

Robot Souls?

AI and Imago Dei

Eve Poole draws our attention to the similarities between ameliorators to our behaviour implanted by God, and those that need to be built into Artificial Intelligence. God's controls on us humans would be called 'junk code' in an AI program – surplus to requirements and possibly flaky. She lists seven human 'junk codes' which could bring risk-mitigation to AI and suggests that the Church may be well placed to provide them for programmers and bring a conscientious soul into AI.

Most businesses have had AI on their risk registers for ages, mainly in the context of cyber attacks, industry disruption, and workforce planning; although rarely in terms of existential risk. But recently a spate of experts has come forward to call for a global pause on AI to allow regulation to catch up; and in 2023 the UK Government National Risk Register named AI for the first time as a 'chronic risk.' Have we lost control of AI already? Interestingly,

it seems that the best place to look for advice on this 'Control Problem' or how to stop AI going rogue, is to look at our own design: while it's true that we're a wayward species, we've still managed to survive until this very day, because our own configuration – on average over time – stops our species going off the rails and wiping itself out. And it does this in a highly sophisticated way. But in our zealous attempts only to programme the very best of our

capacities into AI, and amidst a strong cultural commitment to scientific rationalism, we left out all the human bits we were unsure about, or that seemed a bit flaky: we scrupulously avoided all the bits that looked like 'junk' code. But what if these bad bits were actually the good bits? What if these too-human bits were more useful than they appear?

This is where we come in, because as





Christians we have something quite particular to contribute to this debate. In our tradition - as in many religious traditions - we believe that we were made by God; and unless we believe in a trickster God, it would be reasonable to assume that we must therefore be perfectly designed, by a perfect God, for God's perfect ends. In the Christian tradition there's something else: we are made in God's image. Fearfully and wonderfully made. So much so, that when God decided to visit this planet, he used this design, human design, and incarnated as Jesus in order to do so. But we are so used to human failure that we've managed to blame our failure on our design, as though it's design flaws that make us prone to error. But why would God deliberately design in flaws? Might even these so-called flaws be purposeful?

Let's take a look at that cutting room floor. I have identified seven key items of human 'junk code'.

Our Junk Code

1. Free-will
2. Emotions
3. Sixth Sense
4. Uncertainty
5. Mistakes
6. Meaning
7. Storytelling

Let's go through these in turn. The first one, if you think about it, is a disastrous design choice. Free Will?! Letting creatures do whatever they want is highly likely to lead to their rapid extinction. So let's design in some ameliorators; beginning with emotion. Humans are a very vulnerable species,

because their young take nine months to gestate and are largely helpless for their first few years. Emotion is a good design choice because it makes these creatures bond with their children and in their communities to protect the vulnerable. Next, you design in a Sixth Sense, so that when there is no clear data to inform a decision, humans can use their intuition to seek wisdom from the collective unconscious, which helps de-risk decision-making. Then we need to consolidate this by designing in uncertainty. A capacity to cope with ambiguity will stop them rushing into precipitous decision-making, and make them seek others out for wise counsel. And if they do make mistakes? Well, they will learn from them. And mistakes that make them feel bad will develop in them a healthy conscience, which will steer them away from repeated harms

in future. Now that we have corrected their design to promote survival, what motivators are needed for their future flourishing? They need to want to get out of bed on a dark day, so we fit them with a capacity for meaning-making, because a species that can discern or create meaning in the world will find reasons to keep living in the face of adversity. And to keep the species going over generations? We design in a super-power about storytelling. Stories allow communities to transmit their core values and purpose down the generations in a highly sticky way. Stories last for centuries, future-proofing the species through the learned wisdom of our ancestors, and the human species prevails.

We did not choose to design humanity into AI, because it seemed too messy. A robot that was uncertain and made mistakes would soon be sent back to the shop. Or would it? In fact, our design is so clever that AI has smuggled some junk code back in, because it's so useful. Take for example uncertainty. Suppose you've programmed a computer to categorise cancer scans, but some images are blurry. What you don't want is the computer to force a Yes/No category on a faulty image, because it could mean the difference between life and death. So if it has no categories other than Yes/No, you need to add in a measure of doubt. A discipline called Bayesian AI uses the wisdom of crowds to task all the artificial neurons involved with categorising the image, so that the consensus image has a % probability attached to it depending on the degree of agreement. Variable agreement triggers human intervention to check the image so that it's not wrongly classified. Another redeemed design-flaw is mistake-making. In AI, Reinforcement Learning is used in programming so that algorithms can improve through trial and error, in the same way that we learn from our mistakes. But this is not

yet serving the dual purpose it serves in humans, which is to develop a moral conscience too.

On reflection, we can see that our junk code is not an accident or a mistake, but part of a rather clever defensive design. If this code is how we try to solve our own 'control' problem as a species, might we find wisdom in it for solving those problems for AI? For example, one of the challenges of the day is how best to train up AI. At the moment the development of AI isn't

'At the moment the development of AI isn't informed by what the wisdom traditions know about the formation of moral character'

informed by what the wisdom traditions know about the formation of moral character. Having parented children for on average far longer than pretty much any other species ever, we've collectively got some excellent parenting smarts. When our children are too young to understand, we tend to guide them through brusque and often negative commands, designed to keep them safe: No! Naughty! Stop! Hot! In formal ethics, we see in this the tell-tale signs of a deontological or rules-based ethic: Thou Shalt Not. As our young children grow aware of consequences, we start the regime of mild threat: Santa won't come! If you don't eat your peas, you won't get any pudding! No pocket money if you don't tidy your room! The most famous formal articulation of this kind of ethic is Utilitarianism, which is about optimising outcomes or 'the greatest good for the greatest number.' But as soon as we lose our kids to nursery or school, we know they'll encounter all kinds of novel situations, and we won't be there to advise them. So we focus on character, or what tends to be known in the trade as virtue ethics, in the hope that good kids make good decisions.

This is akin to the journey that AI is on. The complexity of its programming has already forced a move from simple rules to the calculation of best outcome, but we're not yet throwing ourselves into

the development of its moral character. This renders the current goal of AI as quite unashamedly the development of a master-race of psychopaths, because any attempt to program in conscience is being deliberately avoided as unnecessary, undesirable, or impossible. Instead, AI favours a default ethic that isn't sophisticated enough to deal with the complexities of life. And this might prove problematic as more and more of our shared life is outsourced to AI in order to optimise tax spend, because utilitarianism is the settled ethic of most western democracies. It's popular as a public ethic because it's so transparent: everyone can see the outcomes and judge them, so it's perfect for accountability. Its unquestioned popularity also makes it an obvious choice for the default programming of AI, given that it's already the ethical default in capitalism too. Indeed, it's not even seen as an 'ethic' – it's just obvious, in the way that a 'business case' is considered a no-brainer even if this is also a classic piece of specifically utilitarian thinking, that ends justify means. However, we've recently had a collective experience of where this kind of ethic falls dramatically short.

Do you remember when it first dawned on us at the start of the coronavirus pandemic that the UK government was pursuing a herd immunity strategy? A strategy that would mean knowingly sacrificing the elderly, the disabled and the weak, in order to save the majority of the population? In utilitarian terms this makes complete sense and would save a lot of money. But as humans we hold on to the idea that we are somehow special and precious, and that even those who are not 'useful' to society deserve dignity and respect. So we were disgusted by this, and there was justifiable public outrage at the very thought. This gut feel we have in our programming is also why we continue to resist eugenics and cloning, and to police embryology and medical policy. But as soon as you try to articulate what it is about humans that merits this special treatment you enter quicksand. Unless you happen to believe in God



'The Church is an institution of meaning-making par excellence'

Image: CofE

and feel that humans have a particular vocation, you can only really argue that we're special because we're the species currently in charge, and we write the rules. We're not 'the best' species on any objective measure, unless you factor in our design by God in his own image. So to believe in humans, you have to believe in God; and I suspect it won't be too long before that penny drops in the minds of more of our secular friends.

Meanwhile, as Christians we ought to weigh in on the moral development of AI, because it's the Church's particular core-competence. While individual parents up and down the centuries have laboured hard to develop moral character in their own children, institutionally it's the faiths who've held that charge at the level of community and for society as a whole. So much so that this role is embedded in the very nature of how this charge is described when it's given out to the leaders of the Christian faith community. For instance, in the Church of England, when the bishop hands a charge to a new incumbent, they say: 'receive this cure of souls, which is yours and mine.' So in the Church, everyone involved in ministry is curing souls. Cure is one of those words that can just mean care, or it can have the slightly more pointed moral sense, of trying to improve souls too, by making them better. And in this

sense the rituals of the Church line up behind this core objective, to restore souls to God, and old-fashioned words like formation are used to describe the process and practices by which a person progresses on this spiritual journey back towards God.

But what is this soul that we're setting about curing? Well, no-one honestly knows the answer to that question, but many over the years have had a stab at defining it. These days it's often conflated with consciousness, and held just to be religious jargon for it; although of course the concept itself pre-dates Christianity, and like consciousness it remains safely undefined. Is it a Form or a Substance; does it transmit or transpond; is it a Bird or a Plane? In 1907, one enterprising researcher famously tried to weigh the soul, with Duncan MacDougall reporting that the difference post-mortem amounted to 21 grams. But I'm quite keen on principle that we avoid defining the soul, for reasons that will become apparent, although one thing I'm absolutely clear about is that the junk code I've already mentioned betrays the existence of a soul. They are evidence of one. And these hallmarks of soul are the very things that the Church seeks to nurture and to rightly align, when it's engaged in the business of soul-curing.

Specifically, in terms of those individual items of junk code, the Church is an institution of meaning-making par excellence, and is the guardian of the stories that we regard as definitive for humanity. We tell them back to our community, day in day out, through a liturgy structured to keep the full range of them at the forefront of our collective worship, using preaching and exegesis to keep their meaning fresh in every generation, and we've been doing this for over two thousand years. The Church also takes our emotions very seriously, with formal and public rites for all the peak emotional moments of our lives: births, marriages, and deaths; with music, art and architecture to tool our emotions throughout the liturgical year. The Church teaches us to check our Sixth Sense or intuition, through prayer and the seeking of wise counsel, to make sure we're not misled by false instincts, but it's careful to acknowledge that intuitions are real, and could even be the voice of God. The Church gets an A* for its sterling work on helping us with our mistakes, too. We start every service apologising for them and promising to do better. Routines of repentance take our remorse and mobilise it towards therapeutic reparation, all within a narrative that promises us that we are ultimately forgiven for all our mistakes through the death of Christ on the Cross. And this

provides the backdrop for the Church's teaching on free will as a precious gift that must be carefully nurtured, away from sin and towards the exercise of our status as the redeemed people of God. And uncertainty? In Christianity we have four Gospels that cheerfully differ from one another; as well as a Saviour who is both God and Man and part of a rather mysterious Holy Trinity that is indivisible – and frankly completely perplexing as a concept. Faith is after all the daily schooling of uncertainty to render it purposeful. So whether or not you subscribe to the particular brand of morality that the Church offers, it is expert at the cure of souls, such that anyone attempting to copy humans could find no better authority to advise on their formation.

In 1939 in St Andrews, JRR Tolkien delivered a lecture about fairy stories. He talked about their therapeutic use in helping readers to reflect more deeply on their own world by being exposed to an internally consistent and rational fantasy world. In that world, they can see resolution, and experience the consolation of a happy ending. Tolkien coins the phrase 'eucatastrophe' to describe the 'piercing glimpse of joy' we get at 'the joyous turn' when the hero avoids peril and the story resolves. Stories, and particularly the fantasy worlds of Science Fiction, are how we've hitherto been exposed to robots and AI; indeed it's rare to have such a treasury of imagined futures to guide our thinking. But of course as narrative they have to follow the logic of a story arc, entertaining us with heroes and villains and with jeopardy and threat, before delivering us a final victory. In our current conversation about AI we seem to be in two camps: AI is a Comedy that will resolve in our favour, or AI is a Tragedy that will lead to our demise. Tolkien reminds us that the choice is


ours. Our design gives us the free will to intervene and decide; to bring about that moment of eucatastrophe. But to do so, we need to enter the story, and to make some big decisions about the characters involved. Who do we need to be, and who do we need 'them' to be? Can we risk making them better than us?

Right now, we need people to train up AI. We know that tools like ChatGPT behave rather like your average intern: they are extraordinarily keen, but you really do need to check their work. Are your staff ready to supervise them well? As in coaching, the trick to using AI is about getting ever more excellent at prompts and questions, and thinking ahead about interesting assignments that will train them up. It's about honing your instincts so that you're not taken in by plausible answers. It's about thinking of the wider impacts of using AI on those around you, and it's about sharing advice and best practice so that we all get better at it. In short, it's all about nurturing your own junk code to make up for AI's current lack of it. Here is an example.

Your organisation is unique. It's unique because of its very particular mixture of history, culture and leadership, even within its designated sector or geography. And your human staff know this. They pick it up through the stories they are told on the day they arrive and at every office party or away day they have been on ever since. Successes and failures; heroes and villains; who makes it and who doesn't. Staff antennae are acutely attuned to pick up cultural tells,

and they can tell you, instinctively, what will work and what won't. This is too nuanced ever to be fed into an AI, because it changes as the people around you change. It changes with the weather, the news, the seasons. Your people can ask an AI to write a training plan, a website, a strategy presentation: it will do it in seconds. But only your staff know how to tweak this into something that will land in your context. In the past we have frowned on water-cooler chit-chat as wasting valuable work time. But maybe these kinds of interactions are precisely where your staff must go, in order to fine-tune their intuitions about

your culture through exactly this sort of office gossiping. It's another thing that we're in danger of losing as more people work remotely.

But what about the longer term? There will be a transition in many workplaces to a more AI-enabled workforce, and that will mean re-training, re-skilling or exit for some. But before you ask ChatGPT to write your 10-year plan for you, have a look at that list of junk code. Where is it already acting as risk-mitigation in your organisation, even if sometimes it feels like whimsy or waywardness? What more could you do to nurture it? Because if we only prioritise the competencies we've already programmed into AI, there will be no reason to keep humans in the workplaces of the future. And if we lose these competencies, we stand to lose our humanity and our very souls too. 

'Right now, we need people to train up AI. We know that tools like ChatGPT behave rather like your average intern: they are extraordinarily keen, but you really do need to check their work.'



Eve Poole is the author of Robot Souls and a previous Chair of Faith in Business. Following earlier careers at Deloitte and Ashridge Business School, she has written several other books including Leadersmithing and Buying God, and is currently a Lay Canon of York Minster.