Evolutionary Psychology and Artificial Intelligence

Wilson, Holly; Rauwolf, Paul; Bryson, Joanna J.

The SAGE Handbook of Evolutionary Psychology

Published: 25/11/2020

Peer reviewed version

Evolutionary Psychology and Artificial Intelligence: The Impact of Artificial Intelligence on Human Behaviour

Keywords: Trust; Artificial Intelligence (AI); Cooperation; Transparency; Agent Based Modelling; Information Cost; Technology Policy; Cultural Evolution; Information Communication Technology (ICT)

Holly Wilson
University of Bath
hlw69@bath.ac.uk

Paul Rauwolf
Bangor University
p.rauwolf@bangor.ac.uk

Joanna Bryson
Hertie School of Governance
jjb@alum.mit.edu

Abstract

Artificial Intelligence (AI) presents a new landscape for humanity. Both what we can do, and the impact of our ordinary actions is changed by the innovation of digital and intelligent technology. In this chapter we postulate how AI impacts contemporary societies on an individual and collective level. We begin by teasing apart the current actual impact of AI on society from the impact that our cultural narratives surrounding AI has. We then consider the evolutionary mechanisms that maintain a stable society such as heterogeneity, flexibility and cooperation. Taking AI as a prosthetic intelligence, we discuss how—for better and worse—it enhances our connectivity, coordination, equality, distribution of control and our ability to make predictions. We further give examples of how transparency of thoughts and behaviours influence call-out culture and behavioural manipulation with consideration of group dynamics and tribalism. We next consider the efficacy and vulnerability of human trust, including the contexts in which blind trust in information is either adaptive or maladaptive in an age where the cost of information is decreasing. We then discuss trust in AI, and how we can calibrate trust as to avoid over-trust and mistrust adaptively, using transparency as a mechanism. We then explore the barriers for AI increasing accuracy in our perception by focusing on fake news. Finally, we look at the impact of information accuracy, and the battles of individuals against false beliefs. Where available, we use models drawn from scientific simulations to justify and clarify our predictions and analysis.

1 INTRODUCTION

Artificial intelligence (AI) impacts our behaviour. Intelligence, whilst not in itself defining humanity, is one of our key characteristics, inextricably linked to everything from our implicit survival strategies to our explicit self-concepts. Intelligence is also integral to the social institutions on which contemporary existence depends. AI is a set of technologies that extend human intellectual capacities such as perception, action, categorisation and pattern recognition. Some systems, termed *autonomous*, may do all of these things at once without human intervention, though only after human or human-institutional inception.

Determining the impact of AI is imperative for two reasons: for empowering us to adapt to and optimise our existence with AI; and for developing AI and AI policies which maximise benefit and minimise harm to our society. Due to the myriad definitions of AI, we begin by establishing what we mean by the term, at least in the scope of this chapter. Intelligence is the capacity to do the right thing at the right time; thus artificial intelligence refers to non-living artefacts that demonstrate such capacities (Bryson, 2019). By this definition, AI has been prevalent for decades, albeit less advanced than at present, and not so apparent to the public eye. Present awareness of AI has been magnified by two different recent outcomes: first the sudden prevalence of anthropomorphic capacities such as conversational-speech recognition and generation, or automobile driving; and second the use of AI technology as part of a global assault on democracies, and through them key institutions to maintaining peace under the present global order, such as the EU and the North Atlantic Treaty Organization (NATO).1

We focus here on how behaviour is shaped by contemporary intelligent technology; first by our cultural understanding of AI, then by the technological reality of AI. Many of our drives and behaviours, such as tribalism, sex, and resource procurement, are sculpted by thousands of years of evolution in disparate environments (McDonald et al., 2012; Buss and Schmitt, 1993; Kramer and Ellison, 2010). Therefore, in this chapter we explore from an evolutionary perspective how AI – ubiquitous in the modern world – impacts both individual and collective human behaviour. We put emphasis on the predictions from several published models that explain how information transmission facilitates intelligence between intelligent agents. We consider how these theories predict changes to individual and collective behaviour as the information transmission is magnified or its quality improved. We then compare, at least qualitatively, these predicted alterations to present societal trends. We finish with recommendations for guiding AI development and interaction to maximise adaptation, progression and overall benefit to the individual and society.

More specifically, in Section 2 we initially tease apart the current actual impact of AI on society from the impact that our cultural narratives surrounding AI have. We consider the evolutionary mechanisms that maintain a stable society such as heterogeneity, flexibility and

cooperation. Given that AI can constitute a prosthetic intelligence, we discuss the consequences of how it enhances our connectivity, coordination, equality, distribution of control and our ability to make predictions. We also give examples of how transparency of thoughts and behaviours may influence call-out culture, as well as behavioural manipulation with consideration of group dynamics and tribalism. In Section 3, we discuss how AI may exacerbate the societal problems created by inequality. We place focus on the vulnerability of human trust in Section 4. We consider the contexts in which blind trust in information is either adaptive or maladaptive in an age where the cost of information is decreasing. We then bring the focus back onto trust in AI, and how we can calibrate trust so as to avoid over-trust and mistrust adaptively, using transparency as a mechanism. Finally, in Section 5, we explore the barriers to AI increasing accuracy in our perception by focusing on fake news. We consider the impact of information accuracy and the battles of individuals against false beliefs.

2 IMPACTS OF AI THUS FAR

The narratives within a culture can have as much impact on behaviour as at least some objective realities (Hammack, 2008). For this reason, we begin by examining the current narrative surrounding AI. According to Social Representation Theory (SRT), when we encounter a new or unknown phenomenon, we construct a representation of it based on collective narratives and interpersonal communication (Moscovici, 1981, 2001). It seems clear that public gaps in understanding AI are often filled by fear-mongering entertainment shows like Black Mirror, magazine articles on the feats of Alpha Go, and propaganda from businesses (Elish and Boyd, 2018). Indeed, media exposure to science fiction has been found to predict fear of AI above and beyond demographic variables (Liang and Lee, 2017).

Problematically, such representations assume AI with capacities far beyond the current feasibility – and in many cases, beyond the computationally tractable – instilling awe and fear (Bryson and Kime, 2011). An investigation of narratives surrounding the impact of AI revealed that the most common visions elicited anxiety (Cave et al., 2019). One such vision was that by becoming over-reliant on AI and machines, we will replace the need for humans in jobs, relationships and socialising. Emotional arousal increases the efficacy of information spread (Berger, 2011); this mechanism is posited to have evolved to transmit fitness-relevant information; i.e. information relevant to survival which can help organisms avoid dangers (Nairne et al., 2009) or direct resources within a population (Teste et al., 2009). Yet the adaptive benefits of this in contemporary cultures are questionable, now that the mechanism is known and manipulable. These narratives may distract from or obscure the real problems and utility of AI, resulting in sub-optimal allocation of energy and resources (Bryson and

Kime, 2011; Elish and Boyd, 2018). In the next few sections, we discuss the mixture of benefits and potential problems we face with AI.

2.1 Flexibility, Cooperation, Coordination, Perception: Humanity's Survival Mechanisms

AI *does* present a new landscape for humanity – yet, despite popular rhetoric, this does not necessarily pose an existential threat (Gent, 2015; Müller and Bostrom, 2016). Throughout our evolutionary history, humans have succeeded in adapting to changing environments (Gilligan, 2007; Richerson et al., 2005). Machines, in contrast, are typically fragile and short lived. As such, dangerous machines or technology are unlikely to be allowed to persist in their damaging behaviour long enough to destroy humanity as a whole, although arguably technologically mediated impacts such as climate change or hate crimes may already be costing lives. Cooperation and flexibility by means of heterogeneity (diversity) are two mechanisms that enable adaptation, survival and progress in unstable, changing environments (Brown et al., 2011; Lahr, 2016; Smaldino et al., 2013). Here we discuss the impact of AI on human behaviour within the context of these two mechanisms.

There are two core hypotheses as to the role our nervous systems evolved to fill: the sensory-motor view, to link senses to actions; and the action-shaping view, to coordinate the body's micro acts into macro acts (Godfrey-Smith, 2017). These hypotheses are by no means exclusive of each other. Likewise, our development of AI, amongst other things, has greatly enhanced our abilities to sense, act and coordinate. Our society has become increasingly complex; we are connected world-wide, with perturbations in one part of our global system impacting many others. AI can be considered as our prosthetic nervous system, a tool we have developed to selectively mediate the strength of edges between each node. As an example, we can deploy hardy robots to explore subterranean and underwater environments inaccessible to humans, acting and perceiving on our behalf (Siles and Walker, 2009), growing new edges of our agency. Machine learning (ML), the ability to learn to perceive and categorise patterns based on input data, also constitutes a prosthetic perception. With adequate computational resources, ML for example allows us to search across larger ranges of data than a human might otherwise be able to internalise, and to consider more candidate patterns. Enhanced perception, assuming accuracy, means a species has more knowledge with which to react better to events in its environment. This constitutes an increase in human collective intelligence (Eagle and Pentland, 2003).

2.2 AI Increases Connectivity Which Facilitates Coordination but Also Transparency

An enhanced capacity for coordination also results from the increased connectivity that AI and Information and Communication Technologies (ICT) more generally facilitate. These technologies afford communication and coordination on a scale that human societies have never encountered before. This can be expected to have – and be having – myriad impacts on collective and individual behaviour, not all of which we have yet recognised (Bryson, 2015). Through increasing the number of individuals we can connect with, and decreasing temporal and spatial constraints on doing so, AI creates a capacity for highly agile cooperation. Cooperation and group-level investment as a whole is known to increase with capacity to communicate, because this capacity allows for the increased probability of discovering mutually beneficial equilibria (Roughgarden et al., 2006). This cooperation may occur at historically large scales, but also at small scales, and with higher frequency of change in aggregation and direction. Social media platforms, for example, have facilitated mass organisation of not only protests but also disaster recovery which may not otherwise have been feasible (Gerbaudo, 2018; Starbird and Palen, 2011; Vieweg et al., 2010). Additionally, increased access to knowledge enables us to predict threats or more quickly become aware of disasters (Chavan and Khot, 2013).

Connectivity also has the effect of increasing behaviour transparency. Social media pages reveal the 'likes', 'dislikes' and actions of a population. Problematically within this context, humans are often driven to take on characteristics of a group, and strive to behave according to group norms (Terry and Hogg, 1996), which may have a homogenising or polarising effect, particularly during periods when competition and identity politics are steep (McCarty et al., 2016). Before we were so connected via the digital world, the group norms we had access to were of far smaller scale. This facilitated diversity – a global heterogeneity of group norms. Now, our access to the large scale, combined with behaviour transparency, is widely believed to homogenise behaviours and preferences (Morris, 2002). This is true even without taking into account potential conformity induced by physical, political, or economic threats for unacceptable behaviour, which we discuss further below. However, the data created by our digital activity is also made transparent for use by the political and business, as well as the social, realms. This can have positive impacts as well, if it is used to, for example, provide better public services or ensure greater customer satisfaction.

2.3 AI Facilitates Behaviour Prediction and Manipulation

Ordinary collectives such as companies use algorithms on big data sets to predict our behaviour with increasing accuracy (Zuboff, 2015). This seems qualitatively different to our already adept prediction capacities: to navigate our social world, we use mental models of people to infer and predict the beliefs, actions and intentions of others (Bradford et al., 2015).

This ability, alongside language (Smith, 2010), has been critical for facilitating our species' large-scale social cooperation. However, AI has enhanced our ability to predict the individual and collective beyond past capabilities. This can aid us to better allocate our time and resources. For example, by anticipating that the number of people about to use a road is beyond capacity, a navigation app may direct a proportion of people along a different route, altering their behaviour advantageously for the collective. At an individual level, the 'optimal' level of sleep (Hao et al., 2013), exercise (Spring et al., 2013), socialising and nutrition (Franco et al., 2016) can be predicted then recommended. Unfortunately, businesses looking to maximise profits will tailor their product – or just monetising efforts related to their product – to the individual based on predictions. This can be with an incentive to reshape behaviour to make it even more easily predictable, or otherwise less expensive for the company, for example when insurance companies demand digital access to evidence of healthy living (Raber et al., 2019). This, whilst not explicitly detrimental to cooperation, may increase homogeneity and therefore reduce group advantages and societal robustness (Fisher, 1930; Shi et al., 2019). It also speaks to an additional issue with behaviour transparency: capacity for behavioural control.

Both the aforementioned examples of collective and individual behaviour tweaking and recommendations indicate a shift in autonomy; it seems AI may be increasing collective agency but decreasing opportunity and drive for individual decision-making. In fact, a survey of 970 respondents revealed that a core concern surrounding AI technologies is the loss of human agency and input into decisions (Anderson et al., 2018). At one level, behavioural control may or may not shift to autonomous artificial agents, but either way at a higher level it shifts to the individuals or organisations who can monitor data and deploy any such agents. This enables execution of potentially regressive social policing, albeit some well-intentioned. Take for example, the rise of helicopter parenting (Lee et al., 2014), or AI-powered predictive policing systems (Meijer and Wessels, 2019). We do not claim such behaviour manipulation is unique to AI. There are varying views on whether manipulation techniques (not all of which necessarily use AI) are ethical, when used for example to promote health (Behavioural Insights Team, 2010) or reduce debt (Behavioural Insights Team, 2012). We see the sense in policies (such as that of the Institute of Electrical and Electronics Engineers (IEEE), 2019) recommending that behaviour manipulation may be ethical in contexts where all the following hold: it can be beneficial for the individual and/or society, transparency is provided as to the nature of the manipulation, and the subject or a responsible adult representative of the subject has consented. This (arguably, cf. Simkulet, 2019) leaves the individual some control over the decision, at least at a higher level. For example, clinicians, especially under cognitive load, can demonstrate bias towards ethnic-minority patients, resulting in sub-optimal interaction, diagnosis and treatment (Stone and Moskowitz, 2011). Evidence suggests that two tasks – perspective taking or categorising oneself to be in a shared group with the

ethnic minority – can reduce bias. In this scenario, consented behavioural manipulation could involve an app-based intervention, where the clinician chooses to engage with a bias-reduction task prior to seeing the patient.

This same public or semi-private – and sometimes implicit – communication of preferences can be used deliberately to determine the personality types of individuals, and also their voting inclinations (Gelman et al., 2016; Kosinski et al., 2013; Wu et al., 2015). Such information has obvious applications for those interested in the outcomes of elections, which have apparently been deployed with some success. Such techniques were reportedly originally designed and deployed to break up terrorist networks in conflict regions in the Middle East. The technique consists of identifying like-minded individuals with minority opinions that are available locally or otherwise convenient to those doing the aggregating, then introducing these individuals to each other and encouraging their political participation (Piette, 2018). AI-powered search can produce the 'coincidence' of good numbers of like-minded individuals in one place, convincing them all that their position is secretly in the majority – a secret being kept by the political status quo which must therefore be attacked. Relatively simple AI allows the identification and coordination of such target individuals; ICT allows the application of such power from a distance and across borders. Without laws and technological enforcement for transparency, such manipulation may be done invisibly.

2.4 AI as Prosthetic Memory

In the past, humans strove to both remember and to be remembered, yet the time and monetary cost of data retention were barriers to doing so. Now, however, these costs have reduced to the point where we no longer need to be picky about the quality of what we select to retain. This has its advantages and disadvantages. Biological brains forget (Kraemer and Golding, 1997). In humans, such a lack of pruning memories, inability to forget, or information overload can result in deficits in executive functioning, and an inability to escape past autobiographical memories (Parker et al., 2006). Using AI-driven techniques, even with surplus data, we can still successfully store, reorganise, classify and retrieve the relevant information we require from large data stores. However, this prosthetic memory of digital expression of opinions and beliefs – whether political, religious or otherwise – combined with increased connectivity has resulted in, amongst other things, a resurgence of call-out culture and public shaming (Hess and Waller, 2014; Tucker, 2018; Webb et al., 2016).

The urge to normalise groups by exiling or denouncing the credibility of individuals who diverge is a facet of tribalism prevalent throughout our history (Bechtel, 1991; Burns, 2003). Novel though with the AI landscape is the relative immortality of digital memory. The social costs (and benefits) to acting or thinking in diverging ways have increased; it is hard to be forgotten. Thus again, we are homogenised by our evolutionary urge

to remain in the safety of the tribe. Mayer-Schönberger (2007) advocates finding an equilibrium: by giving data an expiry date after which it will automatically be deleted, we can maximise the benefits of a precise prosthetic memory, whilst preserving the right to be forgotten. However, such arbitrary truncation would also be an end to history, unless history were still retained in non-digital format. Even if exceptions were made for those considered public individuals, as is now the case for certain privacy laws, this could have an unintended effect of reducing social mobility, as those in situations of prominence, privileged with being knowable, would be more likely to garner further attention and opportunity.

Historical data also has applications ranging from the social sciences to monitoring the impacts of government policies. In an era where behaviour modification might be practised by subterfuge, accurate historical data may be the only way to detect malicious actors working subtly over time. The rights to freedom of opinion and thought are enshrined in the Declaration on Human Rights (United Nations General Assembly, 1948), but without an option of perceivable expression such rights may be of limited value. There is substantial research being conducted in anonymisation, including for data extraction and analysis – whether this proves mathematically tractable remains to be seen. Uncompromised cybersecurity of not only data storage but transmission would also be necessary for any digital records to have even a hope of remaining private.

The claim that homogenisation is a side effect of AI may seem demonstrably false given the increase in identity politics and political polarisation. Coincidence is not necessarily causal, and even if there is a causal link, it may be difficult to untangle. Polarisation is known to be correlated with wealth inequality, and to have been so since before the advent of ICT and AI. Whether AI is presently contributing to inequality will be considered in the next section. But with respect to homogenisation, it is worth saying that both processes may well occur at the same time – rather than a plethora of perspectives, we may find strong forces towards conformity with one of a small number of tribes. Again, this is the opposite of what was anticipated with access to the Internet and cheap self-publication, and there is also evidence of societal fragmentation (Pentland, 2015).

3 AI AND INEQUALITY

Whether or not we become more homogenous in our beliefs and opinions, inequality in access to resources and quality of life is increasing. We discuss here what is known about how inequality is driven, and consider the impact of AI and the consequences for the collective.

First, it should be observed that globally inequality has been falling, an effect driven primarily by the very poorest. The World Bank reports more than a third of humanity moving out of extreme poverty since 1980 (Roser and Ortiz-Ospina, 2017), a shift that has been facilitated by ICT including AI, as populations have had more access to useful information such as weather predictions, fair prices and how to obtain government support. Further,

efforts to communicate political and economic situations, and means to coordinate protest, are leading at least some governments in rich areas such as the member states of the Organisation for Economic Co-operation and Development (OECD) to adopt policies that have been demonstrated to reduce inequality within populations by increasing wealth distribution.

Nevertheless, an individual's access to public goods such as schools, physical security, utilities and health depends to a large degree on their geographic identity. Access to ICT and AI is no exception (Robinson et al., 2015; Sujarwoto and Tampubolon, 2016), although it is also influenced by other factors such as age. As we globally become more dependent on such tools, the populations without access are exposed to higher risks of inequality. By moving towards equality of access to these technologies, individuals may have improved job opportunities and information in regard to the socioeconomic, political and cultural context in which they live.

The increase in inequality may result in reduced social cohesion as it seems to be correlated to reductions in social mobility and increases in political polarisation (McCarty et al., 2016). ICT and AI may also reduce the sort of localised social cohesion that is critical to many forms of well-being and political engagement, by diverting social attention to others with shared interests in topics that are not geographically centred. In this, it continues and expands on trends of mass media known since the beginning of the information age with, for example, the advent of national newspapers (Perlman and Sprick Schuster, 2016). For a substantial fraction of our society, the time we spend engaging with others as real, physical equals is replaced with ever-more-engaging digital entertainment. Research shows that whilst some US teens felt digital technologies connected them with others, others felt it resulted in a lack of in-person contact in their lives (Anderson and Jiang, 2018). The interaction that would once have taken place in person is now conducted through technology.

There is a danger that inequality produces an elite who no longer identify with the majority of individuals (Atkinson, 2015). This can be the effect not only of lack of social mobility and understanding, but even of simple spatial segregation (Cassiers and Kesteloot, 2012). As discussed earlier, digital social media provides a ready platform for disinformation, including caricaturised and exaggerated distortions of others. An elite may also falsely assume that it can consolidate power by extracting wealth from its own neighbours and nearest contenders. However, inequality breeds instability as the whims of small coalitions or even individual actors are unpredictable (Scheidel, 2017). When greater collective action is required, solutions become more predictable and stable, ironically better ensuring the maintenance of rank order at the higher end of society. Power over a collective is not only an animal thrill, it is also a mechanism of security because it ensures more individuals are invested in a mutually beneficial outcome (Terkel, 1974). Trust is a factor in individuals collectively investing in mutually beneficial outcomes, yet not only does inequality breed distrust (Barone and Mocetti, 2016) but, even when trust is held, it is vulnerable to exploitation and damage to the individual.

Again, there is no clear evidence to date, but rather active investigation, as to whether and in what ways AI may be affecting inequality. It seems evident that any technology that reduces the cost of distance will also facilitate inequality through no particular malfeasance but simply by allowing excellent businesses to dominate larger territories. In the case of some technologies now such as finance, media, pharmaceutical, aerospace, and of course digital, this is approaching the limit case of single corporations dominating markets globally. This same phenomenon may explain the similar surge in developed-world inequality witnessed in the late 19th and early 20th centuries (Atkinson, 2015; Bryson, 2019).

As mentioned in Section 2, another concern is that AI may alter employment. Whilst the common concern is that 'robots will take all the jobs', this seems highly unlikely as 'all the jobs' is not defined. There is an infinite number of ways we can better each other's lives. People tend to employ each other when the economy is good, though this may be seen as tautological. But the advent of intelligent technology has been associated with an increase in demand for both highly skilled and very low-skilled and low-waged work, with a declining demand and therefore wages for the intermediate workers (Acemoglu and Autor, 2011). Acemoglu and Autor (2011) suggest that technology plays two roles in wages: in allowing all to be more productive, this somewhat increases the value of being highly skilled, but flattens any advantages of moderate skill as people become more exchangeable. Worryingly, recent decades seem to be dominated by the latter effect (Acemoglu and Autor, 2011).

4 THE VULNERABILITY OF HUMAN TRUST

Subjective evaluations of trustworthiness are deeply tied to how humans navigate the world. The likelihood of a monetary transaction is largely dependent on the trust a buyer has in a seller (Kim et al., 2008; Ponte et al., 2015). Trust in politicians affects voter turnout (Grönlund and Setälä, 2007) and electoral results (Hetherington, 1999). When evaluating a person's face on multiple dimensions, subjective trustworthiness is one of the most predictive measures for overall evaluation (Oosterhof and Todorov, 2008). Further, once trustworthiness is perceived, it is relatively robust, persistently affecting behaviour (Delgado et al., 2005). If AI alters our capacity or predisposition to trust other humans, it will clearly have a deep impact on society.

4.1 Baseline Proclivities to Blindly Trust

Using simple economic games, it has been shown that individuals will blindly trust a stranger even when it leaves them vulnerable (Berg et al., 1995). In a one-shot Trust Game (TG), an investor is given some monetary windfall and must decide whether to keep the windfall, or give some fraction to a trustee. The fraction offered by the investor is then multiplied by some

factor (typically three) and the trustee can then offer a fraction of the multiplied investment back to the investor. Human investors tend to blindly trust their partner by offering non-zero investments, even though it is in the trustee's best interest to return nothing to the investor. Whilst trust is moderated by several factors, including framing effects (Burnham et al., 2000), geography (Johnson and Mislin, 2011), gender (Buchan et al., 2008), risk preferences (Fehr, 2009) and whether participants are selected from a student population (Johnson and Mislin, 2011), individuals consistently tend to blindly trust, investing some of their windfall (Johnson and Mislin, 2011).

Significant research has sought to understand how blindly trusting others might be ecologically rational. This proclivity to trust has been explained as adaptive in relatively small populations where individuals have reputation cues of their partners (Boero et al., 2009; Masuda and Nakamura, 2012) or if the chance of knowing a partner's strategy exceeds some threshold (Manapat and Rand, 2012; Manapat et al., 2013; McNamara et al., 2009; Rauwolf and Bryson, 2018). The common theme is that blindly trusting another can be adaptive if someone, somewhere, has a chance of having information about a player, including indirectly (e.g. if a population is known to share trust-related characteristics by some sort of contagion effect or enforcement). Interestingly, on the other hand, individuals will not trust others when they know the others are likely to return (reciprocate) an unfair amount in the TG, even if that amount would still make trusting them beneficial.

4.2 When Blind Trust Is Valuable

We have previously contributed work demonstrating a generally advantageous but unstable evolutionary dynamic that typically establishes trust. This demonstration was in the context of simulations with computational agents playing one-shot TGs with each other. Agents were given several potential partners for playing each game, and chose which agent to play, if any. As in the natural world, the investing agent knew the reputation for historic pay-off of some partners, but did not know the pay-off of others – information was partially occluded. Each actor learned three things socially from the strongest players: the levels of trustworthiness in unknown players, the demanded level of reciprocation for known players, and their own reciprocation rate. Rauwolf and Bryson (2018) demonstrate that this simple dynamic is sufficient to generate trust. When the known pay-offs of partners are sufficiently low, it can be in an agent's interest to blindly select a partner whose history is unknown, provided that the population has evolved high enough levels of trust, which tends to coevolve with high levels of reciprocation, and of course assuming sufficient extra benefit from mutual development of the public good. In these cases, blind trust can be more valuable to the agent than walking away with their monetary windfall, trusting no one. On the other hand, trust will not evolve if

the reciprocation rate of many players is known, because it becomes better to stick with the best-known available return rather than to take a chance with the unknown.

The insight from this work is that a willingness to blindly trust others increases competition between others, lowering prices by increasing the reciprocation rate. It is well-known that creating competition between sellers lowers prices. But, by being willing to trust those whose information is unknown, the pressure of competition is increased. Not only do sellers need to compete with other sellers whose information is known, they now need to compete with those whose information is unknown. This tends to lower the market price even further. This is related to work on outside options (André and Baumard, 2011). The value of a sale is contingent upon the other options of a prospective buyer. If an individual is willing to go elsewhere, even by blindly trusting a stranger, then the market is forced to adjust and the buyer's life is improved.

4.3 Information Cost Reduces Benefits of Trust

Importantly, whilst the adaptive models of trust require that some information is available, trust fails if information is fully transparent (Manapat et al., 2013; Rauwolf and Bryson, 2018). By definition, the act of trusting another requires some uncertainty in the outcome (Yamagishi, 2011). If information is fully transparent, then there is no need to trust another; rather, each individual can make an informed decision. The consequence is that there is no selective pressure to evolve or learn the strategies and beliefs associated with the riskier behaviour, but these are what bind a local community together.

We are currently living in an age where the cost of information is dramatically decreasing. As a result, the adaptive benefits of trust are becoming increasingly obsolete. This may be for the best – we may have a more predictable environment with even higher rewards for 'good' behaviour. However, we should also be concerned about this being another force for homogeneity, and further loss of individual capacity to deal creatively with localised crises. The institution of trust may be just a consequence of our inability to perfectly control our peers, but it may also serve an adaptive advantage by reducing our responsibility to do so. Society does not need to come up with plans to handle every contingency, because a desperate individual can always take advantage of the availability of trust without first seeking social approval of their plans.

There are indications that in the near term moving from trust to full information is problematic in other ways. Because we will not just 'shut off' our historic psychological choice making, replacing trust with information may mean that extant prejudices become more rigidly a part of our behaviour. Not only is trust deeply tied to how individuals make decisions, subjective trust is often biased and founded on unhelpful signals. People find

attractive individuals more trustworthy (Wilson and Eckel, 2006). Individuals will invest more if a profile picture has a smiling face (Scharlemann et al., 2001) or is visually perceived as more trustworthy (Bente et al., 2012). The trust individuals place in profile information is often incorrect (Toma, 2010). More generally, individuals perform close to chance when predicting deception (Bond and DePaulo, 2006). Ert et al. (2016) show that perceived trustworthiness of an Airbnb option correlates more with the profile photo than the quantified reputation score of that option. This was confirmed by Fagerstrøm et al. (2017), who found that facial expressions in a renter's photo predicted likelihood to rent more than customer ratings. This demonstrates that the transparency of information is not necessarily sufficient to improve behaviour. Individuals must make decisions using that information before it offers an advantage.

4.4 Calibrating Trust in AI

There is considerable discussion these days about trust in AI and trustworthiness for AI. Our own work and that of many in the UK's ethics community more generally has taken a different tack, emphasising that trust is an anthropocentric trait not truly useful for artefacts, where transparency and accountability are more desirable (Boden et al., 2011; Bryson and Theodorou, 2019). Improving transparency in AI reduces the need to trust AI. Yet it is possible that transparency does not affect the consumption of AI when the human consumer projects human-like identity to intelligent technology (anthropomorphises). By doing so, the consumer exposes themselves to exploitation as their established biases concerning the likelihood of trustworthiness are even easier to exploit via designed artefacts than they are by unscrupulous individuals. For example, one might assume that a robot is unlikely to remember everything you say because a person or pet would not, but the robot may in fact not only recall but transmit its full memory – a full record of all interactions or even nearby events – storing these offsite in a corporate cloud. Although in theory the same digital and architected features of AI that make it more powerful as a manipulator should also make it easier to govern, presently (2019) manipulation is outstripping governance.

Acceptance of AI can be increased in many ways, but given the vulnerability of the human trust system, care is needed to ensure trust is extended with consent, and is not exploited (IEEE, 2019). Whilst tapping into the vulnerabilities of how humanity perceives trustworthiness may be efficacious, it can also result in unwarranted trust. People are already found to perceive AI as more objective than human decision-makers, and in some cases to over-rely on AI. For example, in a legal setting, people demonstrated a preference to follow a machine advisor's decision despite a human advisor having judgement of higher accuracy (Logg et al., 2019). In our own research (Wilson and Theodorou, 2019), we have found that in virtual reality (VR), AI actors presented to the participants as human agents were perceived to

be significantly less deterministic than those presented explicitly as robots, despite both being controlled by the same AI system.

It seems that many of us attribute properties to AI that do not exist, at least where that AI reminds us of humans (Sparrow, 2019). Some evidence suggests that anthropomorphising robots increases interaction with them (Waytz et al., 2014) via increased trust resistance (de Visser et al., 2016) and mind attribution. There are uncertainties as to the impact of anthropomorphism when robots are more prevalent and 'normalised' in our society. As we grow familiar with robots in our day-to-day lives, our mental models of robots may become more accurate (Bryson and Kime, 2011). We may perceive with more clarity the distinctions between artificially embodied cognition and humans, or commercial products may be mandated to provide transparency. In these cases, the impact of anthropomorphism may be reduced. Alternatively, viewing robots as human-like may become normalised, and social robotics – believed to be human-like, despite their inhuman, designed capacities – an embedded aspect of our lives.

We suggest that a safer and more long-term stable approach is to work to increase AI transparency whilst simultaneously helping individuals learn to make choices using empirical information. As the information age reduces the need for trust, individuals need to be trained to operate in this information age, rather than reinforcing poor decision-making tendencies based on fallible and manipulable perceptions of trust. An example of increased AI transparency would be to have a QR code attached to each robot that when scanned gives information on the robot's maker, purpose and capabilities. We have also been developing means for allowing users to see the current goals and strategies of a robot or other real-time interactive AI system (Rotsidis et al., 2019; Theodorou et al., 2017), and are presently experimenting to see whether this reduces the moral-hazard aspects of anthropomorphism. Whilst some may see viewing AI as human-like to be an example of a freedom of opinion or even association, we feel strongly that such opinions need to be informed where information is available, in order to avoid unknowing exploitation. Willing exploitation, like the manipulation of emotions that occurs during a motion picture or other work of fiction, is of course a perfectly acceptable part of life and entertainment. We only seek to avoid economic and political manipulation imposed on unknowing others.

5 BARRIERS TO ACCURATE PERCEPTION AND DEVELOPMENTS

Our earlier metaphor for AI as an extension of our nervous system posited that new AI tendrils enhance our ability to sense and perceive. Yet, problematically, the collective intelligence that *could* be garnered from this additional perception is hindered by two key barriers: the fact that data can be inaccurate and misleading, and our own inability to handle and interpret data. In this section we explore the origins, trajectory and impact of inaccurate

information in human communication networks. We note our current inability to correctly deploy, infer and apply meaning from AI. We highlight current avenues for reducing these barriers in order to fully exploit our augmented collective intelligence.

5.1 'Fake News'

Humans have long expressed a desire to record and share information: the first encyclopaedia was written in AD 77 (Gudger, 1924); libraries have been dated back 2,000 years earlier. The arrival of the telegraph and Morse Code in 1835 enabled instantaneous transmission of knowledge across great distances (Burns, 1988); now databases are ubiquitous. Historical records indicate sharing knowledge spurred many innovations (Bessen and Nuvolari, 2016), and on a day-to-day basis, enables individuals to make informed decisions and actions. Unfortunately, not all shared information is accurate. Disinformation and misinformation, which often fall under the misnomer 'fake news', are of rising public concern. Disinformation implies intentional creation and sharing of manipulated or false information, whereas misinformation refers to inadvertent sharing (Lazer et al., 2018).

Fake news is not new: since humans could speak, misinformation has spread via word of mouth. The spread increased and quickened with the arrival of newspapers and pamphlets, then with mass media such as television, finally exploding with the Internet and especially social media (Burkhardt, 2017). What is new is that, compared to past technological mediums, social media largely lacks filtering, editorial judgement and fact-checking (Allcott and Gentzkow, 2017). Further, the communication network is infiltrated by artificially intelligent bots, able to pass as or augment human users, which can be used to quickly deploy, share and spread information across networks (Machado and Konopacki, 2018). Such bots can be used to sway public opinion. In fact, disinformation affects stock prices (Carvalho et al., 2011), political opinions (Howard et al., 2018), and voting patterns (Allcott and Gentzkow, 2017) at least transiently. Evidence shows that accurate stories take longer to spread but have more purchase once spread (Vosoughi et al., 2018). In previous work (Mitchell et al., 2016), we have demonstrated that even error-prone 'gossip' can be a better strategy than direct experiential learning for acquiring true and useful information. The speed of information transmission such as is provided by social media can in some circumstances outweigh the costs of incorrect information, particularly if disinformation can be identified and quickly combated (Panagiotopoulos et al., 2014).

Nevertheless, in at least some contexts, our species seems often to communicate inaccurate rather than accurate information. An analysis of a data set of rumour cascades on Twitter revealed fake news was 70% more likely to be retweeted than the truth (Vosoughi et al., 2018). There are suggestions that fake news is more likely to be novel, and novelty captures human attention. Perhaps worryingly, the average American spends 23.6 hours

online weekly (Cole et al., 2017), and 62% get their news online (Gottfried and Shearer, 2016). Whilst there is evidence that our trust of such news has decreased, exposure alone may have negative impacts. Exposure can prime thinking and conversational topics. When any news enters the conversational sphere, trust in the information increases (Hajli et al., 2014). We humans seem to have a disposition to trust information communicated by word of mouth (Atika et al., 2018; Huete-Alcocer, 2017). Conversations require resources such as time, cognition, and sometimes emotional investment. Actions occurring as consequence of conversations result in further deployment of resources. We know that, as a social species, we respond to such evidence of investment by others in our society (Zahavi, 1977). We posit that individual and collective resources may be consequently directed away from more accurate topics which are of perhaps more importance to our survival and flourishing.

5.2 The Impact of Information Accuracy

In further work with simulations and formal analysis, we offered insight into the environments where humans may be vulnerable to utilising incorrect information (Rauwolf, 2016). Information requires both time and energy to gather. If information gathering comes at a non-trivial cost, then we would expect individuals to truncate their information search after a period of time (Simon, 1956). However, given continual improvements in technology, the cost of information is falling; as a result we might expect individuals to be better informed. Importantly though, even if information is easily obtained, if the processing of that information is costly, limiting information can be advantageous (Rauwolf and Jones, 2019).

Rauwolf et al. (2015) show that when the benefits of group dynamics conflict with the accuracy of beliefs, false beliefs can become the least-costly option. Across a variety of contexts, individuals tend to prioritise relationships with those who share similar values – a trait called value-homophily (McPherson et al., 2001). We have demonstrated that it is in precisely these contexts where individuals can be expected to use incorrect information (Rauwolf et al., 2015). When the social value or benefit provided by the group outweighs the private cost of possessing incorrect information, it is advantageous for the individual to maintain (or at least act on) their false beliefs. Given that the inaccuracy of political and religious beliefs provides virtually no personal cost (Caplan, 2001), but group agreement may provide social value through security, we would expect humanity would struggle to remove false beliefs, particularly in times of resource scarcity and conflict (Stewart et al., 2018).

As the benefits conferred by AI and technology at large continue to improve and secure individual basic needs, individuals will likely pay a reduced private cost for possessing incorrect information across a broadening array of contexts. Regardless of the accuracy of an

individual's beliefs, the basic needs of most individuals are improving. As such, if individuals pay small personal costs for false beliefs, but garner large social benefits for group homogeneity, then we would expect an expanding and resilient battle against false beliefs.

Nevertheless, what costs an individual little in isolation may cost a society a great deal due to aggregate responses, particularly in a democracy (Chote et al., 2016; Lewis, 2017). Whether or not we should strive for increased rate of information transmission in every case (see the discussions of trust and freedom of opinion above), we should almost certainly prefer *accurate* communication, though here, too, inaccuracy can sometimes lead to useful innovation. Disinformation is a global and long-running issue, and there are global initiatives to combat it. For example, Facebook flags potential news stories to be reviewed by third-party fact checkers; and through the messaging system WeChat in China, users can report fake news, which is then checked and flagged. Crucial new initiatives introduce critical literacy into the education curriculum – training children to recognise and question information sources, particularly online (Vasu et al., 2018). Critical-thinking and fact-checking skills, as well as basic understanding of algorithm mechanics and their limitations, could enable the next generation to be better prepared to avoid such scenarios faced by us today (Guess et al., 2019). Fact-checking can be as simple as conducting a Web search on a topic and its source.

6 CONCLUSION

In this chapter we have discussed the impact of AI on contemporary societies. We took a perspective of understanding how the changing social and economic landscape induced by AI interacts with the human information-processing biases which evolved in very different environments. We consider a better understanding of these impacts imperative for our society going forward as we optimise our existence with AI, and for ensuring the AI and the regulations we design to govern its use both maximise benefit and minimise harm. We discussed the impact on both collective and individual human behaviour. Here we summarise the key foci of this chapter. First, the accuracy of narratives surrounding AI could critically impact optimal engagement with AI. Next, we compared AI to a prosthetic nervous system, which increases our perception and agency. AI also enhances our capacity to remember, coordinate, connect and communicate; this has many positive but also some negative outcomes. We considered the impact on freedom and diversity of opinion, political and economic impact, the mechanisms of information spread, and vulnerability of human trust and social coherence. The increased discoverability and predictability facilitated by AI requires serious consideration; there are myriad beneficial and harmful current applications of AI, and no doubt more of both to come. There are also of course many movements to ensure AI is beneficial to our society rather than harmful, though we did not take time to touch on those much here, we view such consideration and efforts as essential. Coordinating and enforcing such pro-social efforts has traditionally been called governance, and we hope this

chapter may contribute to making sensible governance easier to both justify and employ constructively.

REFERENCES

Acemoglu, D. and Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In Ashenfelter, O and Card, D (Eds). volume 4 of *Handbook of Labor Economics*, chapter 12, pages 1043–1171. Amsterdam: Elsevier.

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of Economic Perspectives, 31(2):211–236.

Anderson, J., Rainie, L., and Luchsinger, A. (2018). Artificial intelligence and the future of humans. Pew Research Center.

Anderson, M. and Jiang, J. (2018). Teens, social media & technology 2018. Washington, DC: Pew Internet & American Life Project. Retrieved June 3, 2018, from https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/

André, J.-B. and Baumard, N. (2011). Social opportunities and the evolution of fairness. Journal of Theoretical Biology, 289:128–135.

Atika, A., Kusumawati, A., and Iqbal, M. (2018). The effect of electronic word of mouth, message source credibility, information quality on brand image and purchase intention. EKUITAS (Jurnal Ekonomi dan Keuangan), 20(1):94–108.

Atkinson, A. B. (2015). Inequality: What can be done? Cambridge, MA: Harvard University Press.

Barone, G. and Mocetti, S. (2016). Inequality and trust: New evidence from panel data. Economic Inquiry, 54(2):794–809.

Bechtel, L. M. (1991). Shame as a sanction of social control in biblical Israel: Judicial, political, and social shaming. Journal for the Study of the Old Testament, 16(49):47–76.

Behavioural Insights Team (2010). Applying behavioural insight to health. Technical Report 403936/1210, Cabinet Office, UK Government, London.

Behavioural Insights Team (2012). Applying behavioural insights to reduce fraud, error and debt. Technical Report 408779/0212, Cabinet Office, UK Government, London.

Bente, G., Baptist, O., and Leuschner, H. (2012). To buy or not to buy: Influence of seller photos and reputation on buyer trust and purchase behavior. International Journal of Human-Computer Studies, 70(1):1–13.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. Games and Economic Behavior, 10(1):122–142.

Berger, J. (2011). Arousal increases social transmission of information. Psychological Science, 22(7):891–893.

Bessen, J. and Nuvolari, A. (2016). Knowledge sharing among inventors: Some historical perspectives, In D. Harhoff, K. Lakhani (Eds.), *Revolutionizing Innovation: Users, Communities, and Open Innovation*, page 135. MIT Press, Cambridge MA.

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., Newman, P., Parry, V., Pegman, G., Rodden, T., Sorell, T., Wallis, M., Whitby, B., and Winfield, A. (2011). Principles of robotics. The United Kingdom's Engineering and Physical Sciences Research Council (EPSRC). Retrieved from https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/

Boero, R., Bravo, G., Castellani, M., and Squazzoni, F. (2009). Reputational cues in repeated trust games. The Journal of Socio-Economics, 38(6):871–877.

Bond, C. F. J. and DePaulo, B. M. (2006). Accuracy of deception judgments. Personality and Social Psychology Review, 10(3):214–234.

Bradford, E. E., Jentzsch, I., and Gomez, J.-C. (2015). From self to social cognition: Theory of mind mechanisms and their relation to executive functioning. Cognition, 138:21–34.

Brown, G. R., Dickins, T. E., Sear, R., and Laland, K. N. (2011). Evolutionary accounts of human behavioural diversity. Philosophical Transactions of the Royal Society B: Biological Sciences, 366, 313–324

Bryson, J. J. (2015). Artificial intelligence and pro-social behaviour. In Misselhorn, C., editor, *Collective Agency and Cooperation in Natural and Artificial Systems: Explanation, Implementation and Simulation*, volume 122 of Philosophical Studies, pages 281–306. Berlin: Springer.

Bryson, J. J. (2019). The past decade and future of AI's impact on society. In *Towards a New Enlightenment?* In Cities, D (Eds.), *A Transcendent Decade*, pages 150–185. Madrid: Turner–BBVA.

Bryson, J. J. and Kime, P. P. (2011). Just an artifact: Why machines are perceived as moral agents. In Proceedings of the 22nd International Joint Conference on Artificial Intelligence, pages 1641–1646, Barcelona. Morgan Kaufmann.

Bryson, J. J. and Theodorou, A. (2019). How society can maintain human-centric artificial intelligence. In Toivonen-Noro, M. and Saari, E (Eds.), *Human-Centered Digitalization and Services,* chapter 16, pages 305-323. Singapore: Springer.

Buchan, N. R., Croson, R. T., and Solnick, S. (2008). Trust and gender: An examination of behavior and beliefs in the investment game. Journal of Economic Behavior & Organization, 68(3):466–476.

Burkhardt, J. M. (2017). History of fake news. Library Technology Reports, 53(8):5–9.

Burnham, T., McCabe, K., and Smith, V. L. (2000). Friend-or-foe intentionality priming in an extensive form trust game. Journal of Economic Behavior & Organization, 43(1):57–73.

Burns, R. (1988). The electric telegraph and the development of picture telegraphy. In Papers Presented at the Sixteenth IEE Week-End Meeting on the History of Electrical Engineering, pages 80–86. IET.

Burns, W. E. (2003). Witch hunts in Europe and America: An encyclopedia. Westport, CT: Greenwood Publishing Group.

Buss, D. M. and Schmitt, D. P. (1993). Sexual strategies theory: An evolutionary perspective on human mating. Psychological Review, 100(2):204.

Caplan, B. (2001). Rational ignorance versus rational irrationality. Kyklos, 54(1):3–26.

Carvalho, C., Klagge, N., and Moench, E. (2011). The persistent effects of a false news shock. Journal of Empirical Finance, 18(4):597–615.

Cassiers, T. and Kesteloot, C. (2012). Socio-spatial inequalities and social cohesion in European cities. Urban Studies, 49(9):1909–1924.

Cave, S., Coughlan, K., and Dihal, K. (2019). 'Scary robots': Examining public responses to AI. In Proceedings of the AIES. Retrieved May 24, 2020, from https://doi.org/10.1145/3306618.3314232

Chavan, S. and Khot, T. S. (2013). Efficient and reliable routing algorithm to enhance connectivity in disaster scenario: Abc algorithm. International Journal of Science and Research, 4(5):891-896

Chote, R., Nickell, S., and Parker, G. (2016). Economic and fiscal outlook. Technical Report Cm 9346, Office for Budget Responsibility, London, UK. Available from www.gov.uk/government/publications.

Cole, J., Suman, M., Schramm, P., and Zhou, L. (2017). The 2017 digital future report: Surveying the digital future. Los Angeles, CA: USC Annenberg School Center for the Digital Future.

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., and Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. Journal of Experimental Psychology: Applied, 22(3):331.

Delgado, M. R., Frank, R. H., and Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. Nature Neuroscience, 8(11):1611–1618.

Eagle, N. and Pentland, A. S. (2003). Social network computing. In Dey, A. K., Schmidt, A., and McCarthy, J. F., editors, UbiComp 2003: Ubiquitous Computing, pages 289–296, Berlin, Heidelberg: Springer Berlin Heidelberg.

Elish, M. C. and Boyd, D. (2018). Situating methods in the magic of big data and AI. Communication Monographs, 85(1):57–80.

Ert, E., Fleischer, A., and Magen, N. (2016). Trust and reputation in the sharing economy: The role of personal photos in Airbnb. Tourism Management, 55:62–73.

Fagerstrøm, A., Pawar, S., Sigurdsson, V., Foxall, G. R., and de Soriano, M. Y. (2017). That personal profile image might jeopardize your rental opportunity! On the relative impact of the seller's facial expressions upon buying behavior on AirbnbTM. Computers in Human Behavior, 72:123–131.

Fehr, E. (2009). On the economics and biology of trust. Journal of the European Economic Association, 7(2–3):235–266.

Fisher RA (1930) The genetical theory of natural selection. Clarendon Press, Oxford (2nd edn., 1958, New York, Dover. Variorum edition, Bennett JH (ed), 1999, Oxford University Press) Franco, R. Z., Fallaize, R., Lovegrove, J. A., and Hwang, F. (2016). Popular nutrition-related mobile apps: A feature assessment. JMIR mHealth and uHealth, 4(3):e85.

Gelman, A., Goel, S., Rivers, D., and Rothschild, D. (2016). The mythical swing voter. Quarterly Journal of Political Science, 11(1):103–130.

Gent, E. (2015). Ai: Fears of 'playing god' [control & automation artificial intelligence]. Engineering & Technology, 10(2):76–79.

Gerbaudo, P. (2018). Tweets and the streets: Social media and contemporary activism. London: Pluto Press.

Gilligan, I. (2007). Neanderthal extinction and modern human behaviour: The role of climate change and clothing. World Archaeology, 39(4):499–514.

Godfrey-Smith, P. (2017). Other minds: The octopus, the sea, and the deep origins of consciousness. New York, NY: Farrar, Straus and Giroux.

Gottfried, J. and Shearer, E. (2016). News Use Across Social Medial Platforms 2016. Washington, DC: Pew Research Center. Retrieved 24 May, 2020 from http:// www.journalism.org/2016/05/26/news-use-across-social-me dia-platforms-2016/.

Grönlund, K. and Setälä, M. (2007). Political trust, satisfaction and voter turnout. Comparative European Politics, 5(4):400–422.

Gudger, E. W. (1924). Pliny's historia naturalis. The most popular natural history ever published. Isis, 6(3):269–281.

Guess, A., Nagler, J., and Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. Science Advances, 5(1), eaau4586.Hajli, N., Lin, X., Featherman, M., and Wang, Y. (2014). Social word of mouth: How trust develops in the market. International Journal of Market Research, 56(5):673–689.

Hammack, P. L. (2008). Narrative and the cultural psychology of identity. Personality and Social Psychology Review, 12(3):222–247.

Hao, T., Xing, G., and Zhou, G. (2013). isleep: Unobtrusive sleep quality monitoring using smartphones. In Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems, page 4. ACM.

Hess, K. and Waller, L. (2014). The digital pillory: Media shaming of 'ordinary' people for minor crimes. Continuum, 28(1):101–111.

Hetherington, M. J. (1999). The effect of political trust on the presidential vote, 1968–96. American Political Science Review, 93(2):311–326.

Howard, P. N., Kollanyi, B., Bradshaw, S., and Neudert, L.-M. (2018). Social media, news and political information during the US election: Was polarizing content concentrated in swing states? arXiv preprint arXiv:1802.03573.

Huete-Alcocer, N. (2017). A literature review of word of mouth and electronic word of mouth: Implications for consumer behavior. Frontiers in Psychology, 8:1256.

Institute of Electrical and Electronics Engineers (IEEE) (2019). Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems. Technical report, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, first edition.

Johnson, N. D. and Mislin, A. A. (2011). Trust games: A meta-analysis. Journal of Economic Psychology, 32(5):865–889.

Kim, D. J., Ferrin, D. L., and Rao, H. R. (2008). A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. Decision Support Systems, 44(2):544–564.

Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 110(15), 5802-5805..

Kraemer, P. J. and Golding, J. M. (1997). Adaptive forgetting in animals. Psychonomic Bulletin & Review, 4(4):480–491.

Kramer, K. L. and Ellison, P. T. (2010). Pooled energy budgets: Resituating human energy-allocation trade-offs. Evolutionary Anthropology: Issues, News, and Reviews, 19(4):136–147.

Lahr, M. M. (2016). The shaping of human diversity: Filters, boundaries and transitions. Philosophical Transactions of the Royal Society B, 371(1698):20150241.

Landon-Murray, M., Mujkic, E., and Nussbaum, B. (2019). Disinformation in contemporary US foreign policy: Impacts and ethics in an era of fake news, social media, and artificial intelligence. Public Integrity, 21(5), 512-522.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Schudson, M. (2018). The science of fake news. Science, 359(6380):1094–1096.

Lee, E., Bristow, J., Faircloth, C., and Macvarish, J. (2014). Parenting culture studies. London: Palgrave Macmillan.

Lewis, M. (2017). Why the scariest nuclear threat may be coming from inside the White House. Vanity Fair. Retrieved 24 May, 2020, from https://www.vanityfair.com/news/2017/07/department-of-energy-risks-michael-lewis.

Liang, Y. and Lee, S. A. (2017). Fear of autonomous robots and artificial intelligence: Evidence from national representative data with probability sampling. International Journal of Social Robotics, 9(3):379–384.

Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151:90–103.

Machado, C. and Konopacki, M. (2018). Computational power: Automated use of WhatsApp in the Brazilian elections. Medium. Retrieved 24 May, 2020 from https://feed.itsrio.org/computational-power-automated-use-of-whatsapp-in-the-elections-59f62b857033.

Manapat, M. L., Nowak, M. A., and Rand, D. G. (2013). Information, irrationality, and the evolution of trust. Journal of Economic Behavior & Organization, 90:S57–S75. Evolution as a General Theoretical Framework for Economics and Public Policy.

Manapat, M. L. and Rand, D. R. (2012). Delayed and inconsistent information and the evolution of trust. Dynamic Games and Applications, 2:401–410.

Masuda, N. and Nakamura, M. (2012). Coevolution of trustful buyers and cooperative sellers in the trust game. PLoS One, 7(9):1–11.

Mayer-Schönberger, V. (2007). Useful void: The art of forgetting in the age of ubiquitous computing, (Working Paper No. RWP07-022). Cambridge, MA: Harvard University.

McCarty, N. M., Poole, K. T., and Rosenthal, H. (2016). Polarized America: The dance of ideology and unequal riches. Cambridge, MA: MIT Press, second edition.

McDonald, M. M., Navarrete, C. D., and Van Vugt, M. (2012). Evolution and the psychology of intergroup conflict: The male warrior hypothesis. Philosophical Transactions of the Royal Society B, 367(1589):670–679.

McNamara, J. M., Stephens, P. A., Dall, S. R., and Houston, A. I. (2009). Evolution of trust and trustworthiness: Social awareness favours personality differences. Proceedings of the Royal Society B: Biological Sciences, 276(1657):605–613.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. Annual Review of Sociology, 27(1):415–444.

Meijer, A. and Wessels, M. (2019). Predictive policing: Review of benefits and drawbacks. International Journal of Public Administration, *42*(12), 1031-1039. .

Mitchell, D., Bryson, J. J., Rauwolf, P., and Ingram, G. P. (2016). On the reliability of unreliable information: Gossip as cultural. Interaction Studies, 17(1), 1-18

Morris, N. (2002). The myth of unadulterated culture meets the threat of imported media. Media, Culture & Society, 24(2):278–289.

Moscovici, S. (1981). On social representations. In Forgas, J (Ed.), Social cognition: Perspectives on everyday understanding,8(12):181–209, London: Academic Press.

Moscovici, S. (2001). Why a theory of social representation? The representations of the Social. Bridging theoretical traditions. 8-37

Müller, V. C. and Bostrom, N. (2016). Future progress in artificial intelligence: A survey of expert opinion. In Müller, V. C. (Ed.), *Fundamental Issues of Artificial Intelligence*, pages 555–572. Berlin: Springer.

Nairne, J. S., Pandeirada, J. N., Gregory, K. J., and Van Arsdall, J. E. (2009). Adaptive memory: Fitness relevance and the hunter-gatherer mind. Psychological Science, 20(6):740–746.

Oosterhof, N. N. and Todorov, A. (2008). The functional basis of face evaluation. Proceedings of the National Academy of Sciences, 105(32):11087–11092.

Panagiotopoulos, P., Bigdeli, A. Z., and Sams, S. (2014). Citizen–government collaboration on social media: The case of Twitter in the 2011 riots in England. Government Information Quarterly, 31(3):349–357.

Parker, E. S., Cahill, L., and McGaugh, J. L. (2006). A case of unusual autobiographical remembering. Neurocase, 12(1):35–49.

Pentland, A. (2015). Social physics: How social networks can make us smarter. New York, NY: Penguin.

Perlman, E. R. and Sprick Schuster, S. (2016). Delivering the vote: The political effect of free mail delivery in early twentieth century America. The Journal of Economic History, 76(3):769–802.

Piette, A. (2018). Muriel Spark and fake news. Textual Practice, 32(9):1577–1591.

Ponte, E. B., Carvajal-Trujillo, E., and Escobar-Rodríguez, T. (2015). Influence of trust and perceived value on the intention to purchase travel online: Integrating the effects of assurance on trust antecedents. Tourism Management, 47:286–302.

Raber, I., McCarthy, C. P., and Yeh, R. W. (2019). Health insurance and mobile health devices: Opportunities and concerns. Journal of the American Medical Association (JAMA), 321(18):1767–1768.

Rauwolf, P. (2016). Understanding the ubiquity of self-deception: The evolutionary utility of incorrect information. PhD thesis, University of Bath.

Rauwolf, P. and Bryson, J. J. (2018). Expectations of fairness and trust co-evolve in environments of partial information. Dynamic Games and Applications, 8(4):891–917.

Rauwolf, P. and Jones, A. (2019). Exploring the utility of internal whistleblowing in healthcare via agent-based models. BMJ Open, 9(1).

Rauwolf, P., Mitchell, D., and Bryson, J. J. (2015). Value homophily benefits cooperation but motivates employing incorrect social information. Journal of Theoretical Biology, 367:246–261.

Richerson, P. J., Bettinger, R. L., and Boyd, R. (2005). Evolution on a restless planet: Were environmental variability and environmental change major drivers of human evolution. In In Woketits, F. M. & Ayala, F. J. (Eds.), Handbook of Evolution, 2:223–242.

Robinson, L., Cotten, S. R., Ono, H., Quan-Haase, A., Mesch, G., Chen, W., Schulz, J., Hale, T. M., and Stern, M. J. (2015). Digital inequalities and why they matter. Information, Communication & Society, 18(5):569–582.

Roser, M. and Ortiz-Ospina, E. (2017). Global extreme poverty. Technical report, Our World in Data.

Rotsidis, A., Theodorou, A., Bryson, J. J., and Wortham, R. H. (2019). Improving robot transparency: An investigation with mobile augmented reality. In 28th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), New Delhi. IEEE.

Roughgarden, J., Oishi, M., and Akçay, E. (2006). Reproductive social behavior: Cooperative games to replace sexual selection. Science, 311(5763):965–969.

Scharlemann, J. P., Eckel, C. C., Kacelnik, A., and Wilson, R. K. (2001). The value of a smile: Game theory with a human face. Journal of Economic Psychology, 22(5):617–640.

Scheidel, W. (2017). The great leveler: Violence and the history of inequality from the Stone Age to the twenty-first century. Princeton, NJ: Princeton University Press.

Shi, F., Teplitskiy, M., Duede, E., and Evans, J. A. (2019). The wisdom of polarized crowds. Nature Human Behaviour, 3:329–336.

Siles, I. and Walker, I. D. (2009). Design, construction, and testing of a new class of mobile robots for cave exploration. In Mechatronics, 2009. ICM 2009. IEEE International Conference on, pages 1–6. IEEE.

Simkulet, W. (2019). Informed consent and nudging. Bioethics, 33(1):169–184.

Simon, H. A. (1956). Rational choice and the structure of the environment. Psychological Review, 63(2):129–138.

Smaldino, P. E., Newson, L., Schank, J. C., and Richerson, P. J. (2013). Simulating the evolution of the human family: Cooperative breeding increases in harsh environments. PLoS One, 8(11):e80753.

Smith, E. A. (2010). Communication and collective action: Language and the evolution of human cooperation. Evolution and Human Behavior, 31(4):231–245.

Sparrow, R. (2019). Robotics has a race problem. Science, Technology, & Human Values, 45(3), 538-560.

Spring, B., Gotsis, M., Paiva, A., and Spruijt-Metz, D. (2013). Healthy apps: Mobile devices for continuous monitoring and intervention. IEEE Pulse, 4(6):34.

Starbird, K. and Palen, L. (2011). 'Voluntweeters': Self-organizing by digital volunteers in times of crisis. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '11, pages 1071–1080, New York, NY. ACM.

Stewart, A. J., McCarty, N., and Bryson, J. J. (2018). Explaining parochialism: A causal account for political polarization in changing economic environments. arXiv preprint arXiv:1807.11477.

Stone, J. and Moskowitz, G. B. (2011). Non-conscious bias in medical decision making: What can be done to reduce it? Medical Education, 45(8):768–776.

Sujarwoto, S. and Tampubolon, G. (2016). Spatial inequality and the internet divide in Indonesia 2010–2012. Telecommunications Policy, 40(7):602–616.

Terkel, S. (1974). Working. New York: The New Press.

Terry, D. J. and Hogg, M. A. (1996). Group norms and the attitude-behavior relationship: A role for group identification. Personality and Social Psychology Bulletin, 22(8):776–793.

Teste, F. P., Simard, S. W., Durall, D. M., Guy, R. D., Jones, M. D., and Schoonmaker, A. L. (2009). Access to mycorrhizal networks and roots of trees: Importance for seedling survival and resource transfer. Ecology, 90(10):2808–2822.

Theodorou, A., Wortham, R. H., and Bryson, J. J. (2017). Designing and implementing transparency for real time inspection of autonomous robots. Connection Science, 29(3):230–241.

Toma, C. L. (2010). Perceptions of trustworthiness online: The role of visual and textual information. In Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, pages 13–22. ACM.

Tucker, B. (2018). ‘That’s problematic’: Tracing the birth of call-out culture. Critical Reflections: A Student Journal on Contemporary Sociological Issues, 6.

United Nations General Assembly (1948). Universal declaration of human rights. Technical report, New York.Vasu, N., Ang, B., Teo, T.-A., Jayakumar, S., Raizal, M., and Ahuja, J. (2018). Fake news: National security in the post-truth era. RSIS. In Policy Report. Singapore: Nanyang Technological University. Retrieved 24 May, 2020, from https://www.rsis.edu.sg/wp-content/uploads/2018/01/PR180313_Fake-News_WEB.pdf

Vieweg, S., Hughes, A. L., Starbird, K., and Palen, L. (2010). Microblogging during two natural hazards events: What Twitter may contribute to situational awareness. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI ’10, pages 1079–1088, New York, NY. ACM.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. Science, 359(6380):1146–1151.

Waytz, A., Heafner, J., and Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. Journal of Experimental Social Psychology, 52:113–117.

Webb, H., Jirotka, M., Stahl, B. C., Housley, W., Edwards, A., Williams, M., Procter, R., Rana, O., and Burnap, P. (2016). Digital wildfires: Hyper-connectivity, havoc and a global ethos to govern social media. ACM SIGCAS Computers and Society, 45(3):193–201.

Wilson, H. and Theodorou, A. (2019). Slam the brakes: Perceptions of moral decisions in driving dilemmas. In International Workshop in Artificial Intelligence Safety (AISafety), IJCAI, Macau.

Wilson, R. K. and Eckel, C. C. (2006). Judging a book by its cover: Beauty and expectations in the trust game. Political Research Quarterly, 59(2):189–202.

Woods, L. (2018). ICO reacts to use of data analytics in micro-targetting for political purposes reports: United Kingdom. European Data Protection Law Review (EDPL), 4:381–383.

Wu, Y., Kosinski, M., and Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. Proceedings of the National Academy of Sciences, 112(4):1036–1040.

Yamagishi, T. (2011). Trust: The evolutionary game of mind and society. Berlin: Springer, Berlin.

Zahavi, A. (1977). The testing of a bond. Animal Behaviour, 25:246–247.

Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. Journal of Information Technology, 30(1):75–89.

1 As of this writing, the impact of these assaults is still a matter of urgent research and debate, but the fact of significant, long-term, and on-going expenditure in the attempts has been established in both courts and academic writings (e.g. Machado and Konopacki, 2018; Woods, 2018; Landon-Murray et al., 2019).