Ethics statement for Soul Machines (Draft 3.0)

September 2019

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.

This is a document that represents our thinking about how this should be done, and what our role is.

1 An active engagement in discussions around AI ethics

When it comes to considering the possible impacts of AI on society,there are more questions than answers. The key thing, from our perspective, is that AI companies should be involved in disussing these questions. For our own company, this means two things.

Firstly, we encourage discussion within the company about ethical issues. 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,[1] 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 CFO, 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'[2], a grouping where these topics are discussed internationally.

- We are also involved in university research exploring the possible impacts of AI. This involvement runs through Ali Knott, who works in our neural networks and dialogue groups, on a contract at the University of Otago. At Otago, Ali coordinates the AI and Society Discussion Group,[3] which has members drawn from across the university (including Computer and Information Science, Law, Philosophy, Econonics, Politics, Marketing, Statistics and Psychology), who meet regularly to discuss issues relating to AI's impacts. 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,[4]. This project focusses on the role of AI in policing and criminal justice in New Zealand, and the impact of AI on employment in New Zealand. In each case, we study applications of AI around the world to inform our research. Another researcher in the brain modelling group, Martin Takac, also teaches a course in Science,Technology and Humanity, at Comenius University in Bratislava; this course has a strong emphasis on AI ethics.

Through all these discussions, we have developed positions on three topics.

- Accountability (Section 3): how do we ensure our AI agents behave well towards their users?

[1] https://aiforum.org.nz/
[2] https://www.partnershiponai.org/
[3] https://ai-and-society.wiki.otago.ac.nz
[4] http://www.cs.otago.ac.nz/research/ai/AI-Law/

- Social responsibility (Section 4): how do we ensure our company plays a positive role in society (e.g. in relation to privacy and jobs)?
- Human-centred AI (Section 5): what kind of AI systems do we want in the longer term?

In discussing each of these topics, it is important to bear in mind that AI technologies have huge positive potential. We will begin in Section 2 by setting out the main benefits of AI techniques, as we see them.

2 The positive potential of AI

Several groupings of AI researchers have been established to explore the likely impacts of AI on society. These include academic groups (in particular AI100[5]), industry groups (in particular the 'Partnership on AI'[6]) and political groups (in particular AINow in Washington[7]). All these groups have qualified their concerns about AI by emphasising AI's potential to bring about positive social effects. To summarise these:

- 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.
- 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. We now turn to a discussion of these.

3 Accountable AI agents

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

[5] `https://ai100.stanford.edu/`
[6] `https://www.partnershiponai.org`
[7] `https://artificialintelligencenow.com/`

3.1 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 AI system 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.

We believe implementing principles like these will create a competitive advantage for our dialogue agents. For one thing, such principles facilitate user-centred dialogues, where the user has some control in determining the direction of the conversation, and should thus boost user satisfaction. In addition, if users *know* the dialogue agent adheres to such principles, we believe this will help them to *trust* the agent. Trust is a very valuable commodity. Building trust with a human salesperson is a lengthy process: it only develops gradually, since we can't directly inspect what principles he or she is using. If our dialogue agents operate with transparent sales principles, they are likely to gain user trust more rapidly than other dialogue agents. They may even have a headstart over human sales agents when it comes to establishing customer trust.

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.

3.2 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, which is an issue for many existing face recognition systems (see e.g. Klare *et al.*, 2012).

3.3 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 3.1. 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 4.3).

3.4 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. 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.

4 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 dialogue 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.

4.1 Protection of personal data

One very specific issue relates to the way our dialogue system works. The system monitors 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' the 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 selling 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.

- 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.
- Personal data files will not be shared between our clients, even after anonymisation. That is, if we gather personal data on some user U in a dialogue system built for client A, this will only be useable in subsequent interactions with U on applications owned by A.
- We will keep personal data files on our servers: this way they are usable by our clients (as specified above), but not visible to clients.

- 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.

- 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.

4.2 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 work with the Australian National Disability Insurance Scheme (NDIS) is an example of such an application.

On the other hand, there are certain types of application we positively do not want to develop. These include applications relating to weapons manufacture, pornography, gambling, political disinformation campaigns (and probably others still to be identified).

4.3 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 dialogue agents 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. This is also the view expressed by New Zealand's AI Forum in their inaugural report (AI Forum, 2018). 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[8], 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 can 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

[8] `https://signal.ac.nz/shift/`

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 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'[9]) and LAWS428 ('Law and Emerging Technologies'[10]).

5 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 clearly expressed by commentators like Elon Musk, Max Tegmark, Stephen Hawking and Nick Bostrom. 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; we think some kinds are better than others.

Our approach at Soul Machines 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. (In fact, our Digital DNA project has an explicit goal to model the facial physiognomy of humans from all of the world's ethnic groups.) 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.

We believe this approach is most likely to result in AI systems that humans can understand, and relate to. We also believe it is most likely to lead to systems that have the same kind of 'transparency' as real people do—that is, which can give humanlike explanations of their actions (see our study in Zerilli *et al.*, 2018). 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

[9] `http://www.otago.ac.nz/law/undergraduate/papers/index.html?papercode=LAWS102#2018`
[10] `http://www.otago.ac.nz/law/undergraduate/papers/index.html?papercode=LAWS428#2018`

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 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 partly 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.) But it is also partly for the longer term: we believe that, given the available alternatives, a humanlike AI has the best prospects for a future scenario where AI systems become as intelligent, or more intelligent, than humans.

Not everyone shares the idea that we should aim to build a humanlike AI. Some commentators suggest we should continue to develop 'tools', rather than emulations of whole humans (see e.g. Markoff, 2015). But as AI methods increase in power, it will be increasingly hard to think of AI systems as 'mere tools'. Other commentators argue that it's hard to predict what a humanlike AI system will look like when it is further improved (see e.g. Bostrom, 2014). Will it just be 'like us, but wiser'? Or will it be something we can't relate to at all? We agree this is hard to know. Indeed, we acknowledge that there may be a point where AI systems become so powerful their abilities have to be constrained by government. But we are far from that point at present.

In the meantime, 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.

References

N Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford, UK, 2014.

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.

J Markoff. *Machines of Loving Grace : The Quest for Common Ground Between Humans and Robots*. Ecco Press, 2015.

V Müller and N Bostrom. Future progress in artificial intelligence: A survey of expert opinion. In V Müller, editor, *Fundamental Issues of Artificial Intelligence*, pages 555–572. Synthese Library, 2016.

J Zerilli, A Knott, J Maclaurin, and C Gavaghan. Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy and Technology*, 2018. doi: 10.1007/s13347-018-0330-6.