to come, and those that came before us, and he believed this is so because without such a partnership we would not obtain the required level of insight to develop society itself, but also science, art, and the virtues (Burke, 1790). Contrariwise, the fact that social moral beliefs change over time might make us more cautious and less convinced about our current moral attachments. Some of our ancestors may have believed that slavery was a good thing. We now think their judgment was in error. Do any of our current moral beliefs fall into a similar category? Could technological changes disrupt and harm our capacity to function as moral agents in the present, due to the moral uncertainty they create? Some authors have argued that we should take this possibility seriously (Danaher, 2021; Nickel, 2020; Williams, 2015). We can do this by anticipating and possibly getting ahead of future changes to social morality (Anthis & Paez, 2021).

In short, changes to social morality are changes to people's beliefs and practices about what is good or bad or right or wrong. It is important to study future changes in social morality for both practical and moral reasons.

Technology and mechanisms of moral change

How do changes to people's moral beliefs and practices come about? There are many causal factors potentially at play. The proximate mechanisms underlying social moral

change are likely to be psychological and neurological in nature (Churchland, 2019). But these proximate mechanisms are likely to be influenced by a whole range of more distal mechanisms of moral change. At an abstract level, we could divide those mechanisms into two main classes: material and ideological. Material mechanisms of change concern the interaction people have with their physical environments: how they obtain the resources they need to survive and so on. Ideological mechanisms of change concern the cultural forces changing how people think about their interactions with the world. This can include ideas shared via cultural institutions such as educational institutions and legal institutions. This contrast between material and ideological mechanisms of change is a prominent one in history and the social sciences, but it encompasses a wide variety of theoretical mechanisms of social change (see, for example, the debates about the history of economic growth discussed in Koyama and Rubin (2022)).

This article is not going to address all possible mechanisms of moral change. It is, instead, going to focus on technology as a mediator of moral change. Our goal is to show how technologies might affect future moral beliefs and practices with respect to the values of truth and trust. It is useful to have an organising framework in place at the outset to guide our interpretation and analysis of these case studies. Peter Paul Verbeek's theory of technological moral mediation is one useful starting point (Verbeek, 2012). In a series of books and papers, Verbeek has argued that technology mediates our moral relationship to the world. These ideas have been developed and expanded by others (e.g. Kudina, 2019; Swiestra et al., 2012). As Verbeek puts it himself:

> technologies-in-use help to establish relations between human beings and their environment. In these relations, technologies are not
> merely silent 'intermediaries' but active 'mediators' …By organizing relations between humans and world, technologies play an active, though not a final, role in morality.
> (Verbeek, 2013, pp. 77–78)

Verbeek's work suggests that there are at least two distinct forms of technological moral mediation: (i) pragmatic mediation and (ii) hermeneutic mediation.

Pragmatic moral mediation arises whenever technology affects the decision problems we face. An ordinary human life is full of decisions. Many of these decisions are morally charged. We choose among outcomes with different moral values and between actions that might be morally obligatory, forbidden or permissible. Technology affects our choices in at least two ways. First, technology can add or subtract options from our decision problems. The invention of the cell phone, for example, has given us new opportunities for connection. Whereas once upon a time we would have to wait until we reached the nearest payphone to call a friend, we can now reach out to anyone at any time. This creates new moral choices: would I be bothering the other person if I called them at night? Should I call my partner to let them know that I am okay? Is it okay to ignore a phonecall or text message? Second, technology can affect the costs and benefits of morally charged actions. For example, the invention of cheap and highly effective forms of contraception has, according to some scholars, affected our social moral norms around extramarital sex: by reducing the risks of unwanted pregnancy and sexually transmitted infections, contraception has made people far more willing to engage in extramarital sex and this has in turn reduced the social taboo associated with this practice (Adshade, 2013; Greenwood, 2020).

Hermeneutic moral mediation is different. It arises whenever technology changes how we interpret and understand some aspect of the world and/or our relationship to it. This happens at the level of moral perceptions, concepts and metaphors. Technologies can enable us to see things in a new light and this can alter our moral beliefs and practices. For instance, Verbeek argues that the invention of obstetric ultrasound can change our moral perception of the foetus in utero. By presenting the foetus to us as an entity separate from its mother, ultrasound encourages us to see the foetus as an independent moral being, capable of bearing moral rights and as an object for therapeutic interventions during prenatal care. Some people already had that perception of the foetus in utero, of course, but obstetric ultrasound made it more vivid and salient for more people. Similarly, the invention of smartphones and social media may be changing our perception of the value of our everyday experiences, particularly our social experiences with friends and colleagues. Instead of viewing these experiences as being primarily valuable in and of themselves, people may now see them as being primarily instrumentally valuable: as content to be recorded, shared and possibly monetised. This seems to be true, in particular, of people who make their living as social media influencers and lifestyle bloggers (see, for example, the ethnographic and qualitative studies of such individuals by Arrigada and Bishop (2021), Duffy and Kang (2020), Hund and McGuigan (2019), and Abidin (2016)). The technology has enabled this reinterpretation of the moral value of everyday experiences.

These two mechanisms of moral change are concerned with what we might call the first-order effects of technology on social morality: how technology affects particular moral decision problems (the options available to us; their costs and benefits) and particular moral perceptions (how we interpret events, actions and states of affairs through moralised concepts and ideas). In practice, technological change can lead to second and third (and so on) order effects on social morality. Consider, once more, the example of the stirrup. Following White's argument, the first order effect of

the stirrup was a straightforward instance of pragmatic moral mediation: it gave military leaders the option of using highly effective mounted knights to deliver decisive victories in combat. So effective were they that the moral case for their use became overwhelming: a military leader who failed to use them would not be doing their duty to king and country. But this led to second and third order social moral changes. An entire social-legal institution was established to support the creation and maintenance of mounted knights—the feudal system. This social-legal system came with its own set of moral beliefs and practices concerning social hierarchy, property rights and honour. Being on the lookout for these second and third order effects might be particularly valuable when it comes to anticipating future moral changes.

Stephen Barley's research on technology in the workplace provides a useful framework for understanding some of these higher-order effects (Barley, 2020). Using the *dramaturgical* theory of social relations (according to which social interactions can be understood to follow scripts, and take place on 'stages' with actors playing certain roles), Barley argues that the most socially transformative technologies are ones that disrupt social scripts and the relationships between different social roles. The stirrup and the feudal system is a good example of this. It greatly elevated the social power and status of mounted knights and thus their legal rights and entitlements relative to other social roles. Other technologies can have a more equalising effect between different social roles. For instance, in an ethnographic study, Barley argues that the internet has had an equalising effect on the relationship between car salespeople and customers. In particular, it made it easier for customers to find information about their preferred make and model, compare prices and extract themselves from unpleasant price negotiations. This resulted in significant behavioural and normative changes in car selling. In particular it made the salespeople more honest and transparent in their interactions with customers and less likely to engage in 'sharp' bargaining practices, such as creating a sense of urgency about the need to close a deal. It also resulted in a shift in how car dealers understood the value of what they were doing: the sales 'game' shifted from being about making large margins of profit on each individual sale, to being about the speed and volume of sales (Barley, 2020, ch 2).

In the remainder of this article, we will consider how these different mechanisms of moral change might play out when it comes to the relationships between technology and the human values of truth and trust. Our analysis will follow a common pattern. We will start by detailing the values, describing their different dimensions and the moral beliefs and practices that tend to be associated with them. Then we will consider various ways in which emerging technologies might be affecting those values. Finally, we will consider the future trajectory of those values: will they survive? Will they be radically transformed? Or will we stop valuing them altogether? The goal here is not to predict the future but, rather, to imagine plausible potential future scenarios. We do this through concrete examples, as opposed to abstract theories.

In presenting these case studies, we assume a form value pluralism and value scarcity. In other words, we assume there are many valuable things and that truth and trust are just two among those (others, include, health, pleasure, education, friendship and so on). We allude to this value pluralism at several points. We also assume that, because time, attention and resources are scarce, people often have to trade-off between different values. To use economic language, we assume that there is an 'opportunity cost' inherent in many value-related decisions that people make: in choosing to pursue or promote some values they often have to ignore or deprioritise others. These two assumptions—value pluralism and scarcity—affect our case studies because we believe that changes in how people prioritise or compromise between different values, provided they are sufficiently systematic and widespread, constitute a kind of change in social morality. It is also worth noting that since our case studies concern changes in existing values, we will not discuss how technology might facilitate the identification of new values. For instance, some people argue that the value of 'sustainability' is a relatively new value that has emerged as a result of increased awareness of the environmental impact of technologies (Poel & Taebi, 2022). This may well be true but it is a limitation of our case study method that we are not going to identify such possibilities.

The transformation of truth

Understanding the value of truth

Let's consider, first, the value of truth and how it is affected by technology. Truth is generally understood to be both intrinsically and instrumentally valuable (Horwich, 2006). There is, for instance, a widely-endorsed view among philosophers that having true beliefs about the world is intrinsically valuable, irrespective of the content of those beliefs (Whiting, 2013). So, for instance, knowing that the Earth will eventually be destroyed by the Sun might be depressing, but it is true and it is good to know that it is true. There is also a widely endorsed view that having true beliefs is instrumentally valuable. True beliefs enable you to accomplish your goals and plan your actions. Having the true belief that it is raining outside is more practically valuable than having the false belief that it is not. At least with the true belief you are more likely to bring an umbrella with you when you leave your house.

These claims need to be finessed. Truth is valuable but, consistent with value pluralism, it is not the only thing that

is valuable. Physical health, social intimacy, mental stability (among many other things) are also valuable. Sometimes there can be a tension between these values and truth. These tensions have surfaced in the debate about the intrinsic value of truth. Critics of that position will, for instance, argue that there is little value in accumulating trivial truths. There may be a fact of the matter when it comes to the total number of blades of grass on the lawn outside my window, and I could spend a long time counting them all, but people would surely question the value of acquiring such a trivial true belief. Some people respond to this by arguing that acquiring true beliefs about significant or important matters is intrinsically valuable, but acquiring them about trivial matters is not (Whiting, 2013, 225ff). The problem with this response is that it is not clear where to draw the line between significant and insignificant truths. Some, seemingly trivial and useless truths can turn out to be useful (Flexner, 1939). Alternatively, some people address the trivial truth problem by simply reemphasising that truth is just one among many values and we have to weigh the benefits of acquiring true beliefs against the potential costs to other values. If it would take too much time and energy to count those blades of grass, and if the cost to physical and mental wellbeing would be high, then perhaps it is best not to do so. This can create problems when it comes to the intrinsic and instrumental benefits of truth. Sometimes true beliefs hinder or scupper our plans. There is, for instance, a lot of psychological research on the value of positive illusions (Bortolotti, 2018; Jefferson et al., 2017). The practical importance of this research is disputed but there does seem to be some evidence suggesting that people that are unrealistically optimistic about their health or personal circumstances score more highly on certain measures of well-being and, even more starkly, on certain health outcomes (Murray & Holmes, 1997; Schiavon et al., 2017; Taylor & Brown, 1994). The tension between truth and other values is something that technological change can exacerbate. This increased tension can lead to more people trading the value of truth off against some other, to them more important, value such as personal happiness or autonomy.

To understand how this happens, it is worth considering the psychology of truth. If, given the option, will humans seek out true beliefs instead of false ones? The available evidence is mixed. There is some tendency to seek out true beliefs and, as noted, such beliefs can often be practically necessary, but our commitment to the truth is fickle and not absolute. Decades of research in cognitive science and psychology suggest that there is a significant bias in most people's brains that means they do not focus on persuasion and confirmation more than the acquisition of true beliefs (Stanovich, 2021). In other words, there is a tendency within most people to engage in motivated reasoning and to confirm their existing beliefs and values. They tend to overlook, or explain away, anything that calls those beliefs and values into question. They are keen to persuade others of their beliefs and values and to form tribal loyalties and cohesive identities (Mercier & Sperber, 2017). They are less quick to identify and embrace unwelcome truths. Obviously, these are general patterns, not universal truths, and there is some criticism of and calls for a broadening of Mercier and Sperber's theory of reasoning (for example Prochownik (2019), Mascarenhas (2019), and Dogramaci (2020)), some of which has been addressed by the authors themselves in *Replies to critics* (Mercier & Sperber, 2019). There is individual variation—some people are not so keen to conform—and there is social variation—some societies incentivise and reward the gadflies (we discuss some emerging sub-communities with these properties below). Furthermore, in some cases, external reality serves as an ultimate sanity check on the irrational or illusory beliefs of individuals or groups: if there is some readily confirmable fact of the matter, it is possible for someone to point this out to a group that is otherwise sustaining a false belief. In other words, it is possible to speak truth to power.

Our main claim here is that technological change can modulate our commitment to seeking the truth and thus change how we prioritise and compromise in relation to its value. To understand how this happens, it is worth considering the age-old question: What is truth? To be clear, answering this metaphysical question is not essential to the aims of this article, nor would it be possible to give a compelling answer in a short space. Nevertheless, it is worth surveying some possible answers to get a better sense of how technology can affect the value of truth (for more details, see Glanzberg, 2021). The classic model of truth is the *correspondence model* (David, 2020). On this model, our beliefs are true if they correspond to some objective reality. If I believe that the emperor is not wearing any clothes and it turns out that he is not wearing any clothes, then my belief is true. If he is wearing clothes, it is false. Though it is not without its philosophical problems, this correspondence model often serves well for simple, everyday factual disputes. It becomes more problematic when we are dealing with abstract, theoretical beliefs and/or moral or aesthetic beliefs, which don't obviously map onto some external objective reality. This is why some people prefer an 'epistemic norms' model of truth.[2] This view claims that true beliefs or propositions are those that have passed some validation test that has been agreed upon by a community of epistemic peers (e.g. it is been experimentally replicated, not falsified; it is supported by logically valid arguments and so on).

[2] Roughly, what we have in mind with this term is a 'pragmatic' approach to truth, whereby what counts as a 'true' proposition depends on the epistemic norms within a relevant community/discourse. For more, see Capps 2019.

These views have been debated and refined by philosophers over millennia. Most people, of course, do not think about truth in the rarefied and technical terms of philosophers. There is a limited amount of empirical work on how ordinary people understand the metaphysics of truth. Many philosophers that defend the correspondence theory of truth do so in the belief that it captures the 'folk' conception of truth (Barnard & Ulatowski, 2013), but pioneering empirical work by Arne (1938; Asay, forthcoming) and more recent work by Barnard and Ulatowski (2013, 2019), as well as Reuter and Brun (2022) puts this in doubt. Although non-philosophers do often say things consistent with the correspondence theory, their commitment to it seems to vary depending on the context and nature of the truth claim. For instance, their implicit theory of truth may be different depending on whether they are presented with a claim about mathematical truth or social truth (Barnard & Ulatowski, 2013, 2019). This suggests, in turn, that people may lean into an 'epistemic norms' model of truth, assuming that what counts as a truth varies depending on the epistemic norms of different disciplines and communities.

Fortunately, we do not need to pick and choose between these metaphysical models here. Our claims about the impact of technology on the value of truth work with both models. What is crucial, however, is that in order to say that a community or individual values truth, there must be (a) some commitment, among the members of that group, to acquiring true beliefs and (b) some agreed epistemic process for confirming or validating beliefs. Checking that beliefs correspond to an external reality may often be the most obvious way to validate them, but not be the only way to do so. Our central claim is that technology can affect the value of truth by affecting both our commitment to truth and the processes we follow for confirming true beliefs (an idea also supported by the empirical work of Reuter and Brun (2022)).

This brings us to one last preliminary point. Since commitment to truth depends on a commitment to certain epistemic processes for generating and validating propositions, it follows that commitment to the value of truth often entails commitment to values that support these epistemic processes. For instance, commitment to the value of free speech is often justified because of its link to the truth. John Stuart Mill's famous defence of free speech, in chapter 2 of On Liberty (1859), is the classic statement of this position. But free speech is not the only value connected to truth. In his book *The Constitution of Knowledge*, Jonathan Rauch outlines a set of values that people committed to the process of acquiring true beliefs ought to and tend to affirm (Rauch, 2021, ch. 4). They include the value of *science and experimentation*, *objectivity*, *fallibilism* (that truth claims are defeasible and capable of being proved wrong), *accountability* (if you get something wrong or violate the epistemic norms of the community, you are held to account for this), *pluralism* and *free inquiry* (you welcome curiosity, new propositions and truth claims), *civility* (you prefer to resolve factual disputes through shared epistemic processes and not violence), *institutionalism* (you value fact-checkers and gatekeepers for their role in validating propositions) and so on. We could quibble about the inclusion of some items on this list but it is a plausible survey of some of the values that are connected to the value of truth. The challenge, however, is that although these values are connected to truth they are also dissociable from it. Some of them can be valued for other reasons. For instance, we can value civility and accountability for reasons other than their role in sustaining our commitment to the truth. This too is an important point when it comes to assessing the impact of technology on the value of truth.

How technology affects the value of truth

So how does technology affect the value of truth? Limiting our focus to digital information technologies, three things appear to be happening at the moment that affect our commitment to truth. First, digital information technology is giving people (and governments and corporations) the power to create more *information* and, perhaps more crucially, more *disinformation*. This limits our capacity to agree upon what is true and confirm or validate truth claims. There are a number of different technological mechanisms underlying this trend. One is simply the volume of information that is being generated and shared via digital networks. A lot of this information may be factual but the sheer volume is overwhelming our traditional processes for validating and checking whether the information is true. This has happened before—the creation of the printing press led to a similar information deluge—but not at the same scale or speed. In addition to this, a number of technologies now allow people to create hyperrealistic fake information. Deepfake technologies (audiovisual images created through generative adversarial networks) are the most widely-discussed examples of this. This fake information can fool our traditional validation processes. This can lead to an increased number of false beliefs or, at the very least, an increased amount of scepticism about our capacity to access the truth (Fallis, 2021; Rini, 2020).

Second, digital infrastructures seem to amplify and promote information for reasons other than its connection with the truth. This is where technology intersects with psychology and social institutions. As noted previously, there is a tendency in human psychology to engage in motivated reasoning, to seek out information that confirms our existing beliefs, to generate and sustain tribal loyalties. Information technology allows us to do more of that. We can live inside filter bubbles that reinforce existing beliefs and identities. It's easier for us to avoid unpleasant truths than ever before—the social gadfly pointing out that the emperor

has no clothes cannot pierce the filter bubble or make us care about his 'alternative' facts. Social institutions are also facilitating this move away from truth. The online economy is largely an attentional economy where capturing attention is the main goal, not generating and sustaining true beliefs (Nelson-Field, 2020; Williams, 2019; Wu, 2016). This has had significant impacts on some traditional fact-checking institutions such as journalism. Although there are many good fact-checking journalists and online media organisations, they struggle in the face of increasingly tribal media organisations that do a better job of capturing our attention. The current political economy also does not help. Many democratic institutions reward politicians and parties that appeal to tribal loyalties and 'populism', not truth. A number of countries around the world, with longstanding political institutions that are intended to provide checks and balances against these forms of populism, are now struggling to maintain those checks and balances. Recent experiences in the US are the most prominent, but not the only examples of this (Gurri, 2018; Levitsky & Ziblatt, 2018; Runciman, 2018).

Third, there are also impressive developments in virtual reality and augmented reality (extended reality or XR) technologies. These technologies allow people to create and enter alternative computer-generated worlds or to layer computer-generated information on top of physical reality. Although no one would currently be fooled into thinking that these computer-generated realities are equivalent to physical reality, they are becoming increasingly immersive and realistic, and they do give people the option of 'escaping' into an alternative reality if they find some aspect of physical reality unwelcome. In other words, technology is giving people the power to create or join a virtual or extended reality that matches their own beliefs and preferences (Chalmers, 2022). In a sense, then, if these technologies develop to a sufficient degree, people will get to choose their own truths by choosing their own reality and creating their own 'facts'.

What impact is this having on the value of truth? What impact is it likely to have going forward? We think there are two significant impacts that are worth highlighting and monitoring. First, these technologies are weakening the instrumental value of accessing or being committed to truth. It is increasingly difficult to sort fact from fiction in planning our actions and achieving our goals; it is also increasingly unnecessary. Technology gives us the means to bend some aspects of reality to our will or to enter into an alternative reality that better matches our preferences. The sanity check of external reality is, consequently, losing its motivational salience. It doesn't matter whether the emperor is wearing new clothes, or not. If he isn't, you can create an alternative version of reality in which he is. This enables people to deprioritise the value of seeking truth in their daily lives—to trade it off against other values. Since many people's commitment to the truth is already unstable and fickle, it's likely that they will avail of this opportunity. They will seek out things that make them feel good or happy, that allow them to express their autonomy and creativity, or solidify their tribal identities, instead of seeking out the truth. This is likely, in turn, to corrode many of the values associated with truth. Free inquiry and free speech, accountability for telling the truth, civility in resolving epistemic disputes—each of these values is likely to be less compelling if the instrumental value of seeking the truth is weakened.

We already see some evidence of this happening. Free speech and free inquiry have always been contested but they are now highly contested largely on the grounds that they pose a threat to other values including sense of self and identity cohesion. For instance, many attempts to regulate or limit speech on university campuses (a lot of which is perceived to be in tension with the value of free speech) is defended in terms of the need to protect particular identity groups from speech that threatens their sense of identity (several of the essays in Riley (2020) highlight this phenomenon with specific reference to examples from the recent campus free speech wars). Online discourse also appears to reward those that signal in-group loyalty by expressing outrage against others (Crockett, 2017; Brady et al., 2021). This reinforces tribal identities at the expense of mutual understanding. The value of civility appears to be under threat as a result. Seeking common ground is less important than maintaining standing within a group. Accountability for telling the truth also appears to be dissipating. Donald Trump's ability to flood the information channels with lies and falsehoods, while suffering very few consequences, seems to exemplify this, though there are many other examples. All that said, as noted above, these related values are not necessarily connected with truth. So one potential consequence of these technological changes that we are witnessing is that people will continue to prioritise and promote these values but do so for other reasons. This might lead to subtle shifts in how we understand and enforce truth-related values. For instance, free speech might be valued for its role in creating and forming identities, and not for its connection with truth. Accountability might be valued for its role in naming and shaming people that do not conform to certain group norms, and not for its role in keeping people honest and focused on the truth.

What about technology's impact on the intrinsic value of truth? Ironically, to the extent that technology makes truth more elusive and difficult to validate, it may increase its perceived intrinsic value and salience. As something becomes more scarce, people often end up attaching more value to it. For some people, the scarcity or difficulty associated with accessing a value might make it more appealing and beneficial than alternative more common values. And, indeed, there is some evidence to suggest that such people exist and are forming their own group identities around their

commitment to the value of truth. Certain academic institutions would profess this value system. But there are also online communities of rationalists that dedicate a lot of time and attention to pursuing the truth and, crucially, avoiding the mistakes of motivated reasoning and other psychological biases that draw us away from true beliefs. Julia Galef's book *The Scout Mindset* can be seen as a manifesto for this value system (Galef, 2021). Not only does Galef celebrate the importance of acquiring true beliefs and avoiding motivated reasoning, she also argues that the alleged benefits of trading the commitment to truth off against other values are less appealing than they first seem. For instance, she critiques empirical studies that suggest that positive illusions can be beneficial, arguing that these benefits are overstated and that the instrumental value of truth is still quite high. Although she does not frame her book in these terms, it is possible to see it as a reaction to the impact of technology on the value of truth. The increased scarcity of truth makes it seem more intrinsically valuable than it might otherwise have been. This increased perceived intrinsic value of truth may have knock-on consequences for other values. Epistemic elites, who promise us access to truth or the ability to sort truth from falsehood, may increase in power. For instance, digital auditors who have the ability to identify deepfakes, may (at least among those that still care about truth) be more sought after and more influential. This can have an impact on other values, such as trust. In order to continue to care about truth people may find that they have to place an increased amount of trust in an epistemic elite. This epistemic elite will then need to work to maintain this trust. This may be difficult to do if technology is also impacting on the value of trust. We discuss these potential impacts below.

What are the general lessons we can draw from this? Well, for one thing, digital information technologies are giving us new options that challenge our commitment to the value of truth. Where once it might have been necessary to try to calibrate your beliefs to some shared external reality or some shared set of epistemic norms, this is becoming less necessary as a result of technology. This is an example of pragmatic moral mediation: technology creates new options and new value dilemmas/tradeoffs. The costs of committing to falsity are being reduced; the benefits of committing to truth are not keeping pace. At the same time, these new options affect the perceived scarcity of truth and hence, among some people at least, increase its perceived intrinsic value. This can be seen as an example of hermeneutic moral mediation: the informational infrastructure makes truth seem more morally valuable than it once was. This is, in turn, having an effect on the power of social institutions and organisations that control the flow of information and disinformation. The capacity to speak truth to power is ebbing away; the need to be accountable for truth is disintegrating. Recording police brutality has much less impact in a world of hyperrealistic fake information. Anything that is dissonant to some prevailing orthodoxy can be discounted; anything that is consistent with it can be amplified. If some organisations have an advantage when it comes to wielding the disinformation apparatus, their power will grow. Governments and large technology companies are the obvious example of this. If the infrastructure of disinformation is more widely dispersed, informational anarchy will tend to reign. It would be difficult for truth to sustain its perceived instrumental value in such a world.

Trust

Understanding the value of trust

Let's start, once again, by considering the nature of true as a value. Trust is an integral part of all human activity. In the absence of certainty or immediate person access to the truth we have to choose—consciously or not—who, what and when to trust. We have to trust that others keep their promises; we have to trust authorities to protect our interests; we have to trust an advisor when we are not sure what to do.

Like truth, trust has an important instrumental value. It saves us time and effort, allows us to rely on others to get things done, gives us peace of mind, fosters meaningful intimate relationships, and so forth. In a broader perspective, trust undergirds all human sociality (Churchland, 2011; Sætra, 2021b), and trust shapes interactions between strangers in different societies, and the relationships between citizens and their governments. It could be said, then, that trust is the keystone in our broader value system by facilitating productive cooperation and coordination. If we can trust others, we can enhance our autonomy, happiness, mental well-being, health, relationships and so on. In addition to these instrumental benefits, however, trust also has an important intrinsic value arising from its expressive function in human social relations (Sætra, 2021b). Trust is a way of signaling respect to another person. If we trust someone, we are respecting their honesty, their competence, and their status as a co-equal moral citizen. Given this, it is common to view trust as a necessarily interpersonal value: as something that defines the bond between two or more people (Hawley, 2014, 2019). But trust seems to extend beyond *human* affairs as well. People are inclined to state, for example, that they trust their car, their TV, their phone (Nguyen, 2019). Critics might argue that this is the result of a naive form of anthropomorphism (Reeves & Nass, 1996), and concept creep, and does not involve genuine trust; defenders of this practice argue that it can (Nguyen, 2019; Nickel, 2013).

So the boundaries of trust are contested and this raises an obvious question when it comes to technology: Can we trust machines? If not, can we substitute the value of trust

for some related value in machines? Are these values, in effect, the same thing? To illustrate the problem, consider another value that is closely related to trust: *reliability*. One might say that someone who is trustworthy is also reliable. They conform to our expectations; they follow a pattern. But reliability works in both positive and negative directions: someone can be reliably good or reliably bad. Trustworthiness does not seem to work like this: it is seen as a good thing (Hawley, 2014). So can we trust machines or merely rely on them? Most people accept that trust is, in part, an expectation of (positive) reliability. This implies that trust can be extended to machines. But some have argued that there are distinct reasons underlying that extension of trust to others that exclude machines. For instance, Levine and Schweitzer (2015) argue that there are two sub-types of trust: benevolence-based and integrity-based trust. Trust based on benevolence indicates that the trustee has a reputation for goodness and that they have an inherent desire to help the trustor; trust based on integrity indicates that the trustor believes that the trustee will adhere to "acceptable ethical principles, such as honesty and truthfulness" (Levine & Schweitzer, 2015, p. 99). These two types of trust give rise to distinct *reasons* for expecting some other actor to do as expected: because they have some goodwill toward you or because they are committed to certain ethical principles. Empirical evidence suggests that the two types of trust don't always go together. Prosocial lies, for example, tend to promote benevolence-based trust when intentions are perceived as good, while simultaneously undermining integrity-based trust. Prosocial liars, for example, have been shown to be perceived as more moral than strictly honest individuals (Levine & Schweitzer, 2014). This is where many will argue that machines are no longer able to take part in relationships in which the concept of trust applies (Sætra, 2021b). Few would argue that a garbage compressor is benevolent just because it does as expected, and we suspect that equally few would laud its honesty or decency for doing so. It could, however, be the case that people expect various machines to be able to adhere to a set of codified ethical principles and hence have a kind of integrity-based trust..

While the garbage compressor is a relatively easy case, what about a sophisticated robot? Imagine a robot with cutting edge artificial intelligence (AI) and robotic technology. The robot is a *social* robot, and designed to interact efficiently with human beings through dialogue and mimicking various other human traits and characteristics (Sætra, 2020). You can talk to this robot, the robot may be programmed to want what is best for you, to live up to its word, and to act in ways indicative of goodness in human beings. It may seem, in other words, to be benevolent and to be committed to ethical interactions. Is it possible to have a trusting relationship with such a robot? The answer could go in one of two ways, depending on our understanding of the metaphysical grounding of the properties required for benevolence and integrity. One approach, epitomised by Danaher (2020) and his stance of 'ethical behaviourism' would maintain that if the robot looks and acts like it has benevolence or integrity, then it probably has those properties. A trusting relationship is possible based on performance alone. An alternative, and perhaps more common, approach would see machines as little more than the instruments of human actions (Sætra, 2021a). When we remove the *veil of complexity* introduced by modern AI and its ability to imitate us, we see that these machines are in principle no different from simple machines such as the garbage compressor (Sætra, 2021a). While robots can speak, and can appear to be both funny, loving, and benevolent, this is no more than the result of sophisticated programming and a simulacra of the actions required to be benevolent or have integrity. If we see benevolence and integrity as more than mere surface appearance, and as requiring a more meaningful form of autonomy and more complex motivation, machines—even sophisticated ones—can be argued to be incapable of this. Consequently, whatever reliance we are placing in machines cannot be confused with trust.

This is related to the emerging literature on *overtrust* (perhaps more properly called over-reliance or distorted reliance). This is a disvalue as opposed to a value. This concept applies to our overly optimistic perception of all machines' capabilities (Lee & See, 2004) and could entail that overtrust occurs when I believe my garbage compressor is more reliable than it actually is. This is a disvalue if such overtrust is placed in machines that can jeopardize human safety (Parasuraman & Riley, 1997), such as overly relying on a machine that is supposed to filter poison gases. Recent literature, however, focuses in particular on how humans tend to overtrust *robots*, in particular and more so than other technologies, because of their uncertain place in our ontological schemes (Robinette et al., 2016; Wagner et al., 2018). Overtrusting robots might entail relying on them without appropriate justification, but it might also involve overestimating their intrinsic capacities. In other words, thinking they are proper objects of trust, capable of benevolence and integrity, when, in fact, they are not.

In short, then, trust is an important instrumental and intrinsic value, primarily associated with human social interactions. That said, the boundaries of trust are contested and it is not clear whether it applies beyond human interactions. Trust is often confused with similar, related, values such as reliability. And when taken to an extreme—overtrust—trust can shift from being valuable to being disvaluable.

How technology affects the value of trust

Technology can affect the conceptual understanding of the value of trust (and hence its perceived prevalence and

relevance in our society), and it can also affect how, why, and when humans trust other humans. This leads to a number of impacts on how we promote, pursue and perceive the value of trust. Some of these impacts result in trust being replaced by increased reliance on machines; some result in trust being redistributed between humans and machines; some result in a changed understanding of the value of trust in social life. Let's consider several examples.

First, trust in humans could be replaced by reliance in machines and, in some cases, result in increased distrust of humans. Imagine you are a visitor to a foreign city. You do not know where your hotel is and you are looking for directions. Who can you trust? You might be inclined to trust the person with the tour guide sign as opposed to the person next to him, who looks like a tourist just as yourself. However, had the tour guide not been there, you would perhaps have placed your trust in your co-tourist. New technologies can disrupt this pattern of trusting relationships. In at least some of its applications, AI has become so advanced that machines are today more capable than human beings at providing valuable information in a fast and efficient way (e.g. accurate maps/directions; translations). This leads to a situation in which fellow humans or human experts—previously the best source of truth in a range of cases—are no longer the most accessible or reliable source of expertise. When visiting a new city many people will rely on digital maps and AI recommender systems instead of fellow humans. It's faster and more efficient and avoids awkward social encounters. Similarly, when evaluating a Chess or Go game, even human experts will defer to the judgements of advanced AI systems over their own.

This redistribution of trust is more fundamental than it might at first seem. It's not just that some humans may be seen as untrustworthy; it's that all humans may be perceived as untrustworthy. We have always known that humans are fallible but there is a difference between being the best there is, but fallible, and simply being the best *human*, when machines exceed our capabilities. If machines are consistently more reliable and accurate than humans in certain domains, this will change who we trust in situations where reliable computers are available, reducing trust in humans and substituting it with reliance on machines. It might also simultaneously create a sense of distrust in human abilities in general, and an unwillingness to trust humans even when no machine is readily available, or when no machine exists which exceeds humans at the particular capability or knowledge area in question.

While this goes for trust in cognitive abilities, it might also apply to trust relating to the performance of physical and mechanical tasks. That both machines and other animals surpass us in strength is nothing new, but machines are every day being applied to new settings, and just as they surpass humans at playing chess, machines tend to surpass humans in more serious domains of life as well. In healthcare, for example, AI is now being used to diagnose dementia and in robot surgery (Ding et al., 2019). If the most capable surgeon, for example, is now a robot, who will trust a human surgeon? These considerations are speculative, and technologists often overstate the ability of medical AI and surgical robots, but the future redistribution of trust, away from even expert humans, is plausible based on what has happened in other domains such as chess.

Another way in which technology might affect trust is when machines allegedly reduce or eliminate the need for trust in other humans altogether. Examples of such technologies could include improved lie detectors, and facial and emotion recognition software (Zhang et al., 2020). While some argue that, for example, computer vision-enabled emotion recognition is little more than modern physiognomy or phrenology (Stark & Hutson, 2021), others are already using AI that detects frustration and (allegedly) identifies basic emotions (Zhang et al., 2020). While such systems might be biased and far from perfect, this need not matter much from the perspective of evaluating technology's effect on trust. If such systems work *reasonably* well, that might be enough. In fact, even if they don't work at all, but people *believe* they work, that might be enough to change our attitude toward the value of trust. Whenever people think that they can refer to some form of software, for example, to detect whether someone is lying, the need to trust people is radically reduced. Recent empirical work on algorithmic decision-making systems seems to confirm this trend. For instance, a series of studies by Bigman et al., (2022), found that people are likely to less outraged at discriminatory algorithmic decisions because they are more likely believe that algorithmic systems make unbiased decisions, and hence more likely to trust their outputs.

This, paradoxically, could increase the perceived intrinsic value of trust in humans. If you are willing to trust someone, even when a technological alternative to a human exists, this suggests that you must really respect them. The value of the expressive signal goes up. This is similar to the effect of digital technology on the intrinsic value of truth. But it may also reduce our tendency to value trust in general. "Trust-free" technologies such as blockchains (Hawlitschek et al., 2018), are particularly relevant in this context. Ostensibly, the creators and boosters of this technology promote it as an alternative to trust: we replace our trust in human intermediaries (such as banks and payment companies) with an acceptance of the consensus algorithm of a blockchain infrastructure. However, whether this results in truly a trust-free technology is highly disputed (de Filippi et al., 2020). Some argue that the blockchain simply results in the redistribution of trust, away from traditional third-party intermediaries towards those that own and control the blockchain infrastructure. What applies for blockchain could

also apply to other, allegedly trust-free technologies: reliance on them could simply redistribute trust away from one group of humans and onto another, technically literate group that understands the technologies in question. This is similar to the phenomenon we previously mentioned in relation to epistemic elites and the detection of deepfakes. Either way, technology can have a profound impact on social networks, role-related duties, and power relationships in society. It does so by undermining and disrupting the power of traditional trust networks, either by eliminating the need for trust in humans (and replacing it with reliance on machines) or redistributing trust onto different groups of humans (technological elites) or technology itself.

Yet another mechanism of change is related to the use of deceptive machines, such as social robots that are designed to mimic or copy human behavioural cues (Sætra, 2021b). Even if one is unwilling to accept that social robots are inherently deceptive, they are at the very least interacting with human beings in ways that trigger psychological response mechanisms that were previously reserved to other humans (Sætra, 2020). All humans have become who they are today through a lifetime of learning from interaction with human beings. Some might have grown up in a very safe environment and learnt to trust indiscriminately, while others may have had a more challenging life and have ended up hardly trusting anyone at all. What happens if this social learning involves many interactions with robots? One concern might be that we learn something different when we interact with robots than when we interact with humans (Sætra, 2020). Unless one is willing to argue that machines perfectly mimic human beings, in the myriad of subtle and fundamental ways in which human social interaction occurs, this argument represents a reasonably likely effect of robots on our social behaviour. But this can have profound impacts on trust as well. Sætra (2021b) argues that if trust, as a relational concept does not apply to robots, and is not assumed to apply to them, having lots of social interactions shaped by robots might, once again, entail downplaying the role of trust in general in human society. Rather than trust being seen as an essential or core instrumental social value (the glue that binds together cooperative relations) it may be seen as unnecessary and discardable. We can drop trust in favour of reliance and still unlock many of the same values. But since reliance is a different kind of value, which does not share the same intrinsic expressive content as trust, we may lose an important intrinsic social value, namely, that of respect for others, and this may have a negative impact on the perceived value of relationships more generally.

Two additional consequences of robot deception are relevant in light of the preceding considerations. First, if one assumes that human beings learn from their experiences, as we do, whenever an individual recognizes that they have been deceived, this might lead to this person being less trustful in future interactions. However, not all individuals will recognize that they have been deceived and so some may continue to be as trusting as ever before. This will result in a differential distribution of trust in society: some will see the rise of social robots as a threat to trust and a reason for mistrust; some will continue as before. Society may polarise into those that disvalue trust and become more suspicious of everything around them, and those that continue to trust (perhaps even overtrust) humans and other technologies. Secondly, a more speculative concern relates to the potential long-term evolutionary changes that might follow in a situation where those least likely to trust (both humans and technology) are most successful. In an evolutionary time frame, this might lead to a situation in which humans as a species will be characterised by a less trustful nature (Sætra, 2020). The work of behavioural scientists such as Michael Tomasello, for instance, suggest that humans are innately trusting and cooperative, much more so than our primate cousins. Although this is partly a learnt behaviour, it also seems to be partly genetic (Tomasello, 2016). Over a long enough timescale of interactions with robots, this innate disposition to trust may be eroded. In the meantime, there is plenty of scope for robots (and, perhaps more importantly their manufacturers) to exploit the innate disposition to trust.

Finally, technology might change trust by changing our moral perception of ourselves and others. This could happen through the phenomenon of 'robotomorphy' (Sætra, 2021c), which is a form of hermeneutic moral mediation. Anthropomorphism describes how we attribute human qualities to other entities, such as robots. Robotomorphy is a companion concept which describes how we also tend to attribute robot qualities to human beings. Seeing ourselves as machines has a very long history, and goes back to philosophers such as Hobbes (1946[1651]) and Le Mettrie (1912[1747]), who established and used a mechanistic philosophy to argue that humans are little more than advanced machines. Fast forward to the modern era, and there is no shortage of scientists that liken the human brain to a computer (Piccini & Bahar, 2013), and no shortage of fiction writers that use intelligent machines explore the nature of the human condition (Cave et al., 2020). But metaphors are dangerous. They can mislead or misrepresent reality in both significant and subtle ways. The danger of robotomorphy is that the more we see ourselves as a kind of machine, the less need there is for the concept of trust. As already discussed, we might *rely* on machines, but trust has usually been used to refer to something deeper and exclusive to being with a mind, intentions, drives, and not least a free will. Trust is something that entails a certain element of the unknowable and mystical. Increased robotomorphy might dispel this mystery and change trust in human beings into something more akin to a question of whether

or not we can *rely* on each other just as we rely on a car or a dishwasher.

Conclusion: lessons learned

Having examined our two case studies, it remains to consider whether or not there are similarities in how technology affects trust and truth, and if there are general lessons to be learned here about how technology may impact values in the future.

The two values we have considered are structurally similar and interrelated. They are both intrinsically and instrumentally valuable. They are both epistemic and practical in nature: we value truth and trust (at least in part) because they give us access to knowledge and help us to resolve the decision problems we face on a daily basis. We also see, in both case studies, similar mechanisms of value change at work. The most interesting, to our minds, are the following:

- *Technology changes the costs associated with accessing certain values, making them less or more important as a result* Digital disinformation technology increases the cost of finding out the truth, but reduces the cost of finding and reinforcing a shared identity community; reliable AI and robotics gives us an (often cheaper and more efficient) substitute for trust in humans, while still giving us access to useful cognitive, emotional and physical assistance.
- *Technology makes it easier, or more attractive to trade off or substitute some values against others* Digital disinformation technology allows us to obviate the need for finding out the truth and focus on other values instead; reliable machines allow us to substitute the value of reliability for the value of trust. This is a function of the plural nature of values, their scarcity, and the changing cost structure of values caused by technology.
- *Technology can make some values seem more scarce (rare, difficult to obtain), thereby increasing their perceived intrinsic value* Digital disinformation makes truth more elusive, thereby increasing its perceived value which, in turn, encourages some moral communities to increase their fixation on it; robots and AI make trust in humans less instrumentally necessary, thereby increasing the expressive value of trust in others.
- *Technology can disrupt power networks, thereby altering the social gatekeepers to value* to the extent that we still care about truth, digital disinformation increases the power of the epistemic elites that can help us to access the truth; trust-free or trust-alternative technologies can disrupt the power of traditional trusted third parties (professionals, experts etc.) and redistribute power onto technology or a technological elite.

We also see, in both cases, first and second-order value effects. Technologies first impact on how we make decisions in relation to certain values, the metaphors or concepts we use to understand those values, and then on our relationships with one another, our perceived duties to one another and the power we hold over one another. For instance, we choose to rely on machines rather than trust humans, this leads us to question the nature of trust and whether it can be applied to machines, and it also affects how we interact with fellow humans and the perceived (and actual) power of humans and technology. It's plausible to assume that similar mechanisms of value change will be at work in other case studies.

There are also important overlaps and synergies to consider in relation to the technological disruption of both values. We have not commented on these in much depth in the foregoing analysis; we have, instead, treated the two case studies as being largely independent, occasionally noting connections between. It is worth noting the synergies and overlaps in more detail now. First, and most obviously, there is an interesting tension between the two values and the possible effect of technology on them. We have argued that digital technology undermines the search for truth both by potentially altering the objective reality to which we are trying to conform our beliefs, increasing the volume of information and disinformation, and undermining the epistemic processes we use to verify our beliefs. We have argued that machines, particularly AI and robotics, replace the need for trust in humans with reliance in machines. But if machines are seen as tools of disinformation, it could well be that we are reluctant to rely on them in the stead of humans. We may trust an expert chess AI, but not an (alleged) expert political policy AI. In other words, the disruptive effect of technology on one value (truth) may block the disruptive effect of technology on another value (trust). It could also work the other way, of course. Increased reliance in machines, particularly in cognitive affairs, could undermine the disinformation effects of technology on truth. If we are convinced that the machines are the path to enlightenment, then perhaps truth can retain its instrumental and intrinsic social value (at the expense of trust in humans).

These are complicated matters. One meta-lesson of our two case studies is that in a world of plural values, and plural technologies, the impact of technology on moral change can be complex and interactive. Technology rarely affects one value in isolation from the others. Greater scrutiny of these complex and interactive effects would be beneficial if we are to improve our ability to anticipate and plan for technology-induced moral change.

References

Abidin, C. (2016). Please subscribe!: Influencers, social media, and the commodification of everyday life. Doctor of Philosophy, The University of Western Australia. https://doi.org/10.26182/5ddc899d698cb

Adshade, M. (2013). *Dollars and sex: How economics influences sex and love*. Chronicle Books.

Anthis, J. R., & Paez, E. (2021). Moral circle expansion: A promising strategy to impact the far future. *Futures, 130*, 102756. https://doi.org/10.1016/j.futures.2021.102756

Appiah, K. A. (2010). *The Honor Code: How moral revolutions happen*. WW Norton.

Arriagada, A., & Bishop, S. (2021). Between commerciality and authenticity: The imaginary of social media influencers in the platform economy. *Communication, Culture and Critique, 14*(4), 568–586. https://doi.org/10.1093/ccc/tcab050

Asay, J. (forthcoming). Arne Næss's experiments in truth. *_Erkenntnis_*. https://philpapers.org/rec/ASAANE

Barley, S. (2020). *Work and technological change*. OUP.

Barnard, R., & Ulatowski, J. (2019). Does anyone really think that ⌜φ⌝ is true if and only if φ? In A. Aberdein & M. Inglis (Eds.), *Advances in experimental philosophy of logic and mathematics* (pp. 145–171). Bloomsbury Academic.

Barnard, R., & Ulatowski, P. (2013). Truth, correspondence and gender. *Review Philosophical Psychology, 4*, 621–638.

Bigman, Y. E., Wilson, D., Arnestad, M. N., Waytz, A., & Gray, K. (2022). Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology. General*. https://doi.org/10.1037/xge0001250

Bortolotti, L. (2018). Optimism, agency, and success. *Ethical Theory and Moral Practice, 21*, 521–535. https://doi.org/10.1007/s10677-018-9894-6

Brady, W. J., et al. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances, 7*(33), eabe5641. https://doi.org/10.1126/sciadv.abe5641

Burke, E. (1790). *Reflections on the revolution in France*. J. Dodsley.

Calhoun, C. (2015). *Moral aims: Essays on the importance of getting it right and practicing morality with others*. Oxford University Press.

Capps, J. (2019). The pragmatic theory of truth. In *The Stanford encyclopedia of philosophy (Summer 2019 Edition)*, (E. N. Zalta, Ed.), https://plato.stanford.edu/archives/sum2019/entries/truth-pragmatic/

Cave, S., Dihal, K., & Dillon, S. (Eds.). (2020). *AI narratives: A history of imaginative thinking about intelligent machines*. Oxford University Press.

Chalmers, D. (2022). *Reality+*. Penguin.

Churchland, P. S. (2011). *Braintrust: What neuroscience tells us about morality*. Princeton University Press.

Churchland, P. S. (2019). *Conscience: The origins of moral intuition*. WW Norton.

Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour, 1*(11), 769–771.

Danaher, J. (2020). Welcoming robots into the moral circle: A defence of ethical behaviourism. *Science and Engineering Ethics, 26*(4), 2023–2049.

Danaher, J. (2021). Axiological futurism: The systematic study of the future of values. *Futures, 132*, 102780. https://doi.org/10.1016/j.futures.2021.102780

David, M. (2020). The correspondence theory of truth. In *The Stanford Encyclopedia of Philosophy (Winter 2020 Edition)* (E. N. Zalta, Ed.), https://plato.stanford.edu/archives/win2020/entries/truth-correspondence/

Debrabander, F. (2020). *Life after privacy: Reclaiming democracy in a surveillance society*. Cambridge University Press.

Ding, Y., Sohn, J. H., Kawczynski, M. G., Trivedi, H., Harnish, R., Jenkins, N. W., et al. (2019). A deep learning model to predict a diagnosis of Alzheimer disease by using 18F-FDG PET of the brain. *Radiology, 290*(2), 456–464.

Dogramaci, S. (2020). What is the function of reasoning? On Mercier and Sperber's argumentative and justificatory theories. *Episteme, 17*(3), 316–330.

Duffy, A., & Kang, H. Y. P. (2020). Follow me, I'm famous: travel bloggers' self-mediated performances of everyday exoticism. *Media, Culture & Society, 42*(2), 172–190. https://doi.org/10.1177/0163443719853503

Fallis, D. (2021). The epistemic threat of deepfakes. *Philosophy & Technology, 34*, 623–643. https://doi.org/10.1007/s13347-020-00419-2

Filippi, P. D., Mannan, M., & Reijers, W. (2020). Blockchain as a confidence machine: The problem of trust & challenges of governance. *Technology in Society, 62*, 101284. https://doi.org/10.1016/j.techsoc.2020.101284

Flexner, A. (1939). The usefulness of useless knowledge. *Harpers, 179*, 544.

Galef, J. (2021). *The scout mindset*. Piatkus.

Glanzberg, M. (2021). Truth. In *The Stanford encyclopedia of philosophy* (Summer 2021 Edition), (E. N. Zalta Ed.). https://plato.stanford.edu/archives/sum2021/entries/truth/

Greenwood, J. (2020). *Evolving households: The imprint of technology on life*. MIT Press.

Gurri, M. (2018). *The revolt of the public*. Stripe Press.

Hartzog, W. (2018). *Privacy's blueprint: The battle to control the design of new technologies*. Harvard University Press.

Hawley, K. (2014). Trust, distrust and commitment. *Nous, 48*(1), 1–20.

Hawlitschek, F., Notheisen, B., & Teubner, T. (2018). The limits of trust-free systems: A literature review on blockchain technology and trust in the sharing economy. *Electronic Commerce Research and Applications, 29*, 50–63.

Hobbes, T. (1946[1651]). *Leviathan*. Basil Blackwell.

Hopster, J. (2022). Future value change: Identifying realistic possibilities and risks. *Prometheus*, forthcoming, https://philpapers.org/archive/HOPFVC.pdf

Horwich, P. (2006). The value of truth. *Nous, 40*(2), 347–360.

Hund, E., & McGuigan, L. (2019). A shoppable life: Performance, selfhood, and influence in the social media storefront. *Communication, Culture and Critique, 12*(1), 18–35. https://doi.org/10.1093/ccc/tcz004

Jefferson, A., Bortolotti, L., & Kuzmanovic, B. (2017). What is unrealistic optimism? *Consciousness and Cognition, 50*, 3–11. https://doi.org/10.1016/j.concog.2016.10.005

Klenk, M., et al. (2022). Recent work on moral revolutions. *Analysis, 82*(2), 354–366. https://doi.org/10.1093/analys/anac017

Koyama, M., & Rubin, J. (2022). *How the world became rich*. Polity Press.
Kudina, O. (2019). The technological mediation of morality: Value dynamism, and the complex interaction between ethics and technology. PhD Thesis, University of Twente. https://doi.org/10.3990/1.9789036547444
Kudina, O., & Verbeek, P.-P. (2019). Ethics from within: Google glass, the Collingridge dilemma, and the mediated value of privacy. *Science, Technology, & Human Values, 44*(2), 291–314. https://doi.org/10.1177/0162243918793711
La Mettrie, J. J. O. (1996[1748]). La Mettrie: Machine man and other writings. Cambridge University Press.
Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80.
Levine, E. E., & Schweitzer, M. E. (2014). Are liars ethical? On the tension between benevolence and honesty. *Journal of Experimental Social Psychology, 53*, 107–117.
Levine, E. E., & Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes, 126*, 88–106.
Levitsky, S., & Ziblatt, D. (2018). *How Democracies Die*. Viking.
Mascarenhas, MascarenhasS. (2019). Review of Mercier and Sperber's the enigma of reason. *Teorema: Revista Internacional De Filosofía, 38*(1), 97–106.
Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Harvard University Press.
Mercier, H., & Sperber, D. (2019). Replies to critics. *Teorema: Revista Internacional De Filosofía, 38*(1), 139–156.
Mill, J. S. (1859). *On liberty*. Parker and Son.
Murray, S. L., & Holmes, J. G. (1997). A leap of faith? Positive illusions in romantic relationships. *Personality and Social Psychology Bulletin, 23*(6), 586–604.
Naess, A. (1938)."Truth" as conceived by those who are not professional philosophers (Skrifter Utgitt avDet Norske Videnskaps-Akademi I Oslo Il. Hist.-Filos. Klass 1938 No. 4). I Kommisjon HosJacob Dybwad.
Nelson-Field, K. (2020). *The attention economy and how media works: Simple truths for marketers*. Palgrave-Macmillan.
Nguyen, C. T. (2019). Trust as an unquestioning attitude. *Oxford studies in epistemology,* forthcoming. https://philarchive.org/rec/NGUTAA
Nickel, P. J. (2013). Trust in technological systems. In M. J. de Vries, S. O. Hansson, & A. W. M. Meijers (Eds.), *Norms in technology: Philosophy of engineering and technology.* (Vol. 9). Springer.
Nickel, P. J. (2020). Disruptive innovation and moral uncertainty. *NanoEthics, 14*(3), 259–269. https://doi.org/10.1007/s11569-020-00375-3
Nickel, P., Kudina, O., & Van de Poel, I. (2021). Moral Uncertainty in technomoral change: Bridging the explanatory gap. *Perspectives in Science*. https://doi.org/10.1162/posc_a_00414
Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230–253.
Pelto, P. J. (1973). *The snowmobile revolution: technology and social change in the Arctic*. Waveland Press.
Piccinini, G., & Bahar, S. (2013). Neural computation and the computational theory of cognition. *Cognitive Science, 37*(3), 453–488.
Prochownik, K. (2019). Three questions about the social function of reason. *Teorema: Revista Internacional De Filosofía, 38*(1), 77–86.
Rauch, J. (2021). *The constitution of knowledge*. Brookings Institution Press.
Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.
Reuter, K., & Brun, G. (2022). Empirical studies on truth and the project of re-engineering truth. *Pacific Philosophical Quarterly*. https://doi.org/10.1111/papq.12370
Riley, C. L. (Ed.). (2020). *The free speech wars*. Manchester University Press.
Rini, R. (2020). Deepfakes and the epistemic backdrop. *The Philosophers' Imprint, 20*(24), 1–16.
Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016, March). Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE international conference on human-robot interaction* (*HRI*) (pp. 101–108).
Ross, W. D. (1930). *The right and the good*. Clarendon Press.
Runciman, D. (2018). *How democracy ends*. Profile Books.
Sætra, H. S. (2020). The parasitic nature of social AI: Sharing minds with the mindless. *Integrative Psychological & BehavioralScience*, *54*(2), 308.
Sætra, H. S. (2021a). Confounding complexity of machine action: A Hobbesian account of machine responsibility. *International Journal of Technoethics*. https://doi.org/10.4018/IJT.20210101.oa1
Sætra, H. S. (2021b). Social robot deception and the culture of trust. *Paladyn: Journal of Behavioural Robotics*. https://doi.org/10.1515/pjbr-2021-0021
Sætra, H. S. (2021c). Robotomorphy: Becoming our creations. *AI and Ethics*. https://doi.org/10.1007/s43681-021-00092-x
Schiavon, C. C., et al. (2017). Optimism and hope in chronic disease: A systematic review. *Frontiers in Psychology, 7*, 2022. https://doi.org/10.3389/fpsyg.2016.02022
Stanovich, K. (2021). *The bias that divides us*. MIT Press.
Stark, L., & Hutson, J. (2021). Physiognomic artificial intelligence. *SSRN*. https://doi.org/10.2139/ssrn.3927300
Swierstra, T., Stemerding, D., & Boenink, M. (2009). Exploring techno-moral change: The case of the obesitypill. In P. Sollie & M. Düwell (Eds.), *Evaluating new technologies. The international library of ethics, law and technology.* (Vol. 3). Dordrecht: Springer.
Swierstra, T., & Waelbers, K. (2012). Designing a good life: A matrix for the technological mediation of morality. *Science and Engineering Ethics, 18*(1), 157–172. https://doi.org/10.1007/s11948-010-9251-1
Taylor, S., & Brown, J. D. (1994). Positive illusions and well-being revisited: Separating fact from fiction. *Psychological Bulletin, 116*(1), 21–27.
Tomasello, M. (2016). *A natural history of morality*. Harvard University Press.
van de Poel, I. (2021). Design for value change. *Ethics and Information Technology, 23*, 27–31. https://doi.org/10.1007/s10676-018-9461-9
van de Poel, I., & Kudina, O. (2022). Understanding technology-induced value change: A pragmatist proposal. *Philosophy & Technology 35,* 40 (2022). https://doi.org/10.1007/s13347-022-00520-8
van de Poel, I., & Taebi, B. (2022). Value change in energy systems. *Science, Technology, & Human Values, 47*(3), 371–379. https://doi.org/10.1177/01622439211069526.
Verbeek, P. P. (2012). *Moralizing technology: Understanding and designing the morality of things*. University of Chicago Press.
Verbeek, P.-P. (2013). The moral status of technical artefacts. *Philosophy of Engineering and Technology*. https://doi.org/10.1007/978-94-007-7914-3_5
Wagner, A. R., Borenstein, J., & Howard, A. (2018). Overtrust in the robotic age. *Communications of the ACM, 61*(9), 22–24.
White, L., Jr. (1962). *Medieval technology and social change*. OUP.
Whiting, D. (2013). The good and the true (or the bad and the false). *Philosophy, 88*(2), 219–242. https://doi.org/10.1017/s0031819113000260

Williams, E. G. (2015). The possibility of an ongoing moral catastrophe. *Ethical Theory and Moral Practice, 18*(5), 971–982. https://doi.org/10.1007/s10677-015-9567-7

Williams, J. (2019). *Stand out of our light: Freedom and resistance in the attention economy*. Cambridge University Press.

Wu, T. (2016). *The attention merchants*. Knopf.

Zhang, J., Yin, Z., Chen, P., & Nichele, S. (2020). Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion, 59*, 103–126.