Chapter 15
Artificial Intelligence and Pro-Social Behaviour

Joanna J. Bryson

15.1 Introduction

Collective agency is not a discrete characteristic of a system, but rather a spectrum condition. Individuals composing a collective must invest some resources in maintaining themselves as well as some in maintaining the collective's goals and structures. The question of how much to invest at which level of organisation is a complex one, for which there may be many viable solutions. For example, one might consider a married parent to be a member of three families—their parents', their partners' parents', and the new one they have created with their partner; a citizen of a village, state and country; an employee of at least one organisation, in which they may also be members of either orthogonal or nested teams; and a member of various other voluntary organisations. Some individuals will seek situations with more or fewer such memberships of collectives. Nevertheless, all of us constantly make choices—not always explicit—about how much attention and effort to devote to influencing the behaviour of each collective of which we are members.

Artificial intelligence is ordinarily seen as something quite separate from all the complexity of human social arrangements (Gunkel 2012). We picture AI as also having agency, like a human, then generally dismiss this vision as not possible, or at least not present. Such dismissal of AI is a mistake. *Intelligent* is not a synonym for *human*. Intelligence is just one attribute of humans, many other animals, and even plants (Trewavas 2005). In itself, intelligence does not determine personhood, nor is it sufficient for moral subjectivity. It is neither necessary nor sufficient for the autonomy that underlies moral agency. Mathematics is normally considered to require intelligence (Skemp 1961), yet calculators prove that arithmetic and geometry at

J.J. Bryson (✉)
Department of Computer Science, University of Bath,
Bath, UK
e-mail: jjb@alum.mit.edu

C. Misselhorn (ed.), *Collective Agency and Cooperation in Natural and Artificial Systems*, Philosophical Studies Series 122,
DOI 10.1007/978-3-319-15515-9_15

least can be conducted without a capacity for setting autonomous goals. Plants can autonomously pursue goals, and change their behaviour in response to their environment, but plants are not considered either moral patients deserving protection,[1] nor moral agents responsible for their actions. Therefore, artificial intelligence does not imply any sort of agency. Rather, like any other artefact, AI could be seen as an extension of human agency (Bryson and Kime 2011).

The purpose of this chapter is to examine how technology, particularly AI, is changing human collective behaviour and therefore both our collective and our individual agency. My intention is to be primarily descriptive, but there is of course a normative subtext, which I will attempt to make as explicit as possible. This primarily imposes on the final section of the chapter, Sect. 15.5, and results in some policy recommendations. This chapter's principal normative motivation is that society should better understand itself, so that it can better choose goals for the regulation and governance of AI, privacy and personal data. This is because by using our data, AI can generate predictions of our behaviour, which increases the utility of and propensity for investment at the collective level. These increases can result in changes not only to our societies, but as a consequence to the experience and meaning of being an individual.

In the following sections, I first further describe intelligence and the current state of AI. Next I describe current scientific understanding concerning why humanity is in its unique situation of knowing and therefore having responsibility, and how this relates to our tendency for collective and pro-social action. I will then describe a series of scientific results, some from social simulations, demonstrating the ease with which pro-sociality can evolve, and which elucidate the limits to which we and other species can and should invest in the collective. Finally, I close by using the models from the earlier sections to project consequences of the advances in AI on human culture and human collectives. These predictions will be based simply on extending my description of intelligence and AI to the models of social investment and examining their consequences.

15.2 Intelligence and the State of AI

To draw conclusions concerning the consequences of intelligence, we first need to define the term. For the purposes of this chapter, I will not attempt to capture its ordinary language meaning, but rather will introduce a simple, clear-cut computational definition of the term, which also relates to its characterisation in biology. Intelligence is capacity to

1. express an appropriate action,
2. in real time, and
3. in response to a perceived environment.

[1] Except where plants are seen as either a part of a broader ecosystem, or as a possession of a human.

Each of these components must be explained in turn. *Expressing an action* is necessary for intelligence to be judged—we will not consider any 'inner life' that is not demonstrated through some action, though action may include communication. *Appropriate* implies some goal, so any intelligent system has some metric by which its performance is judged. For biological systems, this is generally something related to survival.[2] For AI, we the makers define the goals. So for a calculator, it is sufficient to respond to button presses without noticing weather events. *In real time* is not a theoretical requirement of intelligence, but rather indicates that I am limiting my consideration to what also might be called cognitive systems (Vernon et al. 2007). It means that the agent exists in a dynamic environment, and can express action quickly enough that that action is generally still appropriate. "Generally" because of course very intelligent systems occasionally have traffic accidents—intelligence is not all-or-nothing, but rather varies in extent. I include the real-time requirement to focus on competences that find an appropriate action according to an agent's own sensing. This is to discriminate from processes like evolution or other abstract mathematical algorithms which may contribute to intelligence but do not produce a direct action outcome. Finally, *in response to a perceived environment* eliminates from consideration objects that act the same way at all times and just happen to sometimes be in an appropriate place and time when they do so. It also emphasises the importance of sensing to intelligence. Intelligence is judged by its actions as they relate to a context; the ability to perceive and discriminate contexts is therefore critical to intelligence.

Collective agency is not necessarily collective intelligence. Agency implies the capacity to be the author of environmental change. This change can be effected by a collective whether or not the authorship or motivation was achieved in a fully distributed way, as we might expect in collective intelligence (Williams Woolley et al. 2010). While intelligence originates change, that change can be effected by other agents that are not the original motivated entity. A captain may determine a team's strategy, a gardener may determine which wall an ivy will cover. On the other hand, observable collective intelligence is necessarily a form of collective agency. A swarm of insects may choose a new hive location (Marshall et al. 2009); a company may sue for changes in law enforcement (Rosenbaum 2014).

An Internet search is a highly intelligent process, requiring enormous capacity for perception—the perceptual ability to categorise billions of web pages based on a context set by search terms, and the action competence to serve one of these billions to your screen. But the agent responsible for the act, and that (principally[3]) benefits from that act is the human that requests the search. Here the expressed action of the individual user couldn't have been achieved without intelligence

[2] The arguments in this chapter hinge on inclusive fitness (Hamilton 1964; Gardner and West 2014) rather than individual survival, but I postpone that discussion here. It is appears in Sect. 15.3.

[3] Search companies record information about searches and the response of users to the web pages served, so those companies are also intelligent and motivated agents that benefit from the act of the search, but they do not originate it.

belonging to the corporation behind the search, but the motive force for the search is entirely individual.

A websearch is just one example of the AI-augmented individual capacities that have come to pervade twenty-first century life. Others include processes on phones that facilitate our communication, scheduling and even picture taking; AI in our word processors that increases our capacity to effectively communicate by checking grammar and spelling; filters on email that detect spam; filters on credit card expenditures that detect possible fraud; and filters on surveillance cameras that recognise faces, license-plate numbers, and even detect the emotions and intentions underlying human voices and gestures (Valstar and Pantic 2012; Griffin et al. 2013; Eyben et al. 2013; Kleinsmith and Bianchi-Berthouze 2013; Hofmann et al. 2014).

This intelligence enhances the agency of both individuals and corporations (see footnote 4) but has not produced a set of independent artificial actors competing with us for resources as imagined in science fiction. This lack of immediate, apparent, competitive threat, plus a heavy cultural investment in the privileges assumed to associate with human uniqueness, lead many to dismiss the possibility of AI, at the very time it is not only present but fundamentally changing our individual and social capacities.

AI has not yet caused significant change to our direct mechanical capacity for action. In terms of physically altering the world, AI requires a robot. The most prominent robots today are mechanisations of machines we can also use without AI, such as vacuum cleaners, cars and other tele-operated vehicles. We are now capable of acting much faster and at a much greater distance than we could before, but this is primarily due to improvements in telecommunication which are largely (though not entirely) independent of AI.

The way in which current AI fundamentally alters humanity is by altering our capacity for perception—our ability to sense what is in the world. Part of this is also due to communication. For example, we can now see what is happening very far away very quickly. But much more than this, we can remember and recall identical or even similar situations to one we presently observe. Other apes can do that too, but with language and subsequently writing, humans have had a special advantage which is that we can recall situations and actions we have not directly experienced. The reason that we can exploit similar rather than just identical previous contexts for recall is because we store this knowledge in abstract models. Abstraction saves storage space, but even more importantly allows for generalisation to new situations (Bishop 2006). In the simplest case we can find a 'near neighbour' context, treat the present one as the same, and expect similar outcomes (Lopez De Mantaras et al. 2005). Beyond this, we can use models to extrapolate to conditions we have not yet seen, so long as variation within the models tends to be continuous, and the new context does not differ extremely from our historical record. In such conditions we can generate novel variations on previous actions to meet the new conditions (Schaal and Atkeson 1998; Huang et al. 2013).

These processes are ordinarily referred to as machine learning. The reason I am describing them rather as perception is this: I want to emphasise that a great deal of intelligence is the problem of learning to recognise the categories of contexts in

which a particular action is appropriate. Or another way to think of this is that with enough experience and a well-structured model for storing and recalling that experience, we can use the past to recognise the present and therefore predict and address the future.

There are two reasons that AI is generating staggering increases in humanity's available intelligence. First, the basic concepts of learning in general and machine learning in particular described above have been understood for decades (Hertz et al. 1991). In those decades our algorithms for building models have been steadily improving—the recently-trendy deep learning is just one of many fundamental improvements made over that time (Jacobs et al. 1991; McLachlan and Krishnan 2008; Hinton et al. 2006; Le Roux and Bengio 2008). Second, we have found ways to both acquire and store the data that makes up experience in digital format. Thus our models are better, bigger and over a vastly wider variety of human experience. For example, we no longer need to guess why and how people will vote or riot—enough of them happily broadcast their intentions and concerns on the Internet. The important thing to understand is that our models have become sufficiently good, that even where our data is biased, often we can compensate for that bias and still make accurate predictions (Beauchamp 2013; Wang et al. 2015; Rothschild et al. 2014).

15.3 Cooperation and Collective Agency

Prosthetic intelligence affects our lives in innumerable ways, most notably simply by allowing us to make more informed decisions, whether by providing more immediate access to restaurant reviews, health care advice or the day's weather forecast. But in this chapter, and in keeping with the rest of this book, I focus on an even more fundamental aspect of human behaviour—our propensity for cooperation and information sharing, and how exponential rates of improvement in our AI may affect these.

The human propensity for cooperation is often seen as unique (Sober and Wilson 1998; Henrich et al. 2001). There is no denying that humans are extraordinary in a number of ways: the extent and variety of our built culture, our language and written histories, and our recent domination of the planet's biomass (Haberl et al. 2007; Barnosky 2008). These indicators of uniqueness are not necessarily or even likely to be independent. For example, our propensity to share information might explain why we have accumulated the culture that allows us to dominate other species. Science considers the simplest viable explanation for any phenomena to be the most likely, so many researchers have been searching for a single-point explanation for human uniqueness.

Cooperation is however not at all unique to humans. Assuming only that we are talking about observed cooperative behaviour, not cooperative intent or forward planning, then cooperation is ubiquitous in nature. For the purpose of this chapter, I will define cooperation as the expression of altruistic behaviour among a collection of individuals. For altruistic I use a standard definition: an action which at least

when executed is net costly to the actor, but provides net benefit to another agent. Although standard (Gintis et al. 2005), this definition is not universally accepted. Some biologists (and philosophers) are only happy to label an action 'altruistic' if over the entire lifetime of the individual its expected net value is costly, a situation which never occurs in nature (cf. Sylwester et al. 2014). However, the current understanding and explanation of cooperation in fields like economics and biological anthropology is that cooperation consists of costly actions that produce a public good. Even if the actor or their relatives are likely to get a disproportionate amount of that good, the fact that it facilitates communal benefits makes it cooperative (Burkart et al. 2014; Silva and Mace 2014; Taylor 2014). These types of explanations have been used to account for cooperation in nature—cooperation that often extends to one-way and even ultimate sacrifices by an individual agent for the collective good (Ackermann et al. 2008; Ferguson-Gow et al. 2014; Carter et al. 2014; Hobaiter et al. 2014).

By this definition, we can see that even the ultimately 'selfish' genes in fact exist entirely in cooperative contexts, collaborating with their competitors to compose multi-gene organisms (Dawkins 1976, 1982). The level of agency we are used to reasoning about as individual, that is macroscopic animals and plants, are the vehicles for hosts of competing replicators—genes, and arguably memes.[4] The vast majority of macroscopic life reproduces sexually, which is to say the individual agent is not replicated at all, but rather manages to replicate just (generally) half of its own genes in each of its offspring (Okasha 2012). However, these offspring are nearly always shared with another organism of the same species, and consequently necessarily share the vast majority of their replicators with both of their parents.

Cooperation between living individuals then is highly adaptive,[5] simply because copies of the same replicators that control the selection of the altruistic behaviour are very, very likely to reside in the individuals that receive the benefit (Hamilton 1964; Gardner and West 2014). This explanation of altruism is currently known as inclusive fitness, but has been mathematically related to the possibly more familiar concepts of kin selection and group selection (Gardner et al. 2011; Marshall 2011). To further complicate matters, social behaviour is in fact often controlled by replicators that are themselves socially communicated, whether in bacteria (Rankin et al. 2010), humans (Schroeder et al. 2014), or human institutions (Sytch and Tatarynowicz 2014).[6] What matters therefore is not overall relatedness, but a robust capacity of socialising behaviours to survive—presumably by replication—into the future.

[4] Memes are the hypothesised replicators for horizontal (non-genetic) transmission of behaviour. Like genes, they have yet to be precisely defined or measured (Mesoudi et al. 2004). It is also not yet clear the extent to which they change in frequency in accordance to Darwinian evolution (El Mouden et al. 2014). Nevertheless, memes are widely acknowledged as a useful abstraction for thinking about the transfer of traits expressed as behaviour between individuals by means other than biological reproduction.

[5] Adaptive in the biological sense of having been facilitating selection. The AI literature sometimes uses the term adaptive to mean plastic or mutable.

[6] Further, humans at least may choose to associate with those with similar gene structure even where they are not family members (Christakis and Fowler 2014).

To return to the conundrum of human uniqueness, my own hypothesis is that human uniqueness results not from a single cause but from a unique conjunction, at least in terms of extent, of two relatively common traits—a reliance on cognition, culture, and memory, found also in the other great apes and probably other long-lived species (Whiten et al. 1999; McComb et al. 2001; Krützen et al. 2005; Perry and Manson 2003); and a capacity for vocal imitation, something no other ape (or monkey) exhibits, but that has evolved several times apparently independently across a range of taxa[7] (Fitch 2000; Bispham 2006; Bryson 2008; Fitch and Zuberbühler 2013). Vocal imitation provides a communication medium sufficiently rich to support the redundancy necessary for an unsupervised learning process like evolution to operate across our vocalisations (Bryson 2009). Evolution over primate vocalisations, where selection is on both utility and memorability, could produce the system of human language (as per Smith and Kirby 2008; Wray 1998; Wray and Grace 2007). Our ape characteristics—long lives and memories, and predisposition to use culturally-acquired behaviour—allowed us to accumulate sufficient data to facilitate this process, and now allow the learning of complex languages.

Thus no one invented language. Language evolved as a public good, and with it an accumulating catalogue of complex, useful concepts—far more than one individual was otherwise likely to discover or invent for themselves (Dennett 2002). Language might be thought of as the first AI—it is an artefact that massively extends our individual levels of intelligence. As I introduced earlier and will argue more forcefully in the final section, taking our definition of AI to include the motivationless, locationless artefacts that are spoken and written language is a more useful and certainly less dangerous extreme than assuming something is not AI if it is not perfectly human-like. Regardless of whether you will accept language as AI, its intelligence-enhancing properties have consequences for the extent of our cooperation, as I discuss in the next section. Language and culture also may have spectacular consequences for human relatedness, as utilised in theoretical biology for computing the probability of altruistic acts due to inclusive fitness. Language and the culture that it facilitates increase the proportion of our replicators that are shared horizontally. This not only impacts the proportion of our relatedness, but also its plasticity, as humans can rapidly find and communicate ideas that discriminate as well as unite (Krosch and Amodio 2014).

The fact that our relatedness depends on socially-communicated replicators has significant ramifications for collective agency. Genes, individual animals, herds, families, villages, companies and religious denominations can all in some sense be said to be agents—they can all act in ways that effect change in the world. Many of these agencies are composed of others, and further at any level at which there can be seen to be action selection, there can also be seen to be evolutionary selection—at least some reenforcement for decisions taken, and some competition with other actors for limited resources (Wilson 1989; Keller 1999). Every such point of selection

[7] The capacity to recognise novel sounds and to learn novel contexts to express sounds should not be confused with the capacity for vocal imitation (Bryson 2009; Fitch and Zuberbühler 2013).

an agent faces presents them with an action-selection conundrum: how much resource (including, for entities that have it, attention) should that collection of opportunities and threats be allocated?

15.4 Factors Determining Investment in the Collective

15.4.1 Problem Specification and Methods

Before we can understand how AI may affect our identities and our societies, we need to form an understanding of how anything can affect these, and in what ways. In this section I address the question that both began this chapter and concluded the previous section: how do agents determine how much resource to devote to which level of the collectives in which they can have an effect? Because its answer hinges on perception and communication, this question will lead into the final discussion of AI's impact on our selves and our collectives.

Let us start by thinking about the problem in terms of a concrete case. An individual is living with a large number of others on a collectivised farm. This farm has been set in competition with other farms, so that whichever farm performs the best will be allocated more resources such as water, seed and fertiliser by the state. Unfortunately, as is often the case in collectivised farming, the system is not very efficient and not everyone is making enough money to have a family. Should our focal individual devote their time to raising their individual status within their own farm, so that their share of that farm's product is increased? Or should they devote their time to ensuring their farm will be more productive, so that the farm receives more income to distribute? Either strategy might reward the individual with the desired level of income. Also, the strategies are not entirely mutually exclusive: some time could be allocated to either, and if the individual is talented at managing then perhaps both could be achieved with the same actions.

In general in biology, wherever we have tradeoffs like these we find a diversity of solutions, with both different species and different individuals within species adopting different mixes of strategies (Darwin 1859). It is important to remember that while evolution is an optimising process, no species or individual is ever optimal. This is for two reasons: both because the world constantly changes, altering the criteria of 'optimum'; and because the number of possible strategies is inconceivably vast. The vast number of available strategies necessitates that any present solution is dependent not only on the optimising force of selection, but also on historical accidents that determined what available variation natural selection has been able to operate over.[8]

[8] The vast numbers of possible strategies is produced by a process called combinatorial explosion, which I explain in more detail in Sect. 15.4.2. The importance of having a varied set of available possible solutions in order for evolutionary selection to proceed is part of the 'Fundamental

Any such accident of variation may lead the locally-optimal strategy between two individuals to be different. For our farmer, the optimal decision for their strategising may depend on contexts local to the farm, such as opportunities for promotion based on the age of the management team, or might change by the year depending on the weather. In a good year, perhaps the best farms will be able to support a good standard of living for all employees, but in a drought it may be essential to be in the management tier. An individual farmer in a particular farm may have a better chance at promotion due to their charisma, or a better chance at a game-changing farming innovation due to their cleverness. The talents and position of close friends or family among fellow employees could also determine the better strategy.

As the example above illustrates, we are unlikely to determine a single optimal level of investment in a particular collective agency for any individual. However, we can describe a set of factors which influence the utility of investment at different levels, and describe models of how these relate to each other. These models can inform us about what strategies are most likely to be chosen, and how these probabilities might change when new technologies can be used to magnify or repress the impact of native characteristics. For example, if a new fertiliser is invented that allows all the farms to produce enough so every individual might be able to have a family, then this might eliminate the need to compete with other farms, and the farmer might best invest their time in ensuring equitable distribution within their own farm.

Factors contributing to individual versus group-level investment can be roughly decomposed into two categories:

1. Environmental: those factors exogenous to any of the agents' replicators, such as the weather, or that most individuals have very little influence over, such as international policy on banking or the environment.
2. Social: factors that influence how a collective can function, such as its capacity for communication, and the behavioural or genetic relatedness of its members (see discussion of inclusive fitness, above).

There is good evidence that variation in environmental context can determine the utility and structure of a collective. For example, spiteful, anti-social behaviour seems to increase in regions with a low GDP (Herrmann et al. 2008a) or scarce biomass (Prediger et al. 2013). Spite is the opposite of altruism—it is the willingness to pay a cost in to inflict a cost on others. This behaviour taken in isolation is necessarily maladaptive, as it hurts not only the individual but also another who almost certainly shares some measure of relatedness. It can only be accounted for if it covaries with some other attribute, for example if expressing spite increases social dominance and thus helps individuals in local competition (Rand et al. 2010; Powers

theorem of evolution' (Fisher 1930; Price 1972), and will be key in the final section of this chapter, Sect. 15.5.

et al. 2012).[9] These results imply that more cooperative behaviour occurs when resources are more prevalent, but doesn't explain a mechanism. Perhaps the costs of a competitive strategy are less attractive when relative status is not essential to survival, or perhaps cooperation is a riskier strategy more often chosen when participants are better resourced.

The objective of this chapter however is to examine the impact of AI on collective agency. While AI certainly does and will continue to affect the workings of our financial markets, our capacity to damage or protect the environment, and so forth; predicting the consequences of this impact requires an understanding of economics and politics beyond the scope of this chapter. Here I focus on what I've just termed the social aspects of investment in the collective. I review what is known about the 'individual' (animal- or vehicle-level) decision to invest in public rather than private goods. Then in this chapter's final section, I examine how prosthetic intelligence might be expected to alter values in these equations to change our level of investment, and as a consequence, our identity.

Much of the evidence presented in this and the previous sections, including the papers just cited by Rand et al. (2010) and Powers et al. (2012), derives from formal models including social simulation. Given this chapter's context in this volume—where simulation has been presented by some (e.g. Arnold 2015) as somehow controversial—I will briefly revisit why and how simulations are now an accepted part of the scientific method.

The role of simulations in science has been at times confused, not only by occasional bad practice (as with any method), but also by claims by some of the method's innovators that simulations were a "third way" to do science (after induction and deduction, Axelrod 1997). However, more recently a consensus has been reached that simulation and modelling more generally are indeed a part of ordinary science (Dunbar 2002; Kokko 2007; Seth et al. 2012). The part that they are is theory building. Every model is a theory—a very-well specified theory. In the case of simulations, the models are theories expressed in so much detail that their consequences can be checked by execution on a computer. Science requires two things: theories that explain the world, and data about the world which can be used to compare and validate the theories. A simulation provides no data about the world, but it can provide a great deal of 'data' about a theory. First, the very process of constructing a simulation can show that a theory is incoherent—internally contradictory, or incomplete, making no account for some part of the system intended to be explained (Axelrod 1997; Whitehouse et al. 2012). Secondly, modelling in general can show us a fuller range of consequences for a theory. This allows us to make specific, formal hypotheses about processes too complex to entirely conceptualise inside a single human brain (Dunbar 2002; Kokko 2007). The wide-spread acceptance of simulations as a part of the scientific method can be seen by their inclusion in the highest levels of academic publication, both in the leading general science journals and in

[9] There is decent evidence that association with dominance is indeed the ultimate evolutionary explanation for spiteful behaviour, see for a review Sylwester et al. (2013).

the flagship journals for specific fields ranging from biology through political science.

Fortunately, a theory expressed formally as a simulation can also be expressed in the traditional, informal, ordinary-language way as well. This is the technique I use to describe the 'outcomes' (implications) of simulations throughout this chapter.

15.4.2 Models of Social Investment

In order for evolution to direct individuals to invest at a collective level two conditions need to hold. First, there needs to be some inclusive-fitness advantage for the replicators involved in this 'directing' (cf. Sect. 15.3 above.) Second, this advantage has to be discoverable, and discovered. As mentioned in the first part of this section, evolution optimises but never finds an optimum, partly because it cannot evaluate all possible candidates due to the infinite size of the candidate pool. The size of this pool derives from the fact that candidate 'solutions' are composed of combinations of available features. The number of possible combinations is exponentially related to the number of features: it is the number of features per candidate (f) raised to the number of possible values for these features (v), or f^v. This problem of combinatorics affects all forms of directed plasticity—that is, any system capable of change which has an evaluation criterion. In the Computer Sciences, this problem is known as combinatorial explosion, and characterises both AI planning and (machine) learning. But the same problem characterises both evolution and cognition, and by 'cognition' I also mean to include both learning and planning, where they are done by an individual over their or its lifetime.

To address the first condition first, inclusive-fitness (IF) benefit has proven a spectacularly complicated concept to reason about, although its fundamental veracity has been demonstrated time and again in both simulation and empirical data (Gardner and West, 2014, for a recent special issue). What makes IF difficult is not only the confound of memetic as well as genetic replicators, but also the problem of net benefit. We share genes with all life, nevertheless predation—and grazing—evolve (Folse and Roughgarden 2012; Ledgard 2001). We tend to favour those with whom we share more relatedness, yet our survival also depends on the stability of the ecosystem to which we are adapted. Still, since the focus of this chapter is on the impact of AI, I will neglect the Gaia-style analysis of ecosystemic agency (see instead Margulis and Hinkle 1997) and focus primarily on collectives consisting of a single species. Even here, IF leads to wildly counterintuitive effects, such as that promiscuity in socially-monogamous animals can lead natural selection to favour strategies that benefit the public good, such as mutual defence and conflict resolution (Eliassen and Jørgensen 2014).

Within species, families, or even swarms of clonal microbia, understanding IF requires consideration of the net benefit of collaboration. The costs of cooperation are not limited to the costs of the altruistic act, but also include the costs of cohabiting with close genetic relatives. These cohabitation costs include competition for

resources ranging from food to shelter to mates, and increased exposure to biological threats such as disease and predation which will specialise to a particular species, immune system, and locale. In large animals the advantages of communal living have long puzzled biologists, with avoidance of predation via 'cover seeking' with a mob being a key hypothesis (Hamilton 1971). However this relationship is also not simple. Large populations also serve to attract predation and sustain disease (e.g. Bischof et al. 2014; Bate and Hilker 2013), though smaller group size does seem to increase predation risk (Shultz and Finlayson 2010). Recently in the megafauna literature there has been a new hypothesis: individuals in populations might benefit from information transmission, of which vigilance against predators is just a special case (Crockford et al. 2012; Chivers and Ferrari 2014; Hogan and Laskowski 2013; Derex et al. 2013). Transmission of behaviour may be at least as important as information about localised threats (Jaeggi et al. 2008; Dimitriu et al. 2014). Note that behaviour itself, when transmitted horizontally (that is, not by genes to offspring), must be transmitted as information via perception (Shannon 2001). But information is just one example of public goods held by non-human species. Others include territory (including food, shelter and even mating resources, Preuschoft and van Schaik 2000; Dunbar et al. 2009), physical shelters, even digestive enzymes (MacLean et al. 2010). Much of this cooperative production is performed by microbia, where in contrast to megafauna, genetic instructions for cooperative behaviour can be exchanged horizontally—even across species—and injected into the cellular organism to change a local population's behaviour (Rankin et al. 2010; Dimitriu et al. 2014).

Cooperation requires not only that the species affords some sort of cooperative behaviour (e.g. the genetic coding for collaboratively building a hive), but also the capacity to detect when it is a good time to invest in such an activity, and further who is or are the best partners with which to engage. This last is of particular interest, because we know that a variety of species appear to shift between cooperative phases of behaviour. Generalised reciprocity, first observed in Norwegian rats, is an increase in expression of altruistic behaviour that follows the observation of others engaged in cooperative acts (van Doorn and Taborsky 2012; Gray et al. 2015). This sort of behavioural flexibility might be thought useful for facilitating the spread of cooperation, since it allows potential cooperators to suppress cooperative behaviour in the presence of free riders that might exploit them. However such an interpretation may be biased. A better model might be more neutral, like our interpretation of the phase changes in collective behaviour exhibited by slime mould as an adaptation to localised environmental stress (Keller and Segel 1970; Leimgruber et al. 2014).

MacLean et al. (2010) have openly challenged the idea that cooperative behaviour (the creation of public goods) is always something to be maximised. They provide a case study of the production of digestive enzymes by the more altruistic of two isogenic yeast strains. The yeast must excrete these enzymes outside of their bodies (cell walls) as they can only directly absorb pre-digested food. The production of these enzymes is costly, requiring difficult-to-construct proteins, and the production of pre-digested food is beneficial not only to the excreting yeast but also to any other yeast in its vicinity.

In the case of single-cell organisms there is no choice as to whether to be free-riding or pro-social—this is determined genetically by their strain. But the two strategies are accessible to each other via a relatively common mutation. Natural selection performs the action selection for a yeast collective by determining what proportion of each strategy lives or dies. MacLean et al. (2010) demonstrate with both empirical experiments and models that selection operates such that the species as a whole benefits optimally. The altruistic strain in fact overproduces the public good (the digestive enzymes) at a level that would be wasteful if it were the only strategy pursued, while the free-riding strain underproduces. Where there are insufficient altruists free-riders starve, allowing altruists to invade. Where there are too few free-riders excess food accumulates, allowing free-riders to invade. Thus the greatest good—the most efficient exploitation of the available resources—is achieved by the species through a mixture of over-enthusiastic altruism and free riding. Why doesn't evolution just optimise the species as a whole to produce the optimal level of enzyme? Because the temporal cost (delay) associated with a single genome discovering a particular production level is greater than the temporal stability of that optimal value, which is of course determined by the dynamics of the ecosystem. In contrast, death and birth can be exceedingly rapid in microbia. A mixed population composed of multiple strategies, where the high and low producers will always over and under produce (respectively) and their proportions can be changed very rapidly is thus the best strategy for tracking the rate of environmental change—for rapidly responding to variation in opportunity.

Bryson et al. (2014) recently proposed that a similar dynamic may explain cultural variation in the extent of apparently anti-social, spiteful behaviour. This variation was originally observed by Herrmann et al. (2008a), but not explained. In the context of an anonymous economic game played in laboratories,[10] some proportion of nearly every population studied chose to punish (to pay a cost to penalise) altruists who were acting in a way that benefited the punishers. This sort of behaviour,

[10]These were public goods games (PGG). Participants were separated by partitions and were unable to directly interact with or identify other group members. They played games in groups of four, with each participant able to either keep all of the endowment received from the experimenter (20 experimental currency units; ECU) or contribute some portion of the endowment to the public good. At the end of a round, all contributions were combined and the sum multiplied by 1.6. The obtained amount was divided evenly amongst all of the group members, regardless of their contribution. The payoff of each participant was calculated by summing up the amount kept and the amount received from the public good. Ten rounds were played as described above, and also ten rounds with the addition of punishment: participants after seeing the contributions of other players to the pubic good and could decide how much they wished to spend on reducing the payoff to other players. Participants could spend up to 10 ECU punishing the other players. Each ECU spent on sanctioning resulted in 3 ECU being deducted from the payoff to the targeted individual. A participant's payoff was calculated by subtracting the amount of ECU spent on sanctioning and the deduction points received from other players from the payoff from the PGG. Received deductions were capped so as not to exceed PGG earnings. Participants did not receive information about who deducted points from their payoff, making punishment anonymous. At the end of the experiment, participants received real money in the local currency in exchange for the total ECU accumulated across all rounds. See further Sylwester et al. (2014); Herrmann et al. (2008b).

termed anti-social punishment (ASP), cannot be accounted for directly in evolutionary models, but must give some indirect benefit (as mentioned earlier, Sect. 15.4.1). Herrmann et al. (2008a) discovered that the propensity for ASP correlates with the gross domestic product (GDP) of the country where the experiments were conducted, and also with its rule of law as measured by the World Values Survey (Inglehart et al. 2004). Using the Herrmann et al data set, Sylwester et al. (2014) discovered that ASP results in a significant increase in variation in the level of investment in public goods, but not in any particular direction. In contrast, altruistic punishment (of free riders) produces a measurable increase in investment, while those receiving no punishment tend not to change their level of investment over repeated rounds of playing the game. This result is particularly striking because of the anonymous nature of the game—because individuals did not know who punished them, they could not tell whether they were being punished by those giving more or less than themselves.[11] Nevertheless, humans seem to be well-equipped to assess social context. We hypothesise that altruistic punishment is more likely to be coordinated, and coordinated punishment is taken as an indication of ingroup identity, signalling the construction of a collective, and this is what results in the increased investment. ASP in contrast signals a conflict over social status, which results in more varied behaviour, and therefore a greater potential rate of change for the society (Fisher 1930; Price 1972).

This series of hypothesised mechanisms for adjusting investment in different levels of agency is key to the purpose of this chapter—to consider how AI changes human collective agency. There are two points at which AI fundamentally changes our social capacities: detecting appropriate contexts for expressing cooperative behaviours an agent already knows, and the discovery or innovation of new cooperative behaviours with or without the contexts for their expression. Both of these points benefit by improved communication and superior perception.

Choosing appropriate partners is a particularly important part of detecting contexts for behaviour. Cooperative behaviour is most sustainable when the benefit received from the agent's cost will be high, and when there is similarly high benefit for low cost likely to be produced by the agent's collaborator(s). Thus where possible, cooperation often takes place in the context of a relationship where both the needs of the other and the likelihood of their reciprocation can be judged. Zahavi (1977) hypothesises that the time one agent spends with another is an honest signal of the value the first agent places on that relationship. Perry (2011) has used this bond-testing hypothesis to explain strange dysphoric games played amongst capuchins—monkeys well-known for both their intelligence and their aggressive coalition behaviour where coalitions are not necessarily formed with close relatives. Atkinson and Whitehouse (2011) suggest that time spent in mutual dysphoric situations underlies human religion, which serves the purpose of assuring human bonding across groups that require mutual support. Taylor (2014) has recently extended the bond-testing model, drawing attention to the fact that many human societies

[11] Those who gave the most or the least to the group could assess the nature of the punishment they received, but our results held even when these were excluded (Sylwester et al. 2014).

require temporally-expensive displays of investment in the lives of others with whom a family may have long-term economic relations, thus guaranteeing each other assistance in times of hardship. The time-costly displays (for example, constructing elaborate gifts) guarantee that a family is not making many shallow investments, but rather has only a few deeply-committed relationships.

In a more general and less specifically-human model than Taylor's, Roughgarden et al. (2006) propose that an explanation for physical intimacy (beyond what is necessary for procreation) may be that intimacy is a means of increased communication of physical status between potential coalition partners, allowing for the discovery of mutually-advantageous equilibria with respect to the extent of cooperative investment. The suggestion is that this intimacy goes beyond mere partner choice and timing to finding sufficient information about potential shared goals to afford new cooperative activities (Roughgarden 2012). Consider the implications of these results on the earlier discussion of human exceptionalism. Language has made humans the most extraordinary communicators in nature, and writing and AI have accelerated these effects. But our exceptional communication is not limited to deliberate or linguistic mechanisms—for an ape, even the amount of our communication by scent is exceptional (Stoddart 1990; Roberts and Havlicek 2011). This could well explain the exceptional extent of our cooperation.

To summarise, these models show that there will always be a tradeoff between investment in the individual and the collective. Individuals (at least some of them) must be sustained for the collective to exist, so investment can never go to the extreme of being fully collective. However there are a large number of situations in nature that are not zero sum—where altruism can evolve because the cost to the individual is lower than the benefit produced multiplied by the number of individuals helped, divided by their relatedness to the altruistic individual (Hamilton 1964). This idea of 'relatedness' is tricky though—it really depends only on how related the individuals are in whatever trait generates their social behaviour. Social behaviour may itself be transmitted socially, even in microbia (Rankin et al. 2010). Also, relatedness is judged based on the pool of others with which the individual competes. So if competition is imposed on a large scale such as when a government forces collective farms to compete between each other, two individuals in the same farm may seem more related than when a drought sets in and the members of the farm are set to competing with each other for survival (Lamba and Mace 2011; Powers et al. 2011).

We have also seen that investment strategies may vary within a population to the benefit of that population overall; provided that the various strategies are accessible to each other, again either by genetic or social transmission of the strategies. We have seen that selecting appropriate partners can increase the benefit-to-cost ratio, and thus support investing more heavily in cooperative, collective strategies. This selection is dependent on being able to perceive the needs and abilities of others. What would be the outcome for cooperative behaviour if we could exactly know the needs and interests—and predict the future behaviour—of our neighbours?

15.5 The Impact of AI on Human Cooperation and Culture

My main objective in this chapter is this: to convince you that AI is already present and constantly, radically improving; and that the threats and promises that AI brings with it are not the threats and promises media and culture have focussed on, of motivated AI or superintelligence that of themselves starts competing with humans for resources. Rather, AI is changing what collective agencies like governments, corporations and neighbourhoods can do. Perhaps even more insidiously, new affordances of knowledge and communication also change what even we as individuals are inclined to do, what we largely-unconsciously think is worth our while. 'Insidious' is not quite the right word here, because some of these effects will be positive, as when communities organise to be safer and more robust. But the fact that our behaviour can radically change without a shift in either explicit or implicit motivations—with no deliberate decision to refocus—seems insidious, and may well be having negative effects already.

As I indicated in Sect. 15.2, we are already in the process of finding out what happens when our ability to read and predict the behaviour of our fellows constantly improves, because this is the new situation in which we find ourselves, thanks to our prosthetic intelligence. Assuming the output of commercial AI remains available and accessible in price, then the models of the previous section tell us we should expect to find ourselves more and more operating at and influenced by the level of the collective. Remember that this is not a simple recipe for world-wide peace. There are many potential collectives, which compete for resources including our time. Also and it is possible to over-invest in many, perhaps most public goods. The models of Roughgarden and Taylor describe not systems of maximal cooperation, but rather systems of maximising individual benefit from cooperation. There are still physical and temporal limits to the number of people with whom we can best collaborate for many human goals (Dunbar 1992; Dunbar et al. 2009). We might nevertheless expect that our improved capacity to communicate and perceive can help us to achieve levels of cooperation not previously possible for threats and opportunities that truly operate at a species level, for example response to climate change or a new pandemic.

Our hopes should be balanced and informed though also by our fears. One concern is that being suddenly offered new capacities may cause us to misappropriate our individual investments of time and attention. This is because our capacity for cooperative behaviour is not entirely based on our deliberating intelligence or our individual capacity for plasticity and change. Learning, reasoning and evolution itself are facilitated by the hard-coding of useful strategies into our genetic repertoire (Depew 2003; Rolian 2014; Kitano 2004). For humans, experience is also compiled into our unconscious skills and expectations. These are mechanisms that evolution has found help us address the problems of combinatorics (see the first paragraph of Sect. 15.4.2). But these same solutions leave us vulnerable for certain pathologies. For example, a *supernormal stimulus* is a stimulus better able to trigger a behaviour than any that occurred in the contexts in which the paired association

between stimulus and response was learned or evolved (Tinbergen and Perdeck 1950; Staddon 1975). Supernormal stimuli can result from the situation where, while the behaviour was being acquired, there was no context in which to evolve or learn a bound for the expression of that behaviour, so no limits were learned. The term was invented by Tinbergen to describe the behaviour of gull chicks, who would ordinarily peck the red dot on their parent's bill to be fed, but preferred the largest, reddest head they could find over their actual parents'. Natural selection limits the amount of red an adult gull would ever display, but not the types of artefacts an experimental scientist might create. Similarly, if a human drive for social stimulation (for example) is better met by computer games than real people, then humans in a gaming context might become increasingly isolated and have a reduced possibility to meet potential mates. The successful use of search engines—quick access to useful information—apparently causes a reduction in actual personal memory storage (Ward 2013). This effect may be mediated by the successful searcher's increased estimation of cognitive self worth. Though Ward describes this new assessment as aberrant, it may in fact be justified if Internet access is a reliable context.

The social consequences of most technology-induced supernormal stimuli will presumably be relatively transient. Humans are learning machines—our conscious attention, one-shot learning, and fantastic communicative abilities are very likely to spread better-adapted behaviour soon after any such benign exploitation is stumbled over. What may be more permanent is any shift between levels of agency in power, action, and even thought as a consequence of the new information landscape. The increased transparency of other people's lives gives those in control more control, whether those are parents, communities or school-yard bullies. But control in this context is a tricky concept, linked also with responsibility. We may find ourselves losing individual opportunities for decision making, as the agency of our collectives become stronger, and their norms therefore more tightly enforced.

The dystopian scenarios this loss of individual-level agency might predict are not limited to ones of governmental excess. Currently in British and American society, children (including teenagers) are under unprecedented levels of chaperoning and 'protection'. Parents who 'neglect' their children by failing to police them for even a few minutes can be and are being arrested (Brooks 2014). Lee et al. (2010 special issue) document and discuss the massive increase over the last two decades in the variety as well as duration of tasks that are currently considered to be parenting. Lee et al suggest that what has changed is risk sensitivity, with every child rather than only exceptional ones now being perceive as 'at-risk', by both parents and authorities. This may not be because of increased behavioural transparency afforded by technology and AI. Another possible explanation is simply the increased value of every human life due to economic growth (Pinker 2012). But what I propose here is that the change is not so much in belief about the possibility of danger, as the actuality of afforded control. Social policing is becoming easier and easier, so we need only to assume a fixed level of motivation for such policing to expect the amount of actual policing to increase.

Another form of AI-mediated social change that we can already observe is the propensity of commercial and government organisations to get their customers or users to replace their own employees. The training, vetting and supervision that previously had to be given to an employee can now be mostly encoded in a machine—and the rest transferred to the users—via the use of automated kiosks. While the machines we use to check out our groceries, retrieve our boarding cards and baggage tags, or purchase post office services may not seem particularly intelligent, their perceptual skills and capacities for interaction are more powerful and flexible than the older systems that needed to be operated by experts. Of course, they are still not trivial to use, but the general population is becoming sufficiently expert in their use to facilitate their replacement of human employees. And in acquiring this expertise, we are again becoming more homogenous in our skill sets, and in the way we spend that part of our time.

With AI public video surveillance our motions, gestures, and whereabouts can be tracked; with speech recognition our telephone and video conversations can be transcribed. The fact some of us but not others spew information on social media will rapidly be largely irrelevant. As better and better models are built relating any form of personal expression (including purchases, travel, and communication partners) to expected behaviour (including purchases, votes, demonstrations, and donations), less and less information about any one person will be needed to predict their likely behaviour (Jacobs et al. 1991; McLachlan and Krishnan 2008; Hinton et al. 2006; Le Roux and Bengio 2008).

Although I've been discussing the likely homogenising impact of increased AI and increased collective-level agency, collective agency is not necessarily egalitarian or even democratic. Again we can see this in nature and our models of the behaviours of animals very similar to us. In non-human primates, troops are described as either 'egalitarian', where any troop member can protest treatment by any other, and conflict is frequent but not violent; or as 'despotic', where interaction is limited by the dominance hierarchy, aggression is unilateral from dominant to subordinate, and fights while few are bloody (Thierry 2007). Which structure a species uses is partially determined by historic accident (phylogeny, Shultz et al. 2011), but also significantly by the species' ecology. If a species' preferred food source is defensible (e.g. fruit rather than insects) then a species will be more hierarchical, as it will under the pressure for safer spatial positions produced by the presence of predators (Sterck et al. 1997). The choice between social orders is not made by the individual monkeys, but by the dynamics of their ecological context.

Similarly, we cannot say exactly the power dynamics we expect to see as a consequence of increasing agency at collective, social levels. However a worrying prediction might be drawn from Rawls (1980), whose theory mandates that a 'veil of ignorance' is necessary to ensure ethical governance. Those in power should be under the impression that any law they make might apply to any citizen, including themselves. Can such ignorance be maintained in an age of prosthetic intelligence? If not, if those in power can better know the likely social position of themselves and their children or even the likely outcome of elections (Wang et al. 2015), how will this affect our institutions? As uncertainty is reduced, can we ensure that those in

power will optimise for the global good, or will they be more motivated—and able—to maintain control?

The answers to these questions are not deterministic. The models presented in Sect. 15.4 make ranges of predictions based on interactions between variables, all of which can change. Our future will be influenced by the institutions and regulations we construct now, because these determine how easy it is to transition from one context into another, just as available variation partially determines evolution by determining what natural selection can select between (see footnote 9). Although many futures may be theoretically achievable, in practice the institutions we put in place now determine which futures are more likely, and how soon these might be attained.

Humans and human society have so far proved exceptionally resilient, presumably because of our individual, collective and prosthetic intelligence. But what we know about social behaviour indicates significant policy priorities. If we want to maintain flexibility, we should maintain variation in our populations. If we want to maintain variation and independence in individual citizens' behaviour, then we should protect their privacy and even anonymity. Previously, most people were anonymous due to obscurity. In its most basic form as absolute inaccessibility of information, obscurity may never occur again (Hartzog and Stutzman 2013). But previously, people defended their homes with their own swords, walls and dogs. Governments and other organisations and individuals are fully capable of invading our homes and taking our property, but this is a relatively rare occurrence because of the rule of law. Legal mandates of anonymity on stored data won't make it impossible to build the general models that can be used to predict the behaviour of individuals. But if we make this sort of behaviour illegal with sufficiently strong sanctions, then we can reduce the extent to which organisations violate that law, or at least limit their proclivity for publicly admitting (e.g. by acting on the information) that they have done so. If people have less reason to fear exposure of their actions, this should reduce the inhibitory impact on individuals' behaviour of our improved intelligence.

Already both American and European courts are showing signs of recognising that current legal norms have been built around assumptions of obscurity, and that these may need to be protected (Selinger and Hartzog 2014). Court decisions may not be a substitute though for both legislation and the technology to make these choices realistically available. Legislating will not be easy. In Europe there has been concern that the de facto mechanism of access to the public record has been removed as search engines have been forced not to associate newspaper articles with individuals' names when those individuals have asked to be disassociated from incidents which are entirely in the past (Powles 2014). As we do come to rely on our prosthetic intelligence and to consider those of our memories externalised to the Internet to be our own, such cases of who owns access to which information will become increasingly complex (Gürses 2010).

The evolution of language has allowed us all to know the concept of responsibility. Now we are moral agents—not only actors, but authors responsible for our creations. As philosophers and scientists we have also professional obligations with

respect to considering and communicating the impacts of technology to our culture (Wittkower et al. 2013). AI can help us understand the rapid changes and ecological dominance our species is experiencing. Yet that same understanding could well mean that the rate of change will continue to accelerate. We need to be able to rapidly create, negotiate and communicate coherent models of our dynamic societies and their priorities, to help these societies establish a sustainable future. I have argued that the nature of our agency may fundamentally change as we gain new insights through our prosthetic intelligence, resulting in new equilibria between collective versus individual agency. I've also described scientific models showing how these equilibria are established, and the importance of individual variation to a robust, resilient, mutable society. I therefore recommend that we encourage both legislatures and individual citizens to take steps to maintain privacy and defend both group and individual eccentricity. Further, I recommend we all take both personal and academic interest in our governance, so that we can help ensure the desirability of the collectives we contribute to.

Acknowledgments Thanks to Lydia Harriss, Catrin Misselhorn, Miles Brundage and Evan Selinger for encouragement and discussions; David Gunkel and Will Lowe for debate; Misselhorn, Brundage, Selinger and Robin Dunbar for comments on an earlier draft; and Thomas König and the University of Mannheim SFB 884, The Political Economy of Reforms, for a quiet office to work in.

References

Ackermann, Martin, Bärbel Stecher, Nikki E. Freed, Pascal Songhet, Wolf-Dietrich Hardt, and Michael Doebeli. 2008. Self-destructive cooperation mediated by phenotypic noise. *Nature* 454: 987–990.

Arnold, Eckhart. 2015. How models fail? A critical look at the history of computer simulations of evolution in cooperation. In this volume.

Atkinson, Quentin D., and Harvey Whitehouse. 2011. The cultural morphospace of ritual form: Examining modes of religiosity cross-culturally. *Evolution and Human Behaviour* 32: 50–62.

Axelrod, Robert. 1997. *The complexity of cooperation: Agent-based models of competition and collaboration*. Princeton: Princeton University Press.

Barnosky, Anthony D. 2008. Megafauna biomass tradeoff as a driver of quaternary and future extinctions. *Proceedings of the National Academy of Sciences* 105: 11543–11548.

Bate, Andrew M., and Frank M. Hilker. 2013. Predator–prey oscillations can shift when diseases become endemic. *Journal of Theoretical Biology* 316: 1–8.

Beauchamp, Nick. 2013. Predicting and interpolating state-level polling using twitter textual data. In *New directions in analyzing text as data workshop*, ed. Ken Benoit, Daniel Diermeier, and Arthur Spirling, London School of Economics, September.

Bischof, Richard, H. Ali Dondas, Kabir Muhammad, S. Hameed, and Muhammad A. Nawaz. 2014. Being the underdog: An elusive small carnivore uses space with prey and time without enemies. *Journal of Zoology* 293(1): 40–48.

Bishop, Christopher M. 2006. *Pattern recognition and machine learning*. London: Springer.

Bispham, John. 2006. Rhythm in music: What is it? Who has it? And why? *Music Perception* 24: 125–134.

Brooks, Kim. 2014. *The day I left my son in the car*. Salon. http://www.salon.com/2014/06/03/the_day_i_left_my_son_in_the_car/. Accessed 3 June 2014.
Bryson, Joanna J. 2008. Embodiment versus memetics. *Mind & Society* 7: 77–94.
Bryson, Joanna J. 2009. Representations underlying social learning and cultural evolution. *Interaction Studies* 10: 77–100.
Bryson, Joanna J., and Philip P. Kime. 2011. Just an artifact: Why machines are perceived as moral agents. In *Proceedings of the* 22nd *International Joint Conference on Artificial Intelligence*, 1641–1646. Barcelona: Morgan Kaufmann.
Bryson, Joanna J., James Mitchell, and Simon T. Powers. 2014. Explaining cultural variation in public goods games. In *Applied evolutionary anthropology: Darwinian approaches to contemporary world issues*, ed. M.A. Gibson and D.W. Lawson, 201–222. Heidelberg: Springer.
Burkart, Judith M., O. Allon, F. Amici, Claudia Fichtel, Christa Finkenwirth, Heschl Adolf, J. Huber, K. Isler, Z.K. Kosonen, E. Martins, E.J. Meulman, R. Richiger, K. Rueth, B. Spillmann, S. Wiesendanger, and C.P. van Schaik. 2014. The evolutionary origin of human hyper-cooperation. *Nature Communications* 5: 4747.
Carter, Alecia J., S. English, and Tim H. Clutton-Brock. 2014. Cooperative personalities and social niche specialization in female meerkats. *Journal of Evolutionary Biology* 27: 815–825.
Chivers, Douglas P., and Maud C.O. Ferrari. 2014. Social learning of predators by tadpoles: Does food restriction alter the efficacy of tutors as information sources? *Animal Behaviour* 89: 93–97.
Christakis, Nicholas A., and James H. Fowler. 2014. Friendship and natural selection. *Proceedings of the National Academy of Sciences* 111(Suppl 3): 10796–10801.
Crockford, Catherine, Roman M. Wittig, Roger Mundry, and Klaus Zuberbühler. 2012. Wild chimpanzees inform ignorant group members of danger. *Current Biology* 22: 142–146.
Darwin, Charles. 1859. *On the origin of species by means of natural selection*. London: John Murray.
Dawkins, Richard. 1976. *The selfish gene*. Oxford: Oxford University Press.
Dawkins, Richard. 1982. *The extended phenotype: The gene as the unit of selection*. Oxford: W.H. Freeman & Company.
Dennett, Daniel C. 2002. The new replicators. In *The encyclopedia of evolution*, vol. 1, ed. Mark Pagel, E83–E92. Oxford: Oxford University Press.
Depew, David J. 2003. Baldwin and his many effects. In *Evolution and learning: The Baldwin effect reconsidered*, ed. Bruce H. Weber and David J. Depew. Cambridge, MA: Bradford Books/MIT Press.
Derex, Maxime, Marie-Pauline Beugin, Bernard Godelle, and Michel Raymond. 2013. Experimental evidence for the influence of group size on cultural complexity. *Nature* 503: 389–391.
Dimitriu, Tatiana, Chantal Lotton, Julien Bénard-Capelle, Dusan Misevic, Sam P. Brown, Ariel B. Lindner, and François Taddei. 2014. Genetic information transfer promotes cooperation in bacteria. *Proceedings of the National Academy of Sciences* 111: 11103–11108.
van Doorn, Gerrit Sander, and Michael Taborsky. 2012. The evolution of generalized reciprocity on social interaction networks. *Evolution* 66: 651–664.
Dunbar, Robin I.M. 1992. Time: A hidden constraint on the behavioural ecology of baboons. *Behavioral Ecology and Sociobiology* 31: 35–49.
Dunbar, Robin I.M. 2002. Modelling primate behavioral ecology. *International Journal of Primatology* 23: 785–819.
Dunbar, Robin I.M., Amanda H. Korstjens, Julia Lehmann, and British Academy Centenary Research Project. 2009. Time as an ecological constraint. *Biological Reviews* 84: 413–429.
El Mouden, Claire, Jean-Baptiste André, Oliver Morin, and Daniel Nettle. 2014. Cultural transmission and the evolution of human behaviour: A general approach based on the Price equation. *Journal of Evolutionary Biology* 27: 231–241.
Eliassen, Sigrunn, and Christian Jørgensen. 2014. Extra-pair mating and evolution of cooperative neighbourhoods. *PLoS One* 9: e99878.

Eyben, Florian, Felix Weninger, Nicolas Lehment, Björn Schuller, and Gerhard Rigoll. 2013. Affective video retrieval: Violence detection in Hollywood movies by large-scale segmental feature extraction. *PLoS One* 8: e78506.

Ferguson-Gow, Henry, Seirian Sumner, Andrew F.G. Bourke, and Kate E. Jones. 2014. Colony size predicts division of labour in attine ants. *Proceedings of the Royal Society B: Biological Sciences* 281: 20141411. doi: 10.1098/rspb.2014.1411.

Fisher, Ronald A. 1930. *The genetical theory of natural selection*. Oxford: Oxford University Press.

Fitch, W. Tecumseh. 2000. The evolution of speech: A comparative review. *Trends in Cognitive Sciences* 4: 258–267.

Fitch, W. Tecumseh., and Klaus Zuberbühler. 2013. Primate precursors to human language: Beyond discontinuity. In *The evolution of emotional communication: From sounds in nonhuman mammals to speech and music in man*, ed. Eckart Altenmüller, Sabine Schmidt, and Elke Zimmerman, 26–48. Oxford: Oxford University Press.

Folse, Henry J., and Joan Roughgarden. 2012. Direct benefits of genetic mosaicism and intraorganismal selection: Modeling coevolution between a long-lived tree and a short-lived herbivore. *Evolution* 66: 1091–1113.

Gardner, Andy, and Stuart A. West. 2014. Inclusive fitness: 50 years on. *Philosophical Transactions of the Royal Society B: Biological Sciences* 369: 20130356.

Gardner, Andy, Stuard A. West, and Geoff Wild. 2011. The genetical theory of kin selection. *Journal of Evolutionary Biology* 24: 1020–1043.

Gintis, Herbert, Samuel Bowles, Robert Boyd, and Ernst Fehr. 2005. Moral sentiments and material interests: Origins, evidence, and consequences, Chapter 1. In *Moral sentiments and material interests: The foundations of cooperation in economic life*, ed. Herbert Gintis, Samuel Bowles, Robert Boyd, and Ernst Fehr, 3–39. Cambridge, MA: MIT Press.

Gray, Peter B., Justin R. Garcia, Benjamin S. Crosier, and Helen E. Fisher. 2015. Dating and sexual behavior among single parents of young children in the United States. *Journal of Sex Research* 52: 121–128.

Griffin, Harry J., Min S.H. Aung, Bernardino Romera-Paredes, Ciaran McLoughlin, Gary McKeown, William Curran, and Nadia Bianchi-Berthouze. 2013. Laughter type recognition from whole body motion. In *Affective computing and intelligent interaction* (ACII), 2013 Humaine Association conference, Geneva, CH, 349–355.

Gunkel, David J. 2012. *The machine question: Critical perspectives on AI, robots, and ethics*. Cambridge, MA: MIT Press.

Gürses, Fahriye Seda. 2010. *Multilateral privacy requirements analysis in online social network services.* Ph.D. thesis, Katholieke Universiteit Leuven, Department of Computer Science.

Haberl, Helmut, K. Heinz Erb, Fridolin Krausmann, Veronika Gaube, Alberte Bondeau, Christoph Plutzar, Simone Gingrich, Wolfgang Lucht, and Marina Fischer-Kowalski. 2007. Quantifying and mapping the human appropriation of net primary production in earth's terrestrial ecosystems. *Proceedings of the National Academy of Sciences* 104: 12942–12947.

Hamilton, William D. 1964. The genetical evolution of social behaviour. *Journal of Theoretical Biology* 7: 1–52.

Hamilton, William D. 1971. Geometry for the selfish herd. *Journal of Theoretical Biology* 31: 295–311.

Hartzog, Woodrow, and Frederic Stutzman. 2013. Obscurity by design. *Washington Law Review* 88: 385–418.

Henrich, Joseph, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath. 2001. Cooperation, reciprocity and punishment in fifteen small-scale societies. *American Economic Review* 91: 73–78.

Herrmann, Benedikt, Christian Thöni, and Simon Gächter. 2008a. Antisocial punishment across societies. *Science* 319: 1362–1367.

Herrmann, Benedikt, Christian Thöni, and Simon Gächter. 2008b. Supporting online material for antisocial punishment across societies. *Science* 319.

Hertz, John, Anders Krogh, and Richard G. Palmer. 1991. *Introduction to the theory of neural computation*. Redwood City: Addison-Wesley.

Hinton, Geoffrey, Simon Osindero, and Yee Teh. 2006. A fast learning algorithm for deep belief nets. *Neural Computation* 18: 1527–1554.

Hobaiter, Catherine, Anne Marijke Schel, Kevin Langergraber, and Klaus Zuberbühler. 2014. 'Adoption' by maternal siblings in wild chimpanzees. *PLoS One* 9: e103777.

Hofmann, Martin, Jürgen Geiger, Sebastian Bachmann, Björn Schuller, and Gerhard Rigoll. 2014. The TUM gait from audio, image and depth (GAID) database: Multimodal recognition of subjects and traits. *Journal of Visual Communication and Image Representation* 25: 195–206.

Hogan, Kelly E., and Kate L. Laskowski. 2013. Indirect information transfer: Three-spined sticklebacks use visual alarm cues from frightened conspecifics about an unseen predator. *Ethology* 119: 999–1005.

Huang, Bidan, Sahar El-Khoury, Miao Li, Joanna J. Bryson, and Aude Billard. 2013. Learning a real time grasping strategy. In *IEEE international conference on robotics and automation (ICRA)*, Karlsruhe, 593–600.

Inglehart, Ronald, Miguel Basánez, Jaime Díez-Medrano, Lock Halman, and Ruud Luijkx (eds.). 2004. *Human beliefs and values: A cross-cultural sourcebook based on the 1999–2002 values surveys*. México: Siglo XXI Editores.

Jacobs, Robert A., Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive mixtures of local experts. *Neural Computation* 3: 79–87.

Jaeggi, Adrian V., Maria A. van Noordwijk, and Carel P. van Schaik. 2008. Begging for information: Mother–offspring food sharing among wild Bornean orangutans. *American Journal of Primatology* 70: 533–541.

Keller, Laurent. 1999. *Levels of selection in evolution*. Princeton: Princeton University Press.

Keller, Evelyn F., and Lee A. Segel. 1970. Initiation of slime mold aggregation viewed as an instability. *Journal of Theoretical Biology* 26: 399–415.

Kitano, Hiroaki. 2004. Biological robustness. *Nature Reviews Genetics* 5: 826–837.

Kleinsmith, Andrea, and Nadia Bianchi-Berthouze. 2013. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing* 4: 15–33.

Kokko, Hanna. 2007. *Modelling for field biologists and other interesting people*. Cambridge: Cambridge University Press.

Krosch, Amy R., and David M. Amodio. 2014. Economic scarcity alters the perception of race. *Proceedings of the National Academy of Sciences* 111: 9079–9084.

Krützen, Michael, Janet Mann, Michael R. Heithaus, Richard C. Connor, Lars Bejder, and William B. Sherwin. 2005. Cultural transmission of tool use in bottlenose dolphins. *Proceedings of the National Academy of Sciences of the United States of America* 102: 8939–8943.

Lamba, Shakti, and Ruth Mace. 2011. Demography and ecology drive variation in cooperation across human populations. *Proceedings of the National Academy of Sciences* 108: 14426–14430.

Le Roux, Nicolas, and Yoshua Bengio. 2008. Representational power of restricted Boltzmann machines and deep belief networks. *Neural Computation* 20: 1631–1649.

Ledgard, Stewart F. 2001. Nitrogen cycling in low input legume-based agriculture, with emphasis on legume/grass pastures. *Plant and Soil* 228: 43–59.

Lee, Ellie, Jan Macvarish, and Jennie Bristow. 2010. Risk, health and parenting culture. *Health, Risk & Society* 12: 293–300.

Leimgruber, Kristin L., Adrian F. Ward, Jane Widness, Michael I. Norton, Kristina R. Olson, Kurt Gray, and Laurie R. Santos. 2014. Give what you get: Capuchin monkeys (Cebus apella) and 4-Year-old children pay forward positive and negative outcomes to conspecifics. *PLoS One* 9: e87035.

Lopez De Mantaras, Ramon, David McSherry, Derek Bridge, David Leake, Barry Smyth, Susan Craw, Boi Faltings, Mary Lou Maher, Michael T. Cox, Kenneth Forbus, Mark Keane, Agnar Aamodt, and Ian Watson. 2005. Retrieval, reuse, revision and retention in case-based reasoning. *The Knowledge Engineering Review* 20: 215–240.

McComb, Karen, Cynthia Moss, Sarah M. Durant, Lucy Baker, and Soila Sayialel. 2001. Matriarchs as repositories of social knowledge in African elephants. *Science* 292: 491–494.

McLachlan, Geoffrey, and Thriyambakam Krishnan. 2008. *The EM algorithm and extensions*, vol. 382, 2nd ed. Hoboken, NJ: Wiley.

MacLean, R. Craig, Ayari Fuentes-Hernandez, Duncan Greig, Laurence D. Hurst, and Ivana Gudelj. 2010. A mixture of "cheats" and "co-operators" can enable maximal group benefit. *PLoS Biology* 8: e1000486.

Margulis, Lynn, and Gregory Hinkle. 1997. *The biota and gaia. In slanted truths*, 207–220. New York: Springer.

Marshall, James A.R. 2011. Ultimate causes and the evolution of altruism. *Behavioral Ecology and Sociobiology* 65: 503–512. doi:10.1007/s00265-010-1110-1.

Marshall, James A.R., Rafal Bogacz, Anna Dornhaus, Robert Planqué, Tim Kovacs, and Nigel R. Franks. 2009. On optimal decision-making in brains and social insect colonies. *Journal of the Royal Society Interface* 6: 1065–1074.

Mesoudi, Alex, Andrew Whiten, and Kevin N. Laland. 2004. Is human cultural evolution Darwinian? Evidence reviewed from the perspective of the origin of species. *Evolution* 58: 1–11.

Okasha, Samir. 2012. Social justice, genomic justice and the veil of ignorance: Harsanyi meets Mendel. *Economics and Philosophy* 28: 43–71.

Perry, Susan. 2011. Social traditions and social learning in capuchin monkeys (Cebus). *Philosophical Transactions of the Royal Society, B: Biological Sciences* 366: 988–996.

Perry, Susan, and J.H. Manson. 2003. Traditions in monkeys. *Evolutionary Anthropology* 12: 71–81.

Pinker, Steven. 2012. *The better angels of our nature: The decline of violence in history and its causes*. London: Penguin.

Powers, Simon T., Alexandra S. Penn, and Richard A. Watson. 2011. The concurrent evolution of cooperation and the population structures that support it. *Evolution* 65: 1527–1543.

Powers, Simon T., Daniel J. Taylor, and Joanna J. Bryson. 2012. Punishment can promote defection in group-structured populations. *Journal of Theoretical Biology* 311: 107–116.

Powles, Julia. 2014. What we can salvage from 'right to be forgotten' ruling. *Wired,* 15 May.

Prediger, Sebastian, Björn Vollan, and Benedikt Herrmann. 2013. Resource scarcity, spite and cooperation. *German Institute of Global and Area Studies (GIGA) working papers* 227. Hamburg.

Preuschoft, Signe, and Carel P. van Schaik. 2000. Dominance and communication: Conflict management in various social settings, Chapter 6. In *Natural conflict resolution*, ed. Filippo Aurel and Frans B.M. de Waal, 77–105. Berkeley, CA: University of California Press.

Price, George R. 1972. Fisher's 'fundamental theorem' made clear. *Annals of Human Genetics* 36: 129–140.

Rand, David G., Joseph J. Armao, Mayuko Nakamaru, and Hisashi Ohtsuki. 2010. Anti-social punishment can prevent the co-evolution of punishment and cooperation. *Journal of Theoretical Biology* 265: 624–632.

Rankin, Daniel J., Eduardo P.C. Rocha, and Sam P. Brown. 2010. What traits are carried on mobile genetic elements, and why? *Heredity* 106: 1–10.

Rawls, John. 1980. Kantian constructivism in moral theory. *The Journal of Philosophy* 77: 515–572.

Roberts, S. Craig, and Havlicek Jan. 2011. Evolutionary psychology and perfume design. In *Applied evolutionary psychology*, ed. S. Craig Roberts, 330–348. Oxford: Oxford University Press.

Rolian, Campbell. 2014. Genes, development, and evolvability in primate evolution. *Evolutionary Anthropology: Issues, News, and Reviews* 23: 93–104.

Rosenbaum, Sara. 2014. When religion meets workers' rights: Hobby Lobby and Conestoga Wood Specialties. *Milbank Quarterly* 92: 202–206.

Rothschild, David, Sharad Goel, Andrew Gelman, and Douglas Rivers. 2015. The mythical swing voter. In *Collective intelligence*, MIT. Unpublished preprint presented at conference. Available on arXiv:1406.7581.

Roughgarden, Joan. 2012. Teamwork, pleasure and bargaining in animal social behaviour. *Journal of Evolutionary Biology* 25: 1454–1462.

Roughgarden, Joan, Meeko Oishi, and Erol Akçay. 2006. Reproductive social behavior: Cooperative games to replace sexual selection. *Science* 311: 965–969.

Schaal, Stefan, and Christopher G. Atkeson. 1998. Constructive incremental learning from only local information. *Neural Computation* 10: 2047–2084.

Schroeder, Kari Britt, Gillian V. Pepper, and Daniel Nettle. 2014. Local norms of cheating and the cultural evolution of crime and punishment: A study of two urban neighborhoods. *PeerJ* 2: e450.

Selinger, Evan, and Woodrow Hartzog. 2014. Obscurity and privacy. In *Routledge companion to philosophy of technology*, ed. Joseph Pitt and Ashley Shew. New York: Routledge.

Seth, Anil K., Tony J. Prescott, and Joanna J. Bryson (eds.). 2012. *Modelling natural action selection*. Cambridge: Cambridge University Press.

Shannon, Claude E. 2001. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review* 5: 3–55.

Shultz, Susanne, and Laura V. Finlayson. 2010. Large body and small brain and group sizes are associated with predator preferences for mammalian prey. *Behavioral Ecology* 21: 1073–1079.

Shultz, Susanne, Christopher Opie, and Quentin D. Atkinson. 2011. Stepwise evolution of stable sociality in primates. *Nature* 479: 219–222.

Silva, Antonio S., and Ruth Mace. 2014. Cooperation and conflict: Field experiments in Northern Ireland. *Proceedings of the Royal Society B: Biological Sciences* 281: 20141435.

Skemp, Richard R. 1961. Reflective intelligence and mathematics. *British Journal of Educational Psychology* 31: 45–55.

Smith, Kenny, and Simon Kirby. 2008. Cultural evolution: Implications for understanding the human language faculty and its evolution. *Philosophical Transactions of the Royal Society, B: Biological Sciences* 363: 3591–3603.

Sober, Elliott, and David Sloan Wilson. 1998. *Unto others: The evolution and psychology of unselfish behavior*. Cambridge, MA: Harvard University Press.

Staddon, John Eric R. 1975. A note on the evolutionary significance of "supernormal" stimuli. *The American Naturalist* 109: 541–545.

Sterck, E.H.M., D.P. Watts, and C.P. van Schaik. 1997. The evolution of female social relationships in nonhuman primates. *Behavioral Ecology and Sociobiology* 41: 291–309.

Stoddart, David Michael. 1990. *The scented ape: The biology and culture of human odour*. Cambridge: Cambridge University Press.

Sylwester, Karolina, Benedikt Herrmann, and Joanna J. Bryson. 2013. Homo homini lupus? Explaining antisocial punishment. *Journal of Neuroscience, Psychology, and Economics* 6: 167–188.

Sylwester, Karolina, James Mitchell, and Joanna J. Bryson. 2014. Punishment as aggression: Uses and consequences of costly punishment across populations. To be resubmitted.

Sytch, Maxim, and Adam Tatarynowicz. 2014. Friends and foes: The dynamics of dual social structures. *Academy of Management Journal* 57: 585–613.

Taylor, Daniel J. 2014. Evolution of the social contract. Ph.D. thesis, University of Bath, Department of Computer Science.

Thierry, Bernard. 2007. Unity in diversity: Lessons from macaque societies. *Evolutionary Anthropology* 16: 224–238.

Tinbergen, N., and A.C. Perdeck. 1950. On the stimulus situation releasing the begging response in the newly hatched herring gull chick (Larus argentatus argentatus Pont.). *Behaviour* 3: 1–39.

Trewavas, Anthony. 2005. Green plants as intelligent organisms. *Trends in Plant Science* 10: 413–419.

Valstar, Michel F., and Maja Pantic. 2012. Fully automatic recognition of the temporal phases of facial actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 42: 28–43.

Vernon, David, Giorgio Metta, and Giulio Sandini. 2007. A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents. *IEEE Transactions on Evolutionary Computation* 11: 151–180.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. Forecasting elections with non-representative polls. International Journal of Forecasting. In press.

Ward, Adrian F. 2013. Supernormal: How the internet is changing our memories and our minds. *Psychological Inquiry* 24: 341–348.

Whitehouse, Harvey, Ken Kahn, Michael E. Hochberg, and Joanna J. Bryson. 2012. The role for simulations in theory construction for the social sciences: Case studies concerning divergent modes of religiosity. *Religion, Brain & Behavior* 2: 182–224.

Whiten, Andrew, Jane Goodall, William C. McGew, Toyoaki Nishida, Vernon Reynolds, Yukimaru Sugiyama, Caroline E.G. Tutin, Richard W. Wrangham, and Christophe Boesch. 1999. Cultures in chimpanzees. *Nature* 399: 682–685.

Williams Woolley, Anita, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *Science* 330: 686–688.

Wilson, David Sloan. 1989. Levels of selection: An alternative to individualism in biology and the human sciences. *Social Networks* 11: 257–272. Special issue on non-human primate networks.

Wittkower, D.E., Evan Selinger, and Lucinda Rush. 2013. Public philosophy of technology: Motivations, barriers, and reforms. *Techné: Research in Philosophy and Technology* 17: 179–200.

Wray, Alison. 1998. Protolanguage as a holistic system for social interaction. *Language and Communication* 18: 47–67.

Wray, Alison, and George W. Grace. 2007. The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua* 117: 543–578.

Zahavi, Amotz. 1977. The testing of a bond. *Animal Behaviour* 25: 246–247.