Soul Machines Ethics Policy

AI technology is advancing rapidly, and is finding a wide variety of new applications in our society, both in industry and in public services. The technology has the potential to change society in many ways (see e.g. Nature editorial, 2016). Soul Machines is at the forefront of advances in AI, and we are very conscious of our responsibility to steer the technology forward in a good direction and be part of the group of companies establishing standards and guidelines.

This is a document that represents our thinking about how this should be done, and what our role is. In Section 1, we set out the general ethical principles we will adhere to. In Section 2, we outline mechanisms within the company that ensure these principles are actively disseminated and discussed. The remaining sections focus on particular principles, in a variety of ways. In some cases we provide additional detail, to emphasise our own approach to certain ethical issues. In other cases, we discuss more practically how principles are applied in our current products and research, or our aspirations about how they will apply to products in development. These latter sections are partly intended to communicate our current ethical policies to the public. But they also function as an ongoing internal review and audit: we envisage they will be updated as AI technology evolves, and as our discussions progress.

1. General principles

Many sets of AI ethics principles have been proposed. A consensus is emerging in the industry around the principles set out in the IEEE's 'Ethically Aligned Design' document (IEEE, 2017), which are as follows.[1]

1. **Human Rights** A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights.
2. **Well-being** A/IS creators shall adopt increased human well-being as a primary success criterion for development.
3. **Data Agency** A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity.
4. **Effectiveness** A/IS creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS.

[1] 'AI/S' stands for 'autonomous and intelligent systems: for our purposes, this could read 'AI systems'.

5. **Transparency** The basis of a particular A/IS decision should always be discoverable.
6. **Accountability** A/IS shall be created and operated to provide an unambiguous rationale for all decisions made.
7. **Awareness of Misuse** A/IS creators shall guard against all potential misuses and risks of A/IS in operation.
8. **Competence** A/IS creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.

We subscribe to these general principles. We discuss Principle 2 (Well-being) in Section 3, Principle 3 (Data Agency) in Section 5, and Principle 5 (Transparency) in Section 4.3. The rest of the document is structured around issues that cross-classify with the IEEE principles.

2. How the company engages with ethical issues

It's important for everyone in the company to keep ethical issues at the forefront of their minds. We have several mechanisms in place to ensure this.

Firstly, we encourage discussion within the company about ethical issues. We include ethical issues in our schedule of 'demo day' presentations to staff. We have an 'ethics' Slack channel, which all employees can access and post to. Each new employee receives a copy of this ethics policy. Ethical concerns aren't a taboo: any employee should feel free to raise such concerns whenever they arise.

Secondly, we actively participate in several groups beyond the company where these questions are addressed.

- We are a member of the AI Forum, a New Zealand-based group bringing together people from industry, policymakers, lawyers and other stakeholders to discuss the opportunities and challenges of AI technologies for New Zealand. Our co-Founder and CBO, Greg Cross, has served on the board of the AI Forum. The AI Forum is, in turn, an active member of the Partnership on AI to benefit people and society, a grouping where these topics are discussed internationally.
- Mark Sagar is on the AIShield board of the Lifeboat Foundation, which was established to discuss ways of safeguarding AI.
- Alistair Knott is a New Zealand delegate for the recently launched Global Partnership on AI, on the 'Responsible AI' working group. Ali also served on the AI Forum 'Law, Ethics and Society' working group, and helped draft the Forum's principles for Trustworthy AI.
- We are also involved in university research exploring the possible impacts of AI. At the University of Otago, Alistair Knott co-founded the Centre for Artificial Intelligence and Public Policy, and coordinates the AI and Society Discussion Group, which has members drawn from across the university. Ali is also co-coordinator of a three-year research project funded by the New Zealand Law Foundation on AI and Law in New Zealand. Martin Takac teaches a course in AI Ethics, at Comenius University in Bratislava.

Most concretely, Soul Machines has an ethics committee. It is this committee that created the current ethics policy. The committee's brief is to keep the policy up-to-date, in light of developments in the company's products and research, and developments in AI technology and AI ethics more widely. The committee reports quarterly to the board of directors. The committee currently consists of Mark Sagar (chair), Chris Liu, Aftab Mathur, Greg Cross, and Alistair Knott.

3. The role of AI in increasing human well-being

Like most groups discussing the impacts of AI on society, we begin by noting its potential to increase human well-being. Here is our own summary of the benefits that AI can bring to human well-being, drawing on recent reports by academic groups (in particular AI100), industry groups (in particular the Partnership on AI) and political groups (in particular AINow).

- AI has the potential to advance science and technology in very useful directions. For instance, it has potential applications in medicine (drug discovery, surgery, medical research), in climate change (modelling current trends, evaluating possible climate-cooling measures, contributing to sustainable energy technologies) and in economics (better models of human decision-making and economic processes).
- AI has the potential to eliminate drudgery. Many human jobs are highly fulfilling, and contribute to the quality of life of the people who do them. But this is not true for all jobs. There are some jobs that are tedious, repetitive, dangerous, and unhealthy. All things being equal, it would be best if people didn't have to do these jobs.
- In many domains, AI systems can provide people with useful personal assistance: they can answer questions, offer advice, and provide positive interactions. They can even provide personal companionship, if humans are not available to do so. (Though we don't regard this as preferable to human companionship, which we consider more valuable.)
- AI technologies increase society's productivity: they allow more work to be done, so that more goods and services are produced. It is important that growth is sustainable, but within this constraint, AI technologies have a useful role in producing goods and services. In particular, they can potentially make goods and services accessible to people who currently have no access to them. For instance, AI technologies might provide medical advice or interventions to communities without access to a doctor or hospital, or educational services to communities whose schools are poorly resourced.

Of course, it is crucial that the AI systems that deliver such benefits also meet many specific requirements, and operate within clear social and political guidelines. In the remainder of the document, we will discuss these. We structure our discussion around four general issues: standards in AI (Section 4), data use (Section 5), socially responsible AI (Section 5.4) and human-centred AI (Section 7).

4. Standards for AI agents

We expect humans to maintain certain standards in their jobs, and in their interactions with other people. These standards can relate to basic human rights (for instance, 'don't exploit people', 'don't show bias towards or against particular social groups'), or to accountability principles (for instance, 'be able to explain your decisions'). As our AI agents ('Digital Humans') engage in interactions with people, we aim to give them certain humanlike standards of this kind.

4.1. Standards in dialogue management

In some cases, ethical standards for an AI system can be built explicitly into its architecture, so it has no choice but to adhere to them. We are exploring standards of this kind in our model of dialogue management, in cases where the Digital Human plays the role of a salesperson. We aim to bake various explicit principles into the dialogue management algorithm. For instance:

- Have the user's plans and goals in mind when giving the user advice, or offering products or services for sale.
- If the user indicates that they are not interested in some topic, do not discuss it with them further.

These principles should help the user *trust* our dialogue agent.

An additional principle for our dialogue system is the following:

- Don't try to fool your user that you are a human.

Again, a principle of this kind might be helpful in establishing trust with users.

Finally, interactions with our Digital Humans can sometimes lead to negative user experiences, stemming from classifier failures or user utterances not catered for in the dialogue script. Normally the problem is just user frustration, but we are mindful there could be cases where a user is upset, rather than just frustrated. We deploy a range of tools to deal with these cases. An important one is the ability to escalate to a human Customer Service agent, which we always aim to include.

4.2. Standards in neural networks

In neural network or statistical systems, principles are less easy to identify and encode. But it is certainly possible to assess machine learning algorithms (and the datasets they work on) for bias of different kinds; this is an active research area. We don't think these kinds of bias are a big issue for any of our technologies yet. But we will keep an eye on the issue. We are particularly concerned that our algorithms for detecting faces and facial expressions work equally well on users of all races and both genders, which is an issue for many existing face recognition systems (see e.g. Klare et al., 2012). Our staffing policy (see Section 6.4) should help us keep on top of this issue.

4.3. Transparency

In various ways, it is important that the operation of AI systems is transparent. We discussed one form of transparency for dialogue agents in Section 4.1. But transparency has other possible meanings. One common meaning is that a system's decisions should be 'explainable' (that is, understandable to a human), and we should know what variables are considered in making these decisions. In our dialogue system, a principle we might consider adding is:

- The customer should be allowed to ask 'Why?' questions to the system, at appropriate places.

Transparency also relates to education: we should make sure that people growing up in a world of AI understand the technologies that are shaping their world (more of this in Section **Error! Reference source not found.**).

4.4. Maintaining standards in AI

Looking beyond our own company, it's important to consider how ethical standards for AI systems can be maintained. There are several possible models here. We might envisage that AI companies are responsible for making their own ethical standards public, with market forces rewarding those who do. But it might also be useful if there were a national or international auditing body for AI, that could certify AI system standards. This auditing process could take different forms. It could be voluntary, geared around award of some industry-standard 'ethical tick', as is done in other industries to showcase environmental friendliness, Fair Trade and so on. Or it could be legislated for. Different mechanisms are probably suitable for different principles.

For our own systems, we think unilateral communication of principles adhered to in dialogue management is the right level of scrutiny for the moment.

5. Data agency

Our Digital Human can monitor the user's gestures, facial expressions and speech inflections at all points during the dialogue, just like a human conversational partner would. This information provides valuable indications about the direction the dialogue should go in, that the system should be responsive to—and this is why we gather it.

Having said that, data about users' emotional responses to different topics is of considerable commercial value in its own right. Many companies make money by extracting information about people obtained from the public domain, or from proprietary systems, and then sell this to interested parties, so they can target products more precisely. Originally, the information that was 'mined' concerned people's online activities—what products they bought, what sites they visited. Now, the information is becoming increasingly personal: for instance there are companies that build accurate measures of

personality types from Facebook pages. Our methods of analysing facial expressions in dialogues give us particularly fine-grained personal data, about users' subtle (perhaps even unconscious) emotional responses to the topics of conversation. This data would be very valuable for product targeting, and is of commercial value in its own right.

We certainly aim to use the personal data extracted during a given conversation with a user, *later in that same conversation*. It's important to do this, to make the conversation responsive to the user, and satisfying to the user: that's why we do it. We also envisage using the data extracted in one conversation with some user in subsequent interactions *with that same user*. Again, this is important for creating a humanlike dialogue system: the system must be able to remember its previous interactions with the user, just like a person would. So we will also store a file of 'factual' personal data for each user.

However, this stored data will only be used for very specific purposes, as specified below.

1. We may use anonymised personal data gathered when users participate in conversations to help train our systems that classify user facial expressions, user emotions, and the intents behind user utterances. Our clients will be able to use these systems, as they are embedded in our dialogue products. But they will not be able to access the raw data from users, except in very specific scenarios, which we'll discuss in Section 5.2.
2. Personal data files will not be shared directly between our clients, even after anonymisation, except in the exceptional scenarios discussed in 5.2. That is, if we gather personal data on some user *U* in a dialogue system built for client *A*, this will only be usable in subsequent interactions with *U* on applications owned by *A*.
3. We will keep relevant personal data files on our servers: this way they are usable by our clients (as specified above), but not visible to clients.
4. If we create a dialogue system for client company *A*, we will share anonymised data from users of this system with company *A*, so that they can do their own analyses on it (case studies, summary statistics, and so on). But we won't share this data with other client companies.
5. If a user asks for her personal data to be 'forgotten', we will comply with this request. The forgotten data may continue to contribute marginally to systems trained on it—until they are retrained. The effect of forgetting will be most visible for users' 'factual' personal data. If the system is asked to 'forget' a user's data, the dialogue agent won't retain factual information learned in earlier interactions.

Most of these data protection principles are required by law in the jurisdictions we operate in. In particular, we comply fully with GDPR legislation. But two principles go beyond legal requirements. Principle 1 (not allowing our clients to access raw data from users, only anonymised data) and Principle 4 (not providing anonymised data gathered with one company to other companies) go beyond the law.

5.1. Seeking permission to use personal data

It's ethically important that users give explicit permission for their data to be used in the ways described above. This ethical requirement is also a legal one, in all the jurisdictions we operate in. We adhere to this requirement: we explicitly require consent to access the user's microphone and web cam. We also provide users the ability to opt out of the collection of anonymized data that may be used in the training of our systems.

5.2. Scenarios where we may envisage data sharing with a client company

There may be some cases where personal information about a user gained during a dialogue could form a useful part of some database of information which is gathered for the benefit of that user, stored externally to Soul Machines. For instance, information gathered during a health-related dialogue could be usefully added to a user's medical record; information from an educational dialogue could likewise be added to a user's educational record. In cases of this kind, we may consider an arrangement where non-anonymised data is shared with a client company. But we envisage this kind of data-sharing would require special permission from the user, in addition to the regular permissions described in Section 5.1.

5.3. Our privacy and data retention policies

Note we also have a separate Privacy Policy and Data Retention Policy. These are distinct from the Data Agency policy described in this section: they serve different purposes (mostly legal ones). Although the policies have different emphases, they don't conflict with each other.

5.4. Facial data captured from actors

Another privacy issue relates to the actors whose faces we scan to produce our Digital Humans. We naturally require these actors to grant us permission to use the data gathered during these scans to produce digital humans that physically resemble them. However, our Digital DNA technology also enables us to blend the faces of many different actors: it is quite likely that the next generation of digital humans produced by this technology won't directly correspond to any one actor, which may reduce intrusions on actors' privacy.

6. Socially responsible AI applications

Just like any technology, AI systems can have harmful effects when they are deployed. Of course, how a commercial system is used is partly a matter for the client who buys it. But in our commercial products (which are human-computer interaction systems), we work quite closely with clients in each domain, so we do have a certain amount of control over which types of application are built. We have several policies relating to applications: some quite specific to our particular kind of dialogue system; others more general.

6.1. Choice of clients

Part of social responsibility for an AI company is in its choice of clients. We want to focus on clients building applications with socially positive impacts. Our early work with the Australian National Disability Insurance Scheme (NDIS) is an example of such an application. We have also partnered with the World Health Organisation in producing an application for helping users to quit smoking. And we are currently building an application for teaching Māori, New Zealand's indigenous language. We are also proactively looking for opportunities to build systems teaching STEM subjects, systems addressing mental health issues in teenagers, and systems for philanthropic organisations.

On the other hand, there are certain types of application we positively do not want to develop, for a range of different reasons. These include applications relating to weapons manufacture, pornography, gambling, the smoking industry, and political disinformation campaigns. (This list may grow as other areas are considered.)

6.2. Our approach towards client-generated content

Even if we are cautious in our choice of clients, it is still possible that clients build Digital Humans that produce objectionable content in their dialogues with users. In allowing our clients to produce their own content, our company is in a similar position to social media providers like Youtube, Facebook and Twitter: the question arises who is responsible for this content. Social media providers have two ways of presenting themselves legally, in relation to client content: one is as a 'publisher' of content, with first-amendment rights to remove user content; the other is as a simple 'platform' for user-produced content, wherein users are responsible for their own content. Companies often present themselves as publishers for some purposes (e.g. removing user content), and as platforms for other purposes (e.g. placing responsibility for content on users). In practice, the distinction between publishers and platforms is more of a continuum than current legal definitions allow, and Internet systems that allow clients to generate their own content occupy positions in between 'publisher' and 'platform', that aren't yet well articulated in the law.

Is Soul Machines closer to a 'publisher' of client content or to a 'platform'? We argue it is closer to a platform. There are two things that make social media companies like publishers. Firstly, their sites include functionality for sharing content between users, in 'recommender systems' that populate users' content feeds. In deciding what content items each user sees, a social media site is behaving like a publisher, making individualised editorial decisions. Secondly, social media companies are the natural port of call for user complaints about content: a user who is upset about a Facebook post will naturally complain to Facebook. Soul Machines does not allow sharing of content between users, as emphasised in Section 5 - so it is emphatically not a 'publisher' of user-generated content in this sense. And Soul Machines' clients are companies, rather than individual users: it is these client companies

that produce Digital Human content. These client companies are the natural first port of call for complaints about this content.

We nonetheless feel some publisher-like responsibility for the material delivered by Soul Machines Digital Humans. We want users to be able to easily report objectional content

produced by Digital Humans - in the first instance, to the client company that produced the Digital Human, and if necessary, to us directly. And we want to be able to take down Digital Humans that produce such content, if the client company is not responsive enough. To support this:

- Legally: we have a clause in our contract forbidding clients from producing Digital Humans that present objectionable content.
- As a monitoring mechanism:
    - The standard Digital Human interface we deliver includes a feature allowing users to provide feedback about any aspect of the Digital Human, to the client company creating the content.
    - Our web page includes a contact email where users of any Soul Machines Digital Human can report objectionable content.

6.3. Cultural and gender diversity in avatar creation

AI dialogue agents are predominantly female - a fact that has worrying potential to reinforce regressive gender stereotypes. As noted in a recent UN report (UNESCO, 2019), dialogue agents typically perform service roles, so it's easy for users to see them as occupying a socially inferior position. In addition, users treat them in ways they would not treat a real person: they are often the targets of verbal innuendo or abuse. These attitudes are particularly pronounced for female avatars, because they pick up on gender prejudices that already exist in society. And they are likely to exacerbate such prejudices. To combat this, there should be greater gender diversity in AI dialogue avatars - and also, ideally, greater diversity in the social roles avatars play.

AI dialogue avatars are also predominantly caucasian. Since white ethnicity is associated with high social status, there is a different dynamic at play here: for some reason, the gender of AI agents patterns with service role stereotypes, while the race of AI agents patterns with high status stereotypes. In any case, the homogeneity of avatars in ethnicity is unfortunate, just as it is for gender. The diversity of avatars should reflect the diversity of the people who use them. This is not just a point of principle: there is evidence that in some domains, there is great benefit in providing an avatar that mirrors the user's ethnicity. (See Liao and He, 2020 for a recent example in psychotherapeutic counselling.) As a company, we want to support and encourage the use of digital humans of all races and genders, and the development of digital humans that don't perpetuate harmful stereotypes.

Digital DNA

Technically, our main tool for achieving age, gender and ethnic diversity is our Digital DNA technology. This technology has the explicit goal of modelling the facial physiognomy of humans from all of the world's ethnic groups, and thus enabling clients to create digital humans with any ethnicity. Digital DNA techniques also allow blending the faces of several human models. This allows the creation of Digital Humans that combine elements from male and female models - and offers the potential for a non-binary treatment of gender in

digital humans. While these are both still aspirational goals, we believe the Digital DNA project pushes usefully towards ensuring representation of all groups in the AI sphere.

Digital Heroes

We are also committed to building digital people occupying social roles other than assistants. A particular focus for us is 'digital heroes': digital twins of celebrities or other high-profile people. To move towards social equity in digital humans, we have a particular interest in creating digital heroes from under-represented cultural or gender groups. A flagship project in this regard is our digital version of the musician will.i.am, which was featured in Robert Downey Jr's Age of AI YouTube series.

Creating cultural- and gender-appropriate content for Digital People

To move towards digital people that don't perpetuate harmful stereotypes, we must also keep an eye on the dialogue content created for Digital Humans. Most dialogue content for our Digital Humans is created by our client companies, as we discussed in Section 6.2. However, we do create some Digital Humans ourselves. We undertake to ensure that content authors for a Digital Human with a given gender and ethnicity include people from that same gender and ethnicity. This issue connects with our staffing policy, as discussed in Section 6.4.

We do deliver certain domain-general dialogue modules as part of our product, relating to error recovery, and casual chat (which includes responses to offensive utterances). The UN report (UNESCO, 2019) notes that avatars often produce problematic responses in these cases: for instance, Siri's by-now notorious 'I'd blush if I could'. It's very difficult to craft good stock responses, but we endeavour to work hard to make sure our Digital Human responses in these areas push productively against age, gender and racial stereotypes.

6.4. Soul Machines' staffing policy

We need staff from a representative range of ethnic, cultural and gender groups. This is partly to ensure due attention to issues of bias in the AI classifiers we build (see Section 4.2) and partly to ensure legitimacy of the Digital Humans we create (see Section 6.3). We endeavour to hire staff who are representative of the intended users of our systems.

6.5. AI systems and human jobs

AI applications are often able to replace jobs currently done by humans—or at least some components of these jobs. Our systems are no exception: for instance, many of the Digital Humans we build perform the same kind of roles as workers in a call centre.

Any new technology replaces existing human jobs. Historically, new technologies have also created a whole set of new jobs—so their ultimate effect is to redeploy the workforce towards new jobs.

We believe the currently available AI technologies mostly have this effect of creating new jobs in new areas, and redeploying the workforce. Redeployment is away from low-skill jobs towards jobs in higher-tech areas. The best way to respond to the job threats posed by current AI technologies is to provide training and retraining programmes in ICT—particularly in Computer Science and AI. To push in this direction, we have various outreach programmes with schools, and an active programme of schools visits. We also support the New Zealand ICT Graduate School's SHIFT programme, which encourages graduates to retrain in Computer Science.

However, we are also aware that AI might be different from other technologies. As AI systems progressively master more human abilities, it may be that they start to take human jobs without creating new ones. It's hard to predict whether this situation will ever arise, but we think it's important to encourage a public debate around this topic, so that we are prepared if it does eventuate.

What should happen if AI systems make substantial inroads into the human job market is ultimately a matter for politics. There are many questions that need to be addressed. For instance:

- How can the wealth generated by AI be distributed across society? We might envisage some form of robot tax levied on companies that employ AI systems to do work that could previously only be done by a human employee. This tax could then be used to fund some kind of living wage, or universal basic income.
- It may also be important to ensure that people who want to work still have that ability, even if their job can be done by an AI system. Again, a robot tax of some kind could be used to subsidise the wages of human employees, so that companies are not financially disadvantaged by retaining human employees. (We might also envisage schemes where companies retaining human employees advertise this fact, and sell their products at a premium, in the same way that companies selling organic or fair-trade products do today.)
- There are some jobs for which humans are likely to be much better suited than machines. Jobs that involve the creation and maintenance of human communities are a clear case in point. We want people to look after our children, to organise social and community activities, and to provide love and care for our elderly and disadvantaged. These jobs are currently underpaid---often, they are part of the voluntary sector, or not recognised as 'jobs' at all. Another possible role for a robot

tax is to move people into fulfilling, well-paid roles that contribute directly to the things we value most: family, extended family and community.

As a company, we want to be involved in the political discussion around how AI advances impact the jobs market. We believe that AI technologists have an important contribution to make in this discussion—especially those technologists who are building AI products.

We also believe it's important that our next generation of policy makers should have some understanding of AI technology. To this end, we are actively encouraging courses in AI tech

in Law degree programmes: see for instance the Otago courses LAWS102 ('Introduction to Law and New Technologies') and LAWS428 ('Law and Emerging Technologies').

7. Human-centred AI

There is an increasing public worry about what AI machines will look like in the future, if or when they get 'more intelligent than people'—a worry expressed by commentators like Elon Musk, Max Tegmark, Stephen Hawking, Nick Bostrom, and Stuart Russell. The consensus amongst AI researchers is that human-level 'general AI' is still some way off (see e.g. Müller and Bostrom, 2016). But nonetheless, we think it's important to plan ahead, and think about the general direction AI research should take. There are several possible kinds of AI we could produce; some kinds might be safer, or more controllable, than others. An important distinction is between AIs that resemble humans, and AIs that don't.

Most current AI research is not focussed on modelling a 'complete' human being. AI systems tend to be developed for particular domains, and solve problems in their chosen domain in ways that are quite different from humans. (This is true, for instance, for DeepMind's Go program, or Google's translation system.) Most AI systems learn directly from internet corpora in some chosen application domain (words, share prices, user purchases), without any direct experience of the real-world objects and events these corpora refer to. These AI systems develop very powerful representations—but in many ways they are alien to the representations we humans use.

Our approach at Soul Machines, on the other hand, is to build AI systems that model actual human cognition as closely as possible. We aim to build AI systems that simulate human beings, both in their physical form, and in their brain and behaviour. It's for this reason we term our systems 'Digital Humans'. Our avatars implement detailed models of human faces and bodies, right down to skeletons, muscles and skin. And their behaviour is driven by biological models of the human nervous system, and high-level models of cognitive function. A particular focus is on a biologically realistic model of emotions, which impacts on every level of behaviour: our avatars' emotions are expressed in subtle facial and behavioural signals. Another focus is on building an 'embodied' intelligence, in which cognitive processes at all levels are referred to perceptual and motor operations, that

connect the AI system to physical world (or some simulation of the physical world). A final focus is on building intelligence in the way that human babies do, by exploring the world in progressively more sophisticated ways. In the developmental trajectory we model, the AI agent first develops basic sensory and motor concepts first, and then uses these to drive progressively more sophisticated explorations of the world. We place particular emphasis on developing representations of typical human plans and intentions, which arise quite early in infant development.

Both humanlike and non-humanlike AI can plausibly be expected to advance towards human-level intelligence. It is not yet clear which variety of human-level AI will be best suited to human needs. It's possible they will both have good uses. It's also possible that no

form of human-level AI will be beneficial for humanity (see e.g. Bostrom, 2014). We recognise that if AI becomes too good, governments may have to step in to stop its deployment in certain cases, or to prevent further research. But there are at least some arguments for wanting a human-level AI technology that *resembles* humans.

Our intention in modelling infant development processes grounded in the physical world is to build a human-like intelligence, which humans can naturally interact with and relate to, so that the knowledge it gains is qualitatively the same as human knowledge. Our motivation is mostly for the short term: given that we are likely to interact more and more with AI systems, we should make these interactions as humanlike as possible. (In particular, we believe that ethical principles are best taught to an AI system using the same methods we use to teach human infants and children. And we believe a model that emulates humans is most likely to provide the kind of 'transparency' that real people do, when they explain their actions; for details, see Zerilli et al., 2019.) But we also have a tentative view on the longer term: if we are going to have AI systems that are as intelligent, or more intelligent, than humans, humanlike systems may be preferable to non-humanlike ones. But we recognise that there are also arguments in the other direction (see e.g. Markoff, 2015). In particular, we recognise that creating AI systems with their own humanlike goals may be problematic in the longer term (see Bryson, 2010; 2018). However, there are also suggestions that building AI systems with their own goals is the best way of ensuring beneficial AI in the long term, provided their ultimate goal is to help humans - see Stuart Russell's influential recent proposal (Russell, 2019). In any case, we undertake to follow these debates, and we're willing to be proven wrong on the longer-term question.

For the moment, though, our working method is to develop systems that are as humanlike and human-centred as possible. By this method, our knowledge of actual humans can serve as a yardstick for as long as possible: for the foreseeable future, we can be guided by studies of how humans develop as infants and children, and how they behave as mature adults.

8. References

N Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, UK, 2014.
Bryson, J. (2010). Robots should be slaves. In Y. Wilks (Ed.), *Close engagements with artificial companions: Key social, psychological, ethical and design issues* (pp. 63–74). Amsterdam: John Benjamins.
Bryson, J. (2018). Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics and Information Technology* 20:15-26.
Nature editorial. Anticipating AI. *Nature*, 532:415, 2016.
AI Forum. *Artificial Intelligence: Shaping a future New Zealand*. Report, 2018.
B Klare, M Burge, J Klontz, R Vorder Bruegge, and A Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6): 1789–1801, 2012.

IEEE. *Ethically Aligned Design (First Edition). A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*. IEEE (2017).
Liao, Y and He, J (2020). The Racial Mirroring Effects on Human-Agent in Psychotherapeutic Conversation. *Proceedings of IUI 2020*.
J Markoff. *Machines of Loving Grace : The Quest for Common Ground Between Humans and Robots*. Ecco Press, 2015.
V Müller and N Bostrom (2016). Future progress in Artificial Intelligence: A survey of expert opinion. In V Müller, editor, *Fundamental Issues of Artificial Intelligence*, pages 555–572. Synthese Library.
Russell, S. (2019). *Human Compatible: AI and the Problem of Control*. Allen Lane.
UNESCO (2019). *I'd blush if I could: closing gender divides in digital skills through education*. UNESCO report GEN/2019/EQUALS/1.
J Zerilli, A Knott, J Maclaurin, and C Gavaghan (2019). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy and Technology* **32**:661–683.