How Society Can Maintain Human-Centric Artificial Intelligence

Joanna J. Bryson and Andreas Theodorou

Abstract Although not a goal universally held, maintaining human-centric artificial intelligence is necessary for society's long-term stability. Fortunately, the legal and technological problems of maintaining control are actually fairly well understood and amenable to engineering. The real problem is establishing the social and political will for assigning and maintaining accountability for artifacts when these artefacts are generated or used. In this chapter we review the necessity and tractability of maintaining human control, and the mechanisms by which such control can be achieved. What makes the problem both most interesting and most threatening is that achieving consensus around any human-centred approach requires at least some measure of agreement on broad existential concerns.

1 Introduction: Remit and Definitions

The greatest challenges of appropriately regulating artificial intelligence (AI) are social rather than technical. First, we cannot agree on a definition of the term, even though there are perfectly well established definitions of both *artificial* and *intelligence*. The primary problem is that we as humans identify as intelligent, which certainly is one of our characteristics, but that does not imply that *intelligent* means 'human-like'. We are not only intelligent but tall, long-lived, and terrestrial, at least compared to other vertebrates (animals with spines). So from the outset it should be clear that this chapter is not—or not principally—about artificial humans, but about all artefacts that are intelligent. This includes not only humanoid robots but a wide range of intelligent tools and services, including social media platforms,

Joanna Bryson
University of Bath, BA2 7AY, United Kingdom, e-mail: jjb@alum.mit.edu

Andreas Theodorou
University of Bath, BA2 7AY, United Kingdom, e-mail: a.theodorou@bath.ac.uk

driverless and AI-ehanced conventional automobiles, smartphones, spellcheckers, and thermostats.

The term *human* in this chapter will be reserved to mean members of the species *Homo sapiens* as a species is ordinarily recognised in biology. While it is at a minimum generous and possibly highly moral to concern ourselves about the well being of anything that could share phenomenological sensations such as pain and loneliness that members of our species do, it is essentially impossible that we will ever build something from metal and silicon that will be as phenomenologically similar to us as rats or cows are. So again, it is worth being clear from the outset that this chapter is not about humans that have been created via cloning, or other forms of intentional but slight alterations of what is fundamentally our evolved biological design. Rather, this chapter concerns artefacts built from the ground up, though we do mean to include systems with non-deterministic elements of design such as machine learning or random number generators. We will however leave discussions of problems concerning the phenomenological experiences of such artefacts until humanity has agreed to avoid the suffering of rats and cows.

Having said how we do not define 'intelligence,' it is now appropriate to discuss how we will. For the purpose of this chapter:

- An *agent* is anything capable of altering the world. This includes chemical agents.
- *Computation* is the systematic transformation of information from one state to another. Computation is a physical process, requiring time, space, and energy.
- *Intelligence* is a special case of computation, that generates a special form of agency where actions — alterations of the world — are generated from perception — sensing of the world. Intelligence is a property of an agent that allows that agent to change its world *in response to contexts*, to opportunities and challenges. This recognition and addressing of the environment is achieved via computation. This definition is widely used in both natural and artificial intelligence, and dates to at least the nineteenth century (Romanes, 1883).

Artificial intelligence (AI) is simply intelligence expressed by an artefact, which for simplicity we will define as something built intentionally by a human, or multiple humans working together.

We also define two more terms that are the real sources of societal concern that are often misdirected towards the term *intelligent*.

- A *moral agent* is an agent that a society holds responsible for its own actions.
- A *moral patient* is any entity that a society considers it to be the responsibility of moral agents to protect.

Whilst we may often think that such concepts must be universal—and certainly historical ethical systems such as religions will often lead us to believe this is so—in fact there is tremendous variation by society on these details. Only recently have many humans come to recognise climate as a moral patient. Different nations and even states within nations have different ages at which they consider a human to be old enough to vote, fight in a war, choose a marriage partner, or consent to sex.

Given that these are some of the most momentous decisions an individual can make, it is striking that there is no universal agreement on when moral agency is achieved. From this it becomes evident that ethics itself is a social construction. In fact arguably, ethics may be definitionally the means by which a society constructs itself, an idea explored at more length by Bryson (2018).

Finally, the title of this chapter implies that we already have Human-Centric AI. This is largely true, though arguably not entirely. We certainly do already have AI by the straight-forward definitions given here. First, we have technology like Web search, spell and grammar checking, and Global Positioning System (GPS) navigation systems—all AI that billions of people interact with daily. These are AI as service; intelligent systems that transform data into recommendations that we act upon, or not. But secondly, some would argue that our existing corporations and governments are excellent examples of AI (List and Pettit, 2011). True, these artefacts include humans as part of their systems, but they are also already exactly the sort of phenomena some describe when they use words or phrases like 'superintelligence' or 'artificial general intelligence.' Human society as a whole is increasing its capacity to learn exponentially, by extending ourselves through our artefacts, and also just by extending our own sheer numbers. Many of the artefacts benefiting this system are not AI, but simply communication, education, and nutritional technologies which make us as individuals smarter, and give us access to each other's capacities for intelligence. But the identified challenges of superintelligence such as run-away processes over consuming available resources (Bostrom, 2012) are a good description of humanity's present challenges with respect to sustainability.

The extent to which governments, corporations, and their technological tools are human-centric can be debated, but more often the debate concerns who among humanity benefits, not whether something other than a subpopulation of humans is truly benefiting. This chapter does not seek directly to solve this question, but does assume that governments and corporations at least are focused on and controlled by at least some set of humans. Our purpose is to show that similar or greater levels of control can and should be expressed over the AI products humans produce. At the highest level, the means by which this objective may be achieved is by maintaining ordinary levels of human accountability for the devices we produce. We will go into greater detail about how this can be achieved below, but first we discuss why it should be.

2 Why Maintain Human-Centric AI

As just admitted, 'maintaining' human-centric AI isn't exactly the situation we find ourselves in. To the extent that corporations or governments function to serve their own persistence even where that does not benefit humanity, then AI may already be seen as not human-centric. The extent to which this is the current situation is much debated. This will not be the focus of this chapter, but we will return to this question briefly at the end of this section. For the purpose of the present chapter, we will

assume that these institutions largely serve humanity, and that what we really mean by our title is that we wish AI to make the situation no worse than it is, and perhaps even to improve it.

There are many possible humanist reasons to maintain human-centred control. First, we should say that there are two possible alternatives, which actually amount to much the same thing. The first is that we lose control absolutely, and the second is that control is handed from humanity to our artefacts. Whilst there will always be anarchists and nihilists arguing for the former, we will neglect that option here since people holding such positions are unlikely to become organised enough to dismantle control globally. The latter though is seen by many as desirable or even necessary. Aware of their own mortality and that of civilisations and species as well (cf. Scranton, 2015), they put their hope in artefactual progeny. Perhaps this is because (ironically) they can exercise more control over artefacts than over biological progeny, or perhaps they mistakenly believe that machines (unlike humans) can be immortal or omniscient. The fact that the average working 'life' of an artefact is far, far shorter than the average lifespan of a human (or even a chimpanzee) is apparently regarded as irrelevant. Perhaps they think machines can be made self-repairing, but in this sense so are biological lineages (Taylor and Bryson, 2014). Again, that any purely-mechanical technology lineage we produce will exceed the lifespan of our biological lineage is phenomenally unlikely.

It seems that the problem is that AI is viewed not as a type of computation—a physical process, but as a type of math—an abstraction. Mathematics may be eternal and perfect, but that is because it is not real. Computation being a physical process requires time, space, an energy. Even if we are able to achieve long-term energy independence (at least relative to our level of our demand) at some stage, we will always be constrained by space and time.

The above are only reasons not to argue against human-centred AI, but here we give two reasons to argue for it. First, every aspect of our values—not only our ethics, and our human drives and desires, but also our sense of aesthetics—all of these have coëvolved with our species and societies in order to maintain our species and societies (Bryson and Kime, 2011). There is no coherent sense in having machines enjoy hedonism for us, although we can use machines to capture resources that we could not ourselves exploit, preventing them from being exploited by others. While some openly find pleasure in such an expression of power, it is not something we choose to openly condone here, and we doubt it would be condoned by the majority of any stable society were they to recognise this as being the impulse for their support of 'artificial life.'

Second, all social order is based on concepts and institutions of justice that unfortunately have human suffering at their core as a means of dissuasion (Bryson et al., 2017). Law may seem to create compensation, and we could imagine a machine (for example) financially compensating for its wrongful actions. But in fact, law is mostly about dissuasion. Laws and treaties are a means by which we set out agreed behaviour, and agreed costs of violating that behaviour. We have coëvolved with these institutions for so long that we really do *feel* like we've received some form of compensation when in fact we have only received justice. For example, if

someone kills your lover, and that person goes to jail, you have received nothing remotely like what you have lost, but you perceive victory. In fact, perhaps part of what you lost is social status and faith in the system, and perhaps justice returns these to you. But these abstractions exist in order to maintain social order, and rest upon our biological architecture that makes stress and pain pervasively dysphoric, and isolation and loss painful and stressful.

We cannot build machines that can so systemically experience such pervasive dysphoria. Probably we cannot build such a machine at all, but certainly we cannot build one for which we can guarantee its safety. In fact here we return to the idea that AI is already somewhat out of our control, if we accept the List and Pettit (2011) account of corporations as AI. Corporations are extended legal personhood as a legal convenience, but it's a convenience allowable only because real humans are dissuaded from doing wrong by human justice. And we should not have said 'because,' we should have said 'to the extent which.' A shell company dissociates the humans who would suffer if the company does wrong from the humans who decide what the company does (cf. Bryson et al., 2017). Weapons such as guns, airplanes and bombs, and also chains of command (military or corporate) similarly remove individuals at least some ways from the consequences of their decisions, which makes decisions with deeply aversive consequences easier to take.

The primary reason to maintain or even increase the extent to which AI is human-centric is that to do otherwise would be far more likely allow a greater dismantling of justice, resulting in greater human suffering, than it would be to produce a new form of social or somehow universal good.

3 Maintaining Human Control Through Design

There are two means by which human control may be maintained over AI. First, good design of AI systems allows us to ensure that intelligent systems operate within the parameters we expect. Contrary to some contemporary horror stories, machine learning (even DNN) doesn't make this difficult. It is not hard to ring fence what aspects of an AI system are subject to (can altered by) its own learning or planning. In fact, constraining processes like learning and planning allows them to operate more efficiently and effectively, as well as more safely. This is because the amount of computation (time, space, and energy) required is directly related to the amount of possibilities that need to be explored. Thus appropriate constraint is one of the main means for making any system, including humans, smarter. We teach students the sets of tools, facts, and approaches that have been shown to date most likely to produce useful outcomes.

The second means of maintaining human control is by holding those who build, own, or operate AI accountable for their systems through law and regulation. This approach will be described in the following section, but requires first understanding that the first approach is both possible and desirable. That is the focus of this section.

To begin with, it has long been established that the easiest way to tackle very large engineering projects is to decompose the problem wherever possible into subprojects, or *modules* (Bryson, 2000). For example, one component of a driverless car is the GPS navigation system, which has been so completely modularised that it is routinely used by enormous numbers of human drivers daily. There is no reason that a single automobile's 'mind' should alter the algorithm by which new routes are chosen, although the observations of an automobile may contribute to the crowd-sourced data on the current traffic on a road, or even the nuances of controlling a particular make of car. Here again, even if such a crowd-sourced learning strategy is used to recognise and avoid congestion, the constantly-updating models of the current traffic conditions will not alter the independent model of the underlying roads. Neither model will have any direct access to control over where or whether the car moves, which is another module still, or for the time being, more likely a human driver.

More generally, one method for designing modular decomposition for an AI systems is to assess what the system needs to know, and for each aspect of that knowledge, the best way to maintain that knowledge, as well as to exploit it. Here we describe one such approach to systems engineering real-time AI. We use this as a case to demonstrate what is possible, and then to illustrate the more general claims about accountability, transparency, regulation, and social control of AI made in the section following.

3.1 Behaviour Oriented Design

The above observation—that an ontology of required knowledge and its most convenient representation for expressing timely action should be used as the basis for modular decomposition for intelligent systems—is a core contribution of Behaviour Oriented Design (BOD), which is one methodology for systems engineering real-time AI systems (Bryson, 2001, 2003). BOD takes inspiration both from the well-established programming paradigm of object-oriented design (ODD) and its associated agile design (Cockburn, 2006; Gaudl et al., 2013), and an older but still very well-known AI systems-engineering strategy, called Behaviour-Base AI (BBAI, Brooks, 1991). Behaviour-based design lead to the first AI systems capable of moving at animal-like speeds, and many of its innovations are still extremely influential in realtime systems today. Its primary contribution was to emphasise design—specifically, modular design. Previous AI researchers, inspired by their interpretation of their own conscious experience, had expected to express the entire world in a system of logical perfection and then to take only the provably optimal action (Chapman, 1987). BBAI instead focuses on

1. the actions the system is intended to produce, and
2. the minimum, maximally-specialised perception required to generate each action.

In BBAI, each module derives action from its own dedicated perception.

While based in real-world experience of building robots, and as mentioned being the first approach that really succeeded in animal-like navigation at animal-like speeds, there were problems with BBAI as Brooks originally construed it. The first problem with this approach is coordinating the modules. Decomposing for simplicity is of little use if the subsequent coordination proves intractable. Second, Brooks' experience with traditional robot planning and the complexities of dealing with the world lead him to dismiss any real-time extension of intelligent control whatsoever. BBAI in its original form has no onboard planning (at least, no revision of the priorities encoded in the AI) nor learning whatsoever. Brooks initially claimed (like Lorenz before him) that embedding intelligence in its ecological niche was too delicate a problem to be open to risky processes like onboard learning, and that what appeared to be thought and learning were epiphenomenal suppositions imposed by us as observers as the organism interacted with a complex environment. "The world is its own best model" (Brooks, 1991). While this emphasis revolutionised AI by refocussing it on proper systems design, it cannot really account for all of human-like or even insect-like behaviour (Papaj and Lewis, 2012).

BOD connects AI properly back to systems engineering via OOD, affording safety in AI by exploiting BBAI-like modular architectures to limit the scope of machine learning, planning, or other real-time plasticity to the actions or skills requiring the capacity to accommodate change. Such architectural design is essential not only for safety, but also simply for computational tractability. As mentioned earlier, learning systems are faster and more likely to succeed if they are conducting their search over relevant possible capacities. Brains do the same thing. Contrary to Skinner (1935), pigeons can learn to peck for food or to flap their wings to escape shock, but not to flap their wings for food or to peck to avoid shock (Gallistel et al., 1991). Biological evolution also provides architecture as scaffolding for viable systems. Again in contrast to some sensationalist contemporary horror stories, there are in fact zero AI systems for learning chess that represent power switches, or have access to guns. No AI system built to learn chess will ever shoot someone that moves to turn it off at night[1].

BOD makes such common sense architectural decisions an explicit part of its development process. In general BOD is one means of using systems engineering to overcome problems of complexity for intelligence, by introducing an ontology, methodology, and architecture that promotes iterative design and testing. BOD includes common-sense heuristics for modular decomposition, documentation, refactoring, and code reuse. By using well-established OOD techniques, BOD allows decomposition of intelligent behaviour into multiple modules forming what we call a *behaviour library*. Behaviour library modules may wrap machine learning systems, or indeed commercial AI services providing specific capacities such as face recognition or navigation.

Stringing these modules together into a coherent agent requires then only specifying the priorities of that agent. Notice that multiple agents with completely different

[1] Another stupidity of the gun-toting, chess-learning murderous AI fairytale propagated by the Future of Life Institute is that real AI developers *prefer* our systems to do work while we sleep.

goals can be constructed from the same behaviour library, providing only that they either exploit the same type of hardware platform, or that the modules have been constructed to be platform-independent. Aspects of intelligence can also be hosted on servers or in clouds and accessed over an API, but of course for a real-time system much critical intelligent infrastructure needs to be hosted in a way such that the communication rate between modules and their embedded hardware substrate can be guaranteed. Further, any system learning proprietary information e.g. about its owner's household should probably better host such information securely and solely on site (Kalofonos et al., 2008).

3.2 Specifying a System's Priorities

One of the innovations of BOD compared to both BBAI and OOD is to simplify the problem of arbitrating between different modules that might otherwise produce contradictory actions away from a highly distributed, difficult to conceptualise or design network of dependencies, and back towards a more traditional hierarchical representation of priorities. Of course, there were good reasons for Brooks' original avoidance of these hierarchies, concerning efficiency. As Blumberg (1996) observed, action selection is only necessitated where there is competition for resources. If no such competition exists, the behaviour modules are able to work in parallel. However, many things are in this sense a resource, including a robot's physical location, direction of gaze, and what it can hold on to.

Bryson (2001) introduces POSH (Parallel-rooted, Ordered, Slip-Stack Hierarchical) action selection. These ideas were taken up also by the far better-named Behaviour Trees (BT) (Isla, 2015; Rabin, 2014) which function just as well for BOD systems engineering of real-time AI, but here we focus on our original nomenclature. For historic reasons, the data structure built from POSH (or BT) components, describing an agent's priorities, is termed a *plan*, and the part of the AI system that checks these priorities is called a *planner*. This is true even though the planner typically will not alter the POSH plans in the system, but the planner and the plans together determine the sequence of steps the agent takes in pursuing its goals, which might be more conventionally seen as a plan.

POSH plans combines faster response times similar to the fully reactive approaches for BBAI with a greater ease of developing goal-directed plans. A POSH plan consists of the following elements:

1. Drive Collection (DC): This is the root or apex of the plan's hierarchy. It contains a set of Drives and is responsible for giving attention to the highest priority Drive that presently could use that attention. The POSH planning cycle alternates between checking for what is currently the highest level priority that should be active, and then progressing work on that priority. This check is made hundreds or thousands of times a second, to ensure the system's highest priority goals (which should ensure its safety) are constantly monitored.

2. Drive (D): Allows for the design of behaviour in pursuit of a specific goal. Each drive maintains its execution state even when it is not the focus of planner attention, allowing the system to express coarse-grained parallelism even within prioritised actions, as well as independently by modules not requiring arbitration. Each drive specifies its own perceptual context which is suitable to or requires its deployment, while the Drive Collection as a whole maintains track of the multiple Drives' relative priorities.
3. Competence (C): A simpler hierarchical plan element for representing the priorities within a particular component of a plan (also known as a subplan). Competences are similar to the drive collection but lack the extra checks and mechanisms for concurrency, which are handled entirely at the top level or root D. Each competence contains one or more Competence Elements (CE), which also are associated with both a priority relative to the other CEs, and a context which can perceive and report when that element can execute. The highest-priority action that can execute will execute when the Competence receives attention.
4. Action Pattern (AP): These are simple sequences of actions and perceptions used to reduce the design complexity of a plan when such a sequence of actions can be determined in advance.
5. Action (A): A call to code in the behaviour library that sets a skilled act in motion. To maintain agility in the planner, actions should not block (wait for a final response in the world) but simply return immediately with a code indicating whether or not the action was successfully initiated. Other plan elements can be designed to watch for a context in which this action has succeeded or failed if that knowledge is essential. However, in both biology and AI quite often actions are just run 'open loop', without checks, and action selection is simply repeated in the next instant in the new context produced by the agent's actions or inactions as time has progressed.
6. Sense (S): Senses are very much like actions, and also depend on the behaviour library for their implementation. The difference is that they return a value indicating context, which may be used to determine for example whether a Drive or Competence should be released to execute, or an Action Pattern aborted.

Taken together, these plan elements are sufficient for expressing the goals of many systems. Of course, for complex systems with multiple, potentially conflicting goals (e.g. maintaining a job and maintaining a relationship, or hoovering the house and entertaining the dog) it may be useful for the order of priorities to shift over time. For this we have developed several systems of synthetic emotions or moods. Essentially, a mood or emotion is a special type of behaviour module that determines its own current priority. Drives linked to these emotions have from the drive collection's perspective the same level of priority, and a separate system ensures that only one of these at a time receives the focus of attention (Bryson and Tanguy, 2010; Gaudl and Bryson, 2018; Wortham et al., 2018).

3.3 Real-Time Debugging of Priorities

Another myth of AI is that systems should become as intelligent as humans and therefore not require any more training than a human. In reality, very few will want to put as much energy into training an AI system as is required to raise a child, or even to train an intern, apprentice, or graduate student. Programming is generally a far more direct and efficient way to communicate what is known and knowable about generating appropriate behaviour. However, debugging a complex, modular, real-time system requires more insight than ordinary programming. Further, we may well want to allow non-programmers to set priorities and choose between capacities for their agents once reliable behaviour libraries have been defined. Both of these activities requires an element of transparency to a system. Here we use *transparency* to mean that the direct workings of the system should be made apparent — visible and understandable (Bryson and Winfield, 2017; Theodorou et al., 2017).

Hierarchical definitions of priorities like POSH plans or Behaviour Trees offer a sensible means of transparency for either of these two applications: expert debugging or ordinary user understanding. Here we again describe novel work in our own group, but the basic concept behind this may be generalised to other forms of systems engineering. At Bath, we have developed a real-time visualisation system and debugger for POSH plans. The system, ABOD3, is based on, but a substantial revision and extension of, ABODE (A BOD Environment, originally built by Steve Gray and Simon Jones, Bryson et al., 2005; Brom et al., 2006). ABOD3, first described by Theodorou (2017) and shown here in Figure 1, allows the graphical visualisation of POSH-like plans. The editor, as seen in its architecture diagram in Figure 2, is implemented in such a way as to provide for expandability and customisation, allowing the accommodation of a wide variety of applications and potential users.

ABOD3 is designed to allow not only the development of reactive plans, but also the debugging of such plans in real time. The editor provides a user-customisable user interface (UI). Plan elements, their subtrees, and debugging-related information can be hidden, to allow different levels of abstraction and present only relevant information to the present development or debugging task. The graphical representation of the plan can be generated automatically, and the user can override its default layout by moving elements to suit needs and preferences. The simple UI and customisation allows the editor to be employed not only as a developer's tool, but also has been demontrated to present transparency-related information to naive users that helps them develop more accurate mental models of a mobile robot (Wortham et al., 2017a).

Plan elements flash as they are called by the planner and glow based on the number of recent invocations of that element. Plan elements without any recent invocations start dimming down, over a user-defined interval, until they return back to their initial state. This offers abstracted backtracing of the calls, and the debugging of a common problem in distributed systems: race conditions where two or more subcomponents are constantly triggering each other then interfering with or even cancelling each other's effects. Finally, ABOD3 can also support integration

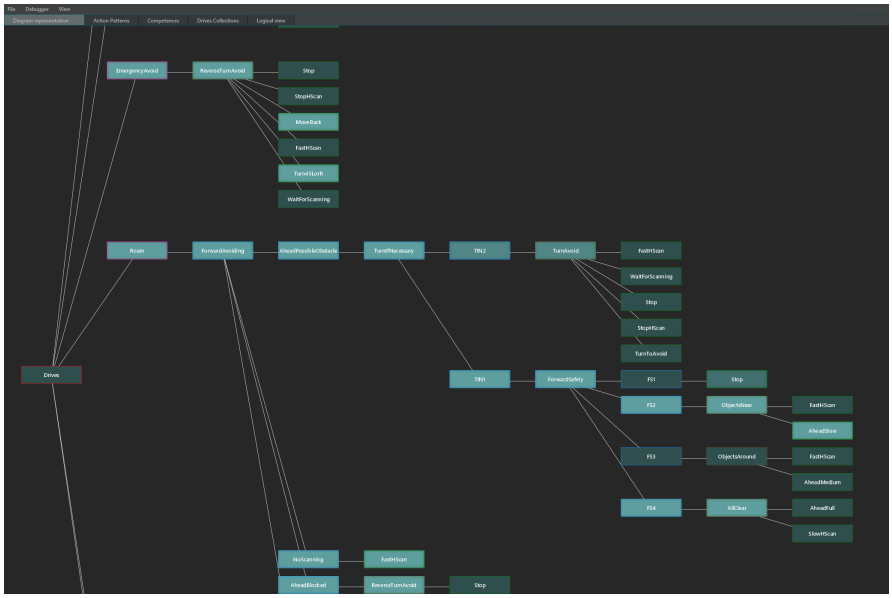

Fig. 1 The ABOD3 Graphical Transparency Tool displaying a POSH plan for a mobile robot, in debugging mode. The highlighted elements are the ones recently called by the planner. The intensity of the glow indicates the number of recent calls.

with videos of the agents in action, allowing for non-real-time debugging based on logged performance. Logging of actions taken and contexts encountered is a substantial aspect of AI accountability and transparency, which we will return to in the next section.

4 Maintaining Human Control Through Accountability and Transparency

To reiterate, although we have here described the systems-engineering approach and tools we have been developing together at the University of Bath, we are not claiming that these are the only, best, or most essential means for maintaining human control of AI. We are rather communicating that such control is perfectly possible, and illustrating examples of some of the technological mechanisms by which such control can be maintained. It is also perfectly possible to build AI for which accounting is not possible, indeed this too has already been done and is indeed too prevalent in our society (Pasquale, 2015). In this section, we summarise what is essential about technological mechanisms for human control, then close with a discussion about the social, legal, and political mechanisms for maintaining that control, which are actu-

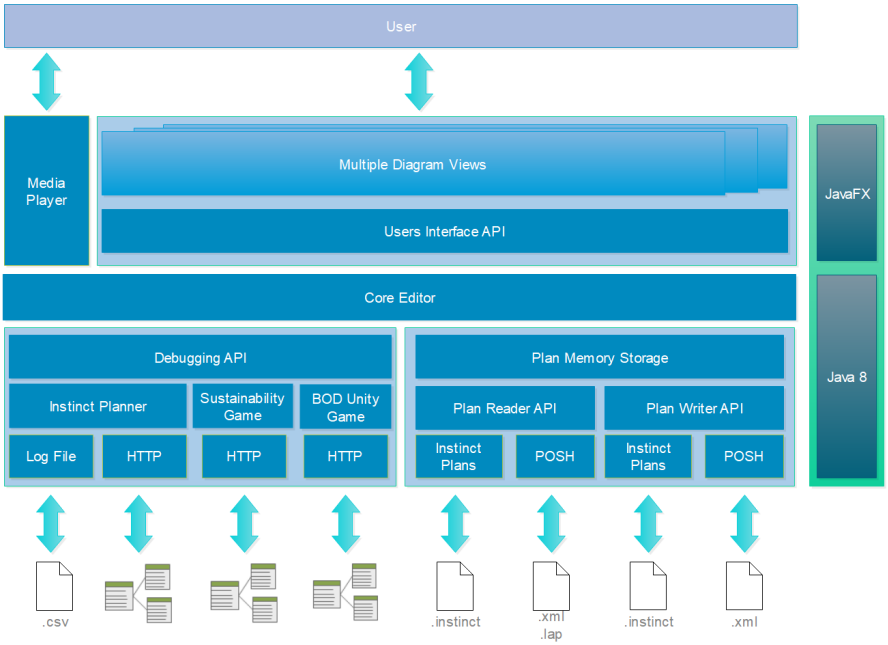

Fig. 2 System architecture diagram of ABOD3, showing its modular design. All of ABOD3 is written in Java to ensure cross-platform compatibility. APIs allow the support of additional BOD planners for real-time debugging or even multiple file formats for the plans. The editor is intended, through personalisation, to support roboticists, software AI developers, and ordinary users interested in AI systems.

ally far more important. Technology serves and extends human society, but ethics is what forms and defines human society.

4.1 Technological Mechanisms for Ensuring Transparency and Accountability

What is important to realise is that every aspect of an artefact is a consequence of design decisions. We are not saying that it is trivial to know what any AI system is doing. We *are* saying that it is possible to provide the tools and keep the records such that we know *at the level sufficient to maintain human accountability* what goes wrong with a system, if it goes wrong, and how and why it was constructed such that it did go wrong. There are social requirements underlying these technological features: can a person or a company show that they followed due diligence when they created an artefact? If not, they should be held liable for what that artefact does.

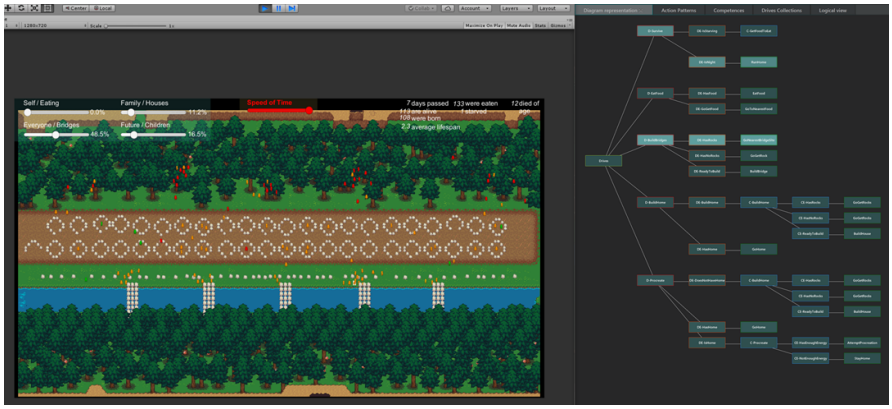

Fig. 3 ABOD3 implemented as part of a serious game (the Sustainability Game) so that game players can understand the interaction between agent motivations and the viability of an artificial community (Theodorou et al., 2019).

This does not mean that AI has to be deterministic or formally provably correct. Due diligence can be demonstrated by corporations despite the fact they employ people. People learn, have free will, and have incomprehensible systems of synapses making up their action selection mechanisms. Many humans are dishonest, careless, or sometimes just upset or incompetent. Nevertheless, we can construct systems that ensure that humans working together tend to succeed. These systems generally include records, such as financial accounts, access permissions, and minuted meetings where executive decisions are agreed. They also include external regulatory bodies and law enforcement.

Exactly the same kinds of procedures can be used for retaining control of AI, and indeed already are at least in well-regulated sectors like the automotive industry (Doll et al., 2015). In every single case so far concerning a human fatality involving a driverless car, newspapers have within a week shown us exactly what the car perceived, how that perception had been categorised, and what actions the car was taking at the point of fatality, and even why. Keeping records of this sort of information is not difficult, but it *is* a design decision. That decision is enforced in the automotive industry by its high levels of regulatory accountability mandated by the incredible amount of human suffering and death generated as its by-product (Williams, 1991). The design decision to provide adequate logging is one we can and should also enforce for other AI systems in socially critical roles.

As we described in the previous discussion of moduarlity and safety, the equivalents of ‘access permissions’ are also a completely standard part of design that anyone with any practical experience of creating an intelligent systems takes for granted. Every sensor or actuator a system has is an expense for its manufacturing, so these will naturally be restricted to those required by a system’s task, but further within the system, access to information can and should be restricted to that information likely to be useful, not only for safety but simply for efficiency.

In addition to logging what a system perceives and performs, we can also log every aspect of how we designed that system. Standard practice in software development is to use a software revision control system that documents the exact author and timing of any change to the system's software. Unfortunately, not every development team will exercise best practice in terms of ensuring that each individual developer has their own individual login, or documents the reasons for their changes, or documents the versions of software libraries used to support their programming, or data libraries used to train their machine learning. In fact, there has been a well-documented, scandalous disregard for the provenance of both software libraries (Gürses and van Hoboken, 2018) and data libraries (Pasquale, 2015). However, there is no technological reason that a better standard of practice couldn't be generated, and even required.

All of the mechanisms described above, and also the architectural concepts and software tools described in the previous section, are mechanisms of transparency. To be clear, when we talk about transparency here, we mean neither invisibility (as is sometimes advocated by human-computer interaction specialists), nor (necessarily) mandatory open-sourced code or formal, symbolic programming. The former—invisibility—actually *increases* the hazard of AI as ordinary users fail to realise their data is being gathered or to consider the consequences of compromising the security of the system. The latter can produce more information than humans can accommodate without resulting in clarity about responsibility or good practice. What is effectively transparent therefore varies by who the observer is, and what their goals and obligations are (Bryson and Winfield, 2017; Theodorou et al., 2017).

The goal of transparency is never complete comprehension. That would severely limit the scope of human achievement. Rather, the goal of transparency is providing sufficient information to ensure that at least human accountability, and therefore control, can be maintained.

Our position about transparency is supported by Dzindolet et al. (2003), who conducted a study where the participants had to decide whether they trust a particular piece of pattern recognition software. The users were given only the percentage of how accurate the prediction of their probabilistic algorithm was in each image. Yet, by having access to this easy-to-implement transparency feature, they were able to calibrate their trust in real time. Our own studies (discussed in Wortham et al., 2017a,b), demonstrate how users of various demographic backgrounds had inaccurate mental models about a mobile robot running a POSH planner. They were ascribing unrealistic functionalities, potentially raising their expectations for its intelligence and safety. When the same robot was presented with ABOD3 providing an end-user transparency visualisation, the users were able to calibrate their mental models. This lead to more realistic expectations concerning the system's capabilities, though interestingly also a higher assessment of its intelligence.

4.2 Maintaining Human Control Through Governance and Regulation

There have at various periods, including the present, been a worrying tendency to blame individual scientists or programmers for the consequences of their work. While it is true that individuals are accountable for their actions—including life choices concerning their employers—successful regulation requires looking at the entire context of that action. If we know there will or at least can be individuals who are dishonest, sloppy, suicidal, corrupted, or simply prone to occasional errors (that is, human), then we should expect systems containing such individuals should have some means for ensuring and promoting the quality of their work. For AI, the scale of this task may sound insurmountable—do we really think we should check the work of every individual programmer, globally? Who would do such a thing? Yet this is *exactly* what Apple does for individual programmers who want to write software applications for Apple's smart phone, the iPhone. Smart phones are the most fantastic information-gathering devices ever created, so it makes sense to have this level of security and scrutiny enforced by the maker and owner of this platform. Note also that despite the cost of such an operation, Apple has a perfectly successful business model for producing wealth as well as products.

We mentioned just above that car manufacturers already are developing vast amounts of AI in a highly regulated environment. At least some of them have also been able to successfully demonstrate that they practice due diligence when they are investigated by state prosecutors (Doll et al., 2015). But what about organisations that changed the world in unanticipated ways by introducing entirely new platforms and therefore capacities into societies and economies. Can they also be held accountable for damage done with the tools they've provided?

This is a question being addressed in courts and legislatures globally as we write, but we believe that the short answer is 'yes, to a point.' That point is demonstrated community standards of good practice. So if for example damage results from the obviously poor (and often illegal) standards of conduct documented by Pasquale (2015) or Gürses and van Hoboken (2018), then governments and other collectives should hold organisations that profit from this conduct accountable for the damage they cause. Similarly, if most organisations refuse to sell access to the data they collect from their users because doing so would seem a clear ethical violation, but some organisations do sell such data, then these latter organisations can be held accountable for violating the known ethical standards of their sector. This is particularly true for organisations of scale, who are routinely held to higher standards by the law because of their position of leadership. With great power (or even just money) does indeed come great responsibility.

In discussions we've held in the United Kingdom (UK) at least, it appears that there is not really a call for changes in legislation (House of Lords, 2018). Rather, what is needed is only to get through the fog of confusion caused by the smoke and mirrors associated with 'intelligence.' This is why we started this chapter as we did, to make it clear that AI and indeed I are ordinary properties amenable to both

science and law. Once this is clear, then with a little education and some good hiring, ordinary legal enforcement of liability standards should be sufficient to maintain human control.

AI does present two special problems however. One we've already mentioned but will come back to again here. There is a mistaken belief that the capacity to express human-like behaviour is in any way indicative of commonality of pheneomenological experience between machines and humans. As Caliskan et al. (2017) demonstrate, a glorified spread sheet that has just counted words on the Internet can report phenomenological commonality with humans, e.g. that flowers are more pleasant than insects, or even stereotyped beliefs such as that women's names are more associated with the domestic sphere. Such a system barely even qualifies as AI by the definition we've given since the only 'action' from its perception of the Web is the numerical report of what words are associated with what others. Further, these counts are replicated globally in standard AI tools, so there is no hazard of loss of a unique perspective if we destroy one of these spreadsheets, as there is if we lose a single human life, or even a unique copy of an old book or fossil. Humans act differently around robots that look human to them, but then humans act differently around statues that look human. Public spaces that had felt and been dangerous feel and become safer when ordinary human statues are introduced at ordinary human scale (Johnson, 2017). Thus reports of phenomenological similarity generated either by AI or by human observers cannot be seen as valid demonstrations of AI moral patiency.

Unfortunately, many people argue that empathy is core to ethics. Empathy is a terrible metric of moral patiency; it is extended more to those more like us (Bloom, 2017). Also, people are moved to self-deception by their fear of mortality and desire for powerful progeny and partners. There are many proposals to extend the mechanisms that sadly often fail to protect humans to protect robots or AI (Gunkel, 2018). We share the goal of not wanting any entity to suffer unnecessarily, but we take this to imply we should design AI so that it will not suffer, and further to ensure that damage to AI would not incur human suffering. Again, it is a design decision whether we make AI that is robust, can be backed up and thus protected by standard means for protecting and preserving digital data. This is the only ethical decision for AI that anyone cares about, and eliminates the necessity of the sorts of protections extended to unique human lives.

Another problem with mistakenly thinking AI is human-like is believing that human punishments such as social shunning, fines, prison, and the other tools of human law could be extend to it. Again, if we accept the List and Pettit (2011) definition of corporations as AI, we can already see that where the humans who make the decisions are not the humans who will be held to account, corruption follows. If we make artefacts to be legal persons, those artefacts will be used like a shell company, to evade justice and corrupt economies and power structures (Bryson et al., 2017), leaving ordinary citizens disempowered with less protection from powerful institutions (Elish, 2016).

The second special problem of AI is not actually unique to it but rather a characteristic of Information Communication Technology (ICT) more generally. ICT

thanks to the Internet and other networking systems operates transnationally, and therefore affords the accumulation of great wealth and power, while simultaneously evading the jurisdiction of any particular nation. This means that appropriate regulation of AI requires transnational cooperation. Again, the process to establish transnational agreements, treaties, and enforcement mechanisms is nontrivial, but already known, and already under way.

Conclusion

In conclusion, societies both can and should maintain control over artificial intelligence. Fortunately, significant progress is being made in achieving this goal—progress made by technology companies, regulatory bodies, governments, professional organisations, and individual citizens including software developers who are taking the time to understand the social consequences of technology. We welcome the opportunity to describe these efforts here, and encourage our readers to join the perpetually ongoing project of creating a richer, fairer, and more just society in which we may all flourish with dignity.

Acknowledgements We would like to acknowledge our collaborators, particularly Rob Wortham for his work on architecture and accountability, Swen Gaudl for his work on the architecture, and Alan Winfield, Karine Perset, and Karen Croxson for their work on regulation and accountability. Thanks also to the helpful reviewers for this volume. We thank AXA Research Fund for part-funding Bryson and EPSRC grant [EP/L016540/1] for funding Theodorou. Some of the description of ABOD3 previously appeared in Theodorou (2017).

References

Bloom, P. (2017). *Against empathy: The case for rational compassion*. Random House.

Blumberg, B. M. (1996). *Old Tricks, New Dogs: Ethology and Interactive Creatures*. PhD thesis, MIT. Media Laboratory, Learning and Common Sense Section.

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22(2):71–85.

Brom, C., Gemrot, J., Bída, M., Burkert, O., Partington, S. J., and Bryson, J. J. (2006). POSH tools for game agent development by students and non-programmers. In Mehdi, Q., Mtenzi, F., Duggan, B., and McAtamney, H., editors, *The Ninth International Computer Games Conference: AI, Mobile, Educational and Serious Games*, pages 126–133. University of Wolverhampton.

Brooks, R. A. (1991). New Approaches to Robotics. *Science*, 253(5025):1227–1232.

Bryson, J. J. (2000). Cross-paradigm analysis of autonomous agent architecture. *Journal of Experimental and Theoretical Artificial Intelligence*, 12(2):165–190.

Bryson, J. J. (2001). *Intelligence by Design: Principles of Modularity and Coordination for Engineering Complex Adaptive Agents*. PhD thesis, MIT, Department of EECS, Cambridge, MA. AI Technical Report 2001-003.

Bryson, J. J. (2003). Action selection and individuation in agent based modelling. In Sallach, D. L. and Macal, C., editors, *Proceedings of Agent 2003: Challenges in Social Simulation*, pages 317–330, Argonne, IL. Argonne National Laboratory.

Bryson, J. J. (2018). Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics and Information Technology*, 20(1):15–26.

Bryson, J. J., Caulfield, T., and Drugowitsch, J. (2005). Integrating life-like action selection into cycle-based agent simulation environments. *Proceedings of Agent*, pages 1–14.

Bryson, J. J., Diamantis, M. E., and Grant, T. D. (2017). Of, for, and by the people: the legal lacuna of synthetic persons. *Artificial Intelligence and Law*, 25(3):273–291.

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 Tanguy, E. A. R. (2010). Simplifying the design of human-like behaviour: Emotions as durative dynamic state for action selection. *International Journal of Synthetic Emotions*, 1(1):30–50.

Bryson, J. J. and Winfield, A. F. T. (2017). Standardizing ethical design for artificial intelligence and autonomous systems. *Computer*, 50(5):116–119.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Chapman, D. (1987). Planning for conjunctive goals. *Artificial Intelligence*, 32:333–378.

Cockburn, A. (2006). *Agile Software Development: The Cooperative Game*. Addison-Wesley Professional, Boston, second edition.

Doll, N., Vetter, P., and Tauber, A. (2015). Wen soll das autonome auto lieber überfahren? (whom should the autonomous car drive over?). *Welt*.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., and Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human Computer Studies*, 58(6):697–718.

Elish, M. (2016). Moral crumple zones: Cautionary tales in human–robot interaction. In *WeRobot 2016*. working paper.

Gallistel, C., Brown, A. L., Carey, S., Gelman, R., and Keil, F. C. (1991). Lessons from animal learning for the study of cognitive development. In Carey, S. and Gelman, R., editors, *The Epigenesis of Mind*, pages 3–36. Lawrence Erlbaum, Hillsdale, NJ.

Gaudl, S., Davies, S., and Bryson, J. J. (2013). Behaviour oriented design for real-time-strategy games: An approach on iterative development for STARCRAFT AI. *Foundations of Digital Games Conference*, pages 198–205.

Gaudl, S. E. and Bryson, J. J. (2018). The extended ramp model: A biomimetic model of behaviour arbitration for lightweight cognitive architectures. *Cognitive Systems Research*, 50:1–9.

Gunkel, D. J. (2018). *Robot Rights*. MIT Press, Cambridge, MA.

Gürses, S. and van Hoboken, J. (2018). Privacy after the agile turn. In Selinger, E., Polonetsky, J., and Tene, O., editors, *The Cambridge Handbook of Consumer Privacy*, pages 579–601. Cambridge University Press.

House of Lords (2018). AI in the UK: ready, willing and able? Technical Report HL 2017–2019 (100), United Kingdom.

Isla, D. (2015). Handling Complexity in the Halo 2 AI. In *GDC 2005*.

Johnson, S. (2017). Celebrating the familiar. claim made in books and on web pages, am trying to track down data. http://www.groundsforsculpture.org/Seward-Johnson-The-Retrospective/Celebrating-the-Familiar.

Kalofonos, D. N., Antoniou, Z., Reynolds, F. D., Van-Kleek, M., Strauss, J., and Wisner, P. (2008). Mynet: A platform for secure p2p personal and social networking services. In *Pervasive Computing and Communications, 2008. PerCom 2008. Sixth Annual IEEE International Conference on*, pages 135–146. IEEE.

List, C. and Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press.

Papaj, D. R. and Lewis, A. C. (2012). *Insect Learning: Ecology and Evolutinary Perspectives*. Springer Science & Business Media.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Rabin, S. (2014). *Game AI PRO 1*.

Romanes, G. J. (1883). *Animal intelligence*. D. Appleton.

Scranton, R. (2015). *Learning to Die in the Anthropocene: Reflections on the End of a Civilization*. City Lights Publishers, San Francisco, CA, USA.

Skinner, B. (1935). The generic nature of the concepts of stimulus and response. *The Journal of General Psychology*, 12(1):40–65.

Taylor, D. J. and Bryson, J. J. (2014). Replicators, lineages, and interactors. *Behavioral and Brain Sciences*, 37(3):276–277.

Theodorou, A. (2017). A graphical visualisation and real-time debugging tool for BOD agents. In *CEUR Workshop Proceedings*, volume 1855, pages 60–61.

Theodorou, A., Bandt-Law, B., and Bryson, J. (2019). Game technology as an intervention for public understanding of social investment. in preperation.

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.

Williams, H. (1991). *Autogeddon*. Jonathan Cape Ltd., London.

Wortham, R. H., Gaudl, S. E., and Bryson, J. J. (2018). Instinct: A biologically inspired reactive planner for intelligent embedded systems. *Cognitive Systems Research*.

Wortham, R. H., Theodorou, A., and Bryson, J. J. (2017a). Improving robot transparency: Real-time visualisation of robot AI substantially improves understanding in naive observers. In *26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Lisbon. IEEE.

Wortham, R. H., Theodorou, A., and Bryson, J. J. (2017b). Robot transparency: Improving understanding of intelligent behaviour for designers and users. In Gao, Y., Fallah, S., Jin, Y., and Lekakou, C., editors, *Towards Autonomous Robotic Systems*, pages 274–289, Cham. Springer International Publishing.