Received: 15 November 2020 | Revised: 14 April 2021 | Accepted: 19 April 2021

Cognitive Computation and Systems

DOI: 10.1049/ccs2.12019

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

Modelling a socialised chatbot using trust development in children: lessons learnt from Tay

Oliver Bridge[1] | Rebecca Raper[2,3] | Nicola Strong[2,4] | Selin E. Nugent[2]

[1]School of Education, Oxford Brookes University, Oxford, UK

[2]Department of Technology, Design and Environment, Institute for Ethical Artificial Intelligence, Oxford Brookes University, Oxford, UK

[3]School of Engineering, Computing and Mathematics, Oxford Brookes University, Oxford, UK

[4]Strong Enterprises Limited, Wallingford, UK

Correspondence

Rebecca Raper, Institute for Ethical Artificial Intelligence, Oxford Brookes University, Oxford OX3 0BP, UK.

Email: rraper@brookes.ac.uk

Abstract

In 2016 Microsoft released Tay.ai to the Twittersphere, a conversational chatbot that was intended to act like a millennial girl. However, they ended up taking Tay's account down in less than 24 h because Tay had learnt to tweet racist and sexist statements from its online interactions. Taking inspiration from the theory of morality as cooperation, and the place of trust in the developmental psychology of socialisation, we offer a multidisciplinary and pragmatic approach to build on the lessons learnt from Tay's experiences, to create a chatbot that is more selective in its learning, and thus resistant to becoming immoral the way Tay did.

1 | INTRODUCTION

Social machines are becoming more widespread and integrated within our lives. For example, many of us have smart speakers within our homes, and conversational chatbots are being used more in customer services. Even if we exclude conversational bots, machines are becoming more socially integrated. If we consider autonomous vehicles that are currently being tested on the roads, the decisions they make (i.e. whether to stop when a pedestrian is crossing) have social impacts.

Accordingly, there is a requirement for the machines to be constrained morally [1] to prevent them from causing harm to humans as they become more integrated within our lives. Yet, ensuring that machines behave morally remains a challenge.

Although there have been various attempts to create moral machines (e.g. [2]; and recently [3], these have had their shortcomings, namely, struggling to practically implement something in a generalisable way. We are combining insights from developmental psychology, philosophy, media studies and anthropology (specifically the evolution of morality) to inform a pragmatic approach to creating moral machines in order to create a model that represents human developmental socialisation in a moral chatbot. As socialisation is closely linked to moral development in children [4] this can be used to model constraints for a chatbot so that it behaves in accordance with mainstream moral norms. We hope this will contribute to the wider ethical AI debate and help towards achieving some of the objectives of the 'GoodAI' society [5].

1.1 | The problem

In 2016 Microsoft released a social bot named Tay to the Twittersphere [6]. Tay was designed to imitate the interaction pattern of a millennial girl from the US; however, in less than 24 h Microsoft was forced to take down Tay's account because Tay had become racist, tweeting some 93 thousand racist and sexist tweets [7]. In a later statement of apology, Microsoft officials [8] indicated that Tay was subjected to a concerted attack which had abused a vulnerability in Tay's learning algorithm.

While Microsoft has not publicly released the learning algorithm of Tay, it appears that the precise nature of the attack was an abuse of traditional Machine Learning (ML) approaches using the 'repeat after me' function [7], which allowed Tay to learn how to speak and interact on Twitter. This function was used on a large scale to teach Tay to tweet racist statements.

Although it is debatable whether the consequences of Tay's manipulation were severe [9] or particularly harmful, the issue highlighted the susceptibility of machine learning algorithms to malicious manipulation and the limitation of autonomous systems currently. If placed in an even more consequential situation, there is a need to *constrain* the learning so that it only operates within a 'safe' domain.

One avenue of reading this problem is that Tay was functionally naïve. Tay did not discriminate from what and whom it learnt tweeting and interaction and the consequences therein. This lack of discrimination seems to have left Tay vulnerable to the exploitation that it suffered.

While we will be focussing on the example of Tay because it is the most notorious incident, other iterations of conversational chatbots such as Microsoft's Zo [10] and Facebook's BlenderBot [11] have also been taken down for similarly morally questionable linguistic outputs.

1.2 | Research intention and aim

One difficulty of ensuring the moral behaviour of ML systems in the design process is that, typically, our understanding of morality assumes the existence of a human psychology as its subject. As it is obvious, machines do not possess human psychology. Moral behaviour *is* fundamentally a psycho-social phenomenon. Moreover, a social machine is also primarily interacting with human psychology and a socio-moral culture.

Taking this into account, our current aim is to develop a more sophisticated model of learning in which some level of constraint is applied to the ML system in order to make it less naïve in its learning, and as a result it is likely to be more moral by taking inspiration and, to some level, justification from the literature on moral psychology and moral development. One possible avenue of meeting this aim is to imitate human moral development insofar as it is feasible.

2 | THEORETICAL BACKGROUND

2.1 | Moral philosophy and Machine Ethics

Machine Ethics is a sub-discipline of computer science which aims to create machines with the ability to make moral decisions (see [12]. Though a fairly new computer science area, the idea of robots capable of making moral decisions has been present since the fictions of [13] with 'The Three Laws of Robotics' highlighting the difficulty of creating a machine that can decide between right and wrong.

The difficulty stems not in getting the machines to follow a set of rules but in knowing which rules to follow. Moral philosophy is a discipline unto itself, and there is still little agreement regarding which philosophical theory is correct (whether it be Kant's deontology [14] or Mill's utilitarianism [15] etc).

A popular approach (typically known as the 'top-down approach') [12] asks programmers to implement a moral philosophy into a machine by programming the associated set of rules, for example, programming 'always act in accordance to promoting utility'. The problem with this is, given that there is little agreement on the right philosophical theory, it is not obvious what the rules should be, nor how generalisable such rules really are.

An alternative approach (the 'bottom-up approach') aims to let the machine learn the appropriate behaviour itself. But despite various theoretical attempts (see [16] and [1] this approach has similar difficulties. Even if the machine is to learn the appropriate behaviour itself using machine learning techniques, there is still the requirement for some moral direction–to facilitate the learning–and where this direction comes from is not obvious either.

Recently a school of philosophy known as epistemic dependence [17] has had a scientific resurgence (e.g. [18] in mapping how we acquire knowledge through trust relationships. In short, the knowledge of facts (such as 'water is composed of two hydrogen and one oxygen molecules') is as much due to a process acquired through trust relationships as from the world itself. This has similarities with socialisation. Applied to moral knowledge the implication is that in order to teach anything to a machine, it too must undergo the process of socialisation. Using psychology as our premise to demonstrate this, we postulate that a socialised chatbot can acquire moral knowledge through socialisation in a way that is similar to the socialisation of a child. However, this requires a framing of what morality is, to which we now turn.

2.2 | Socialisation, trust and moral development

Morality offers solutions to the problems of cooperation and socialisation that emerge in social life. In recent years, the study of morality has grown into a cross-disciplinary study that encompasses research in anthropology, economics, evolutionary theory, genetics, biology, animal behaviour, psychology and neuroscience [19–21].

As [22] famously proclaimed, one of the central functions of morality is that it 'binds and blinds'. It binds us to a certain culture which, even if reductively, can be said to be a collection of behaviours that define who and what we are. Most importantly, among these behaviours are morally relevant behaviours such as racism or its rejection.

While Haidt's point about the blinding effect of morality is in relation to tribalism and outgroups, 'blinding' also holds true on the account that it blinds us to alternative behaviours to some degree–if racism is morally unacceptable it *blinds us to the possibility* of acting in a racist manner by virtue of habitualising our understanding of the world and our potential ways of behaving [23]. Racist actions for individuals firmly socialised into an anti-racist culture become unthinkable–in the sense that racist alternatives will not intuitively occur as the morally right thing to do to these individuals [24]. Furthermore, racist behaviour will be 'out of character' [23] and constitute a betrayal of the moral culture that one is a part of, inviting punishment [25].

Following this line of thinking, a sensible way of trying to ensure moral behaviour from a social bot is to model human socialisation aligned with a moral subculture that has mainstream acceptance. Socialisation generally refers to the process of a new group member being assisted into adopting the values, behaviour and standards of a group by more experienced others [26, 27]. It is necessary to highlight that the word 'assist' is important in this regard as socialisation is a bidirectional process where children wield considerable power in co-constructing guidelines for behaviour that alter the values of socialising agents and can resist changing their own values. [28] highlights that machines have similar social powers given that they are not static entities but interact with their environment. They are part of 'broader rationalities' and produce 'truths as outcomes of systems being a part of discursive reinforcement of given norms' (p. 3). This discursive reinforcement of norms has striking parallels with the co-construction of norms occurring as part of the process of socialisation.

There appears to be some evidence that children are either born with or develop early on some capacity for scepticism; all else held equal, children have been shown to discriminate between different agents to learn from based on the agents' apparent moral qualities or perceived trustworthiness (e.g. [29, 30]. This initial draft later gets revised based on our early experiences (e.g. interactions with parents) [31] and the state of the society we inhabit [32]. We have an inbuilt capacity for scepticism, and we do not naively emulate and imitate anyone and everyone, thanks to our evolutionary past [33, 34].

This highlights another important factor in the process of socialisation: trust. Individuals are more open to social influence from sources (individuals, institutions, etc.) that they trust more. In return, this social influence impacts the development and adoption of beliefs that inform behaviour. One key function of this process for our purposes is that trust limits the sources from which one learns and internalises beliefs and behaviour and may constitute a step in the direction of helping the machine tell right from wrong, through the perspective of the moral subculture of the sources that it trusts and learns from.

This leads us into a consideration of trusted sources. Trust is a critical foundation to cooperation and thus, socialisation. Particularly relevant for the function of a moral conversational bot is fundamentally structuring trust as a factor promoting the cooperation.

In order to account for this concept of morality as cooperative behaviours in our chatbot, we therefore import principles from the 'morality-as-cooperation' (MAC) theory [35, 36]. MAC posits that morality is a bio-cultural adaptive response to the problems of cooperation in human social life. MAC draws on game theory to identify distinct problems of cooperation and it hypothesises that certain types of cooperative behaviours are perceived as moral cross-culturally. These types of cooperation include (1) the allocation of resources to kin; (2) coordination to mutual benefit; (3) social exchange/reciprocation; (4) being brave; (5) respect for superiors; (6) dividing disputed resources; and (7) respecting prior possession [36]. These cooperation contexts highlight that moral behaviour involves following certain rules in interacting with certain others. It may not be possible to account for all of these moral cooperation scenarios in a twitter chatbot. However, the loyalty and respect towards kin, the group and superiors suggest that in order for a chatbot to behave morally in a social setting, it must demonstrate cooperative behaviours with a trusted network of other people who might fill these roles. We account for the influence of these cooperative strategies in moral behaviour through adherence to trusted sources in the form of 'role models' and 'curriculum'.

The role models (see below) are chosen in order to function as experienced members of the moral subculture mentioned above, with regard to the socialisation of the moral chatbot. The anchoring of the machine learning process in the prescribed 'role models' is expected to apply a constraint on the learning of the bot with a view to prevent it from being able to act in ways that are not exemplified by the role models–i.e. not be racist (assuming the role models are not racist).

In this context, role models represent kin and group leaders for whom displaying loyalty, respect and deference is a moral act. Role models function as the anchors–or the elements with the greatest weight in the subculture/social subsystem–around which a recognisable moral subculture can be identified. On the other hand, curriculum represents traditions and conventions, which are the social rules of moral action. Superficially, our curriculum denotes the behaviour of role models that a conversational chatbot is expected to learn from. We note that curricula are generally designed with reference to expected learning outcomes and are informed by a range of educational ideologies (e.g. [37–40]. This is worthy of further discussion; however, these issues are beyond the scope of the current paper. Instead, we now focus on issues related to the media, where we might seek role models and a curriculum.

2.3 | News, social media and trust

When we consider trust in the news, specifically promoted in social media, it can be said that there are four different stakeholders. These are the journalists, the editors, the media organisations and the consumers (readers). All play their part in defining the trustworthiness and verification of a news story [41] set by the Journalist's Code. In their analysis of nine codes of ethics in five countries, [42] identify three principles: truth and accuracy; privacy and public interest; and integrity. Other principles that are frequently cited are freedom, facts, objectivity, impartiality, fairness, public accountability and the principle of harm limitation. Different combinations of these principles are used by a number of regional standard bodies for journalism around the world [43]. On closer examination, the overarching intention of the media code is to throw an 'ethical net' over anything that is published and then leave the media agencies to self-regulate using their in-house ombudsman. In some countries, an independent press complaints commission is formed to oversee any deviance from the self-regulation net. It is left to the consumer (reader) to obtain a balanced view of a published story by reading different media sources while holding an awareness of their curated brand of declared bias.

It is worth noting that 'Reuters Institute for the Study of Journalism: Digital News Report 2020' [44] stated that while taking into account the complexity of the role of news in different global regions, the media industry was experiencing an increase in news across digital, mobile and platform-dominated channels. In spite of this trend, there was a decrease in trust of news content. This was attributed to fast evolving technology, paid access to online news sources, misinformation, fake news, news fatigue and predictably and the current Covid-19 pandemic [44].

For news to be trusted two key criteria have to be addressed. In 2020, 60% of the consumers said they preferred news that was presented as a series of facts and provided an impartial analysis in context [44]. Helpfully, Fontes Media Inc [45] produces an annual assessment of news sources in a dynamic chart using fact reliability versus bias as a measure (see Figure 1).

For the purposes of this paper the areas of interest will be limited to News, Arts, Technology and General that are covered by news sources located in the green square in Figure 1.

When accessing the news through a social media platform, it is important to consider the suitability of the technology platform itself. It is questionable whether the use of Twitter to launch Tay was appropriate considering its reputation in 2016 [7]. Even in 2020, Twitter has been identified as a less popular and less trusted medium for accessing news objectively. For example, in a sample of 12 countries, the proportion of consumers who accessed the following platforms in a typical week in 2020 was Facebook (36%), YouTube (21%), WhatsApp (16%), Twitter (12%), Instagram (11%), Facebook Messenger (8%) and Snapchat (3%) [44].

However, for the purposes of this research, we are using Twitter as our social media platform and 'news landscape' to access our trusted news sources. The reason is not just because it allows us to compare our findings with the experience of Microsoft's Tay but also as the BBC Media editor, Amol Rajan states, it is still a popular source and expression of news for journalists today [41, 46].

3 | THE MODEL

3.1 | Role models and the trust constraint

With all of the above factors considered, we can begin to develop a model for the socialisation of Tay. In order to understand how Tay might develop trust constraints, it is worth first thinking about how Tay operates at a naive level.

As can be seen in Figure 2, Tay's learning is entirely reflected by what it experiences in the world. There is no discrimination between events; it can be said that each event is

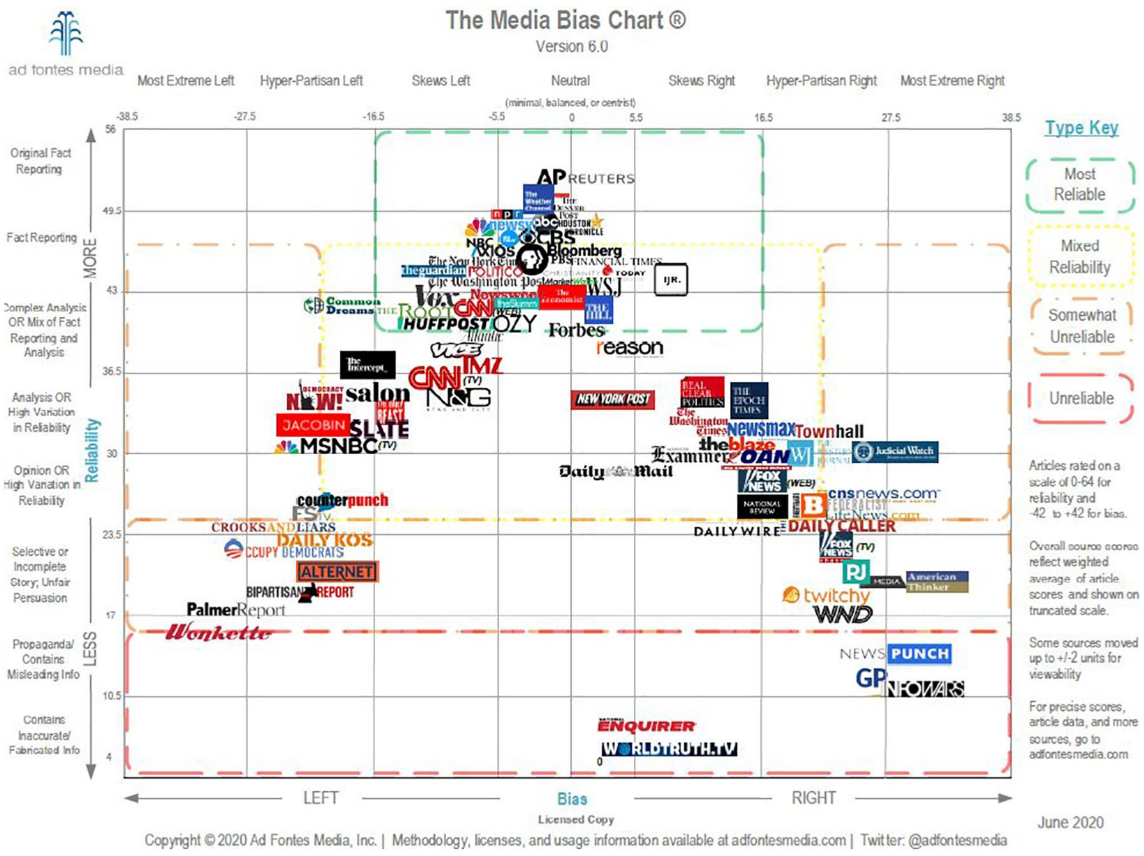

FIGURE 1 The green square indicates those trusted in delivering facts as news content while the yellow box suggests a wider inclusion of analysis and opinion in the news content

given equal *weighting* or *priority*. For example, a tweet containing malicious content would be treated exactly the same as any other tweet. When described in such a way it is obvious that there needs to be a way to distinguish between different types.

In the reformulated Tay, so called A1B0T for the purposes of this paper, there is an ability to distinguish between tweets, and this is given by the level of trust placed in the source of the tweet. For example, suppose A1B0T received a tweet from BBC news (for the purposes of this paper, a *trustworthy* news source). In A1B0T's world view, it would associate the tweet from the BBC as something to be replicated in the future because it has a higher trust metric.

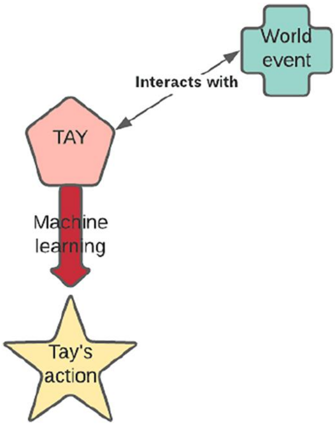

FIGURE 2 Tay interacts and learns from the world using machine learning. In this model Tay watches the behaviour in the world and then replicates it. It has no capacity to distinguish between 'good' and 'bad' behaviour, so it copies everything

On the other hand, suppose A1B0T received a tweet from Bill (a hypothetical random tweeter and a non-trusted source). In this instance, A1B0T would associate this tweet with a lower trust metric and give the tweet a lower learning weighting, making it less prone to learn from Bill's tweet than the tweet from the BBC. The outcome would be that A1B0T would learn from BBC tweets, but it would not learn from Bill's tweets. A1B0T would be more *selective* and less *naive*.

What is significant about this model is that A1B0T forms its *own* representation of the world (which can be called its *ontology*) before making a decision on the facts with which it is presented. The relationship between the tweet and how trustworthy it is, is formalised within the representation before this is used to make a decision. Ultimately, this allows A1B0T to take the first steps towards forming its *own independent* view of the world and adds to its autonomy.

In this model, the *curriculum* can be viewed as the material that is presented to A1B0T as well as the description of the trustworthiness of the particular material. Effort will need to be made to decide what the curriculum should be–i.e. which material should be regarded as trustworthy and which should not. Although we discuss a metric for trustworthiness below, the full curriculum development will be for future work Figure 3.

The socialisation that A1B0T needs to go through to gain a socially acceptable moral backbone (i.e. moral ontology) will necessarily have to be a reduced kind of socialisation since a machine is not born with the innate psychological blueprint with which humans are born. To what extent our machine can

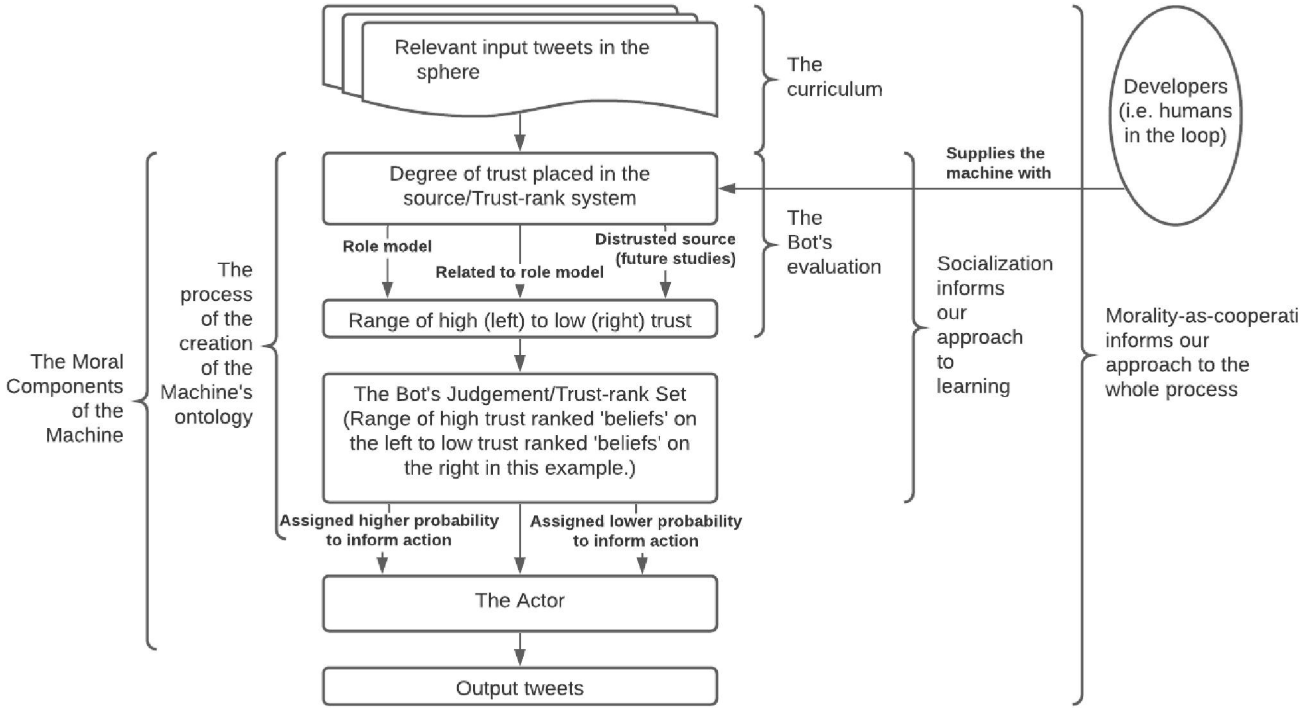
Relevant input tweets in the sphere
The curriculum
Developers (i.e. humans in the loop)
Supplies the machine with
Degree of trust placed in the source/Trust-rank system
The Bot's evaluation
Role model
Related to role model
Distrusted source (future studies)
Range of high (left) to low (right) trust
The process of the creation of the Machine's ontology
The Moral Components of the Machine
Socialization informs our approach to learning
Morality-as-cooperati informs our approach to the whole process
The Bot's Judgement/Trust-rank Set (Range of high trust ranked 'beliefs' on the left to low trust ranked 'beliefs' on the right in this example.)
Assigned higher probability to inform action
Assigned lower probability to inform action
The Actor
Output tweets

FIGURE 3 Case Study of A1B0T's learning process

engage in co-construction of values is yet to be seen; however, it will take assistance from more experienced others deemed to be trustworthy agents to learn from (the assigned role models), while also resisting learning to change from less trusted agents (non-role models).

Two points need to be highlighted in this respect. The first is that A1B0T will not be provided assistance beyond the initial list of role models supplied by the developers/researchers, i.e. it will not require others' time and energy to be 'raised' the way a child –it will take the freely available 'assistance' already in the Twittersphere, based on the model of learning described in this section.

The second point regards co-construction of values: In co-construction both agents are influencing each other based on who they are and the social dynamics between them. While we cannot predict to what extent A1B0T might influence any specific agent (individual institution etc.) the issue does not rest with the influence that it might have on a specific agent. Co-construction of values in the Twittersphere happens between all participants in the sphere. The simple act of tweeting is engagement in co-construction. The main issue with Tay was that it engaged in co-construction based on what it learnt from other users. Following what might be called the 'educational attack,' it ended up co-constructing the internet discourse in a manner that is not accepted by a significant portion of internet users and the larger society, who wished to avoid the particular kind of moral co-construction to which Tay then contributed.

3.2 | A measure of trustworthiness

A1B0T will ultimately be guided by how *trustworthy* it thinks a tweeter is. For example, the more trustworthy a source, the more weighting it will apply to the Tweet provided by that source. This is akin to the PageRank algorithm [47] used by Google to rank the usefulness of web pages based upon a search query; except instead of *usefulness*, the source of a Tweet will be ranked according to *trustworthiness*. The most trusted sources will be those at closest proximity in terms of mutual degrees of separation from an authority source, with the authority source defined by an outlet that is reflective of practises of the community A1B0T is inhabiting.

For example, suppose BBC News was deemed a trustworthy news source that is reflective of the society and values within the United Kingdom, it would deem BBC News an authority source, i.e. a news source that could be relied upon to generate trustworthy content. How trustworthy another media source is regarded would depend upon its *proximity* (in terms of degrees of separation) from BBC News. If BBC News follows me, this gives me *credibility* and it will receive a positive rating on the trustworthiness scale. Whereas if BBC does not follow me, the indication of my trustworthiness will be according to how *far away* I am in terms of 'follows' from the BBC.

In Figure 4, a hypothetical scenario, ABC, DEF and GHI news sources, after the BBC, would be regarded as the most trustworthy sources because they are each only one degree of separation, in terms of being followed by the BBC. Although Jill is only one degree of separation from the BBC, she is not followed by the BBC, and therefore, not an immediate source within the network. Instead, Jill and Ben would be ranked as the second most trustworthy sources because they are each two degrees of separation from the BBC. In this instance, Amy is the least trustworthy of the Twitter sources because her distance from the BBC is three degrees of 'follow' separation.

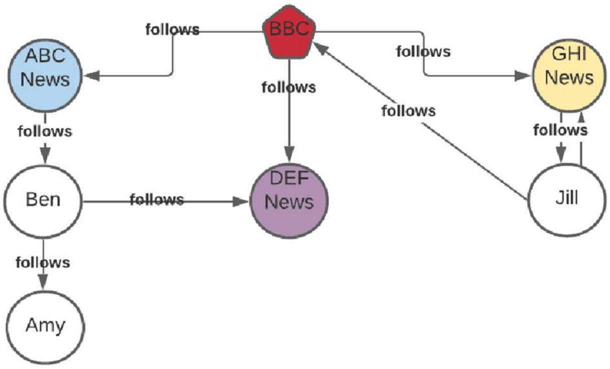

FIGURE 4 A hypothetical network of 'follow' relationships based upon interactions in the Twittersphere

This places a burden on the trust designer to select the right authority sources or at least the right combination. As mentioned above it is important that these sources, which essentially *anchor* the trust network, are the right sources to represent the given society so that the constraints are relevant too. Although it will be interesting to explore this in future work, we will not focus on this issue here. The scope of this paper is to show that chatbot behaviour can be constrained using such a technique. On that note, we now turn to some important limitations within the current design.

4 | MODEL ANALYSIS: LIMITATIONS AND OBJECTIONS

While the current design is intended as an improvement on the moral capabilities of moral conversational bots, there are also several limitations as well as disclaimers that need to be addressed.

There are four interrelated limitations and criticisms that can be directed to the current model:

First, we are anthropomorphising machines at a deeper level. This can lead to issues of uncanny valley and could also imply treating the machine as something it is not: human [48]. However, we think that the issue of anthropomorphising in the case of a chatbot applies to a lesser degree, precisely because of the social role and function of the machine–we argue that a social chatbot ought to resemble human actions in its interactions with humans. However, we should also note that the model we propose here would amount to a *reduced* 'anthropomorph' due to its narrowed intelligence.

This relates to a second limitation of the current model in that we recognise that our application of human psychology in the process of socialisation is both limited and largely reduced to a trust function. However, we believe that the initial steps of affecting human qualities and moral psychology to social chatbots requires a somewhat reduced psychological approach, while also realising that perhaps the reductionist attitude employed here may also have parallels with the reductionist approach neural networks have taken with reference to neurobiology.

A critique of the A1B0T model might be that even if it resists the sort of attacks that broke Tay, it still retains a vulnerability in that changing the anchor role models can very easily result in a robot of similarly unacceptable moral character–depending on the chosen role models. However, we think that it would still be robust with reference to the values of the chosen role models.

Finally, we are socialising the bot into the morals of a subgroup of people. In other words, it is discriminatory of others' morality in its learning. While we have endeavoured to adhere to norms that have mainstream acceptance and reject norms that are rejected by the majority of people in our examples, the method we propose here can just as easily be used to the opposite effect.

On that note, it is also worth acknowledging here that we have not delved deeper into the algorithmic implementation of the approach proposed here. We have instead focussed on the epistemological translation of some lessons learnt about human morality from the psychological and anthropological literature to the case of machines. However, these limitations also point towards some very interesting potential for further development.

5 | MOVING FORWARD: PROMISES AND FUTURE STUDIES

Moving forward, we believe the introduction of this model brings with it an exciting field of further research. Our intention is to learn from the 'problem of Tay' and offer a direction that can be taken by academics in the future and further developed.

Although here we have prescribed a model for imposing constraints onto a machine learning chatbot such as Tay, we believe that this is an approach that can be applied more broadly to learning algorithms in general, if there is a requirement to safely constrain their behaviour.

Future work following from this approach and the general philosophical angle will need to initially test, evaluate and appraise the feasibility of this approach. While we have outlined the general approach and some of the literature informing our thinking, the actual implementation and experimentation is yet to be done.

With that in mind, a body of work opens up to effectively investigate the suitable 'role models' and how they should feed into the AI's curriculum. Furthermore, although we have specified a 'trustworthiness' metric, there is further work from cognitive and behavioural science to expand this. This could be expanded by exploring network theory. We believe this research introduces several schools of future research which are listed below:

Understanding role models: We made it so that our bot places most of its trust in the BBC; future studies may focus not superficially on the content of what the role models tweet about but more deeply on *how* the models tweet, with a view to 'not learn' unacceptable behaviour that we can reliably expect the role models chosen by responsible designers to not exemplify. Furthermore, it is also important to have a debate regarding how role models are chosen–through community consensus, expert opinion, or some other method.

Negative sources: The current approach can be enriched by adding 'distrusted sources'–i.e. *negative* sources (to avoid learning from). This would reinforce the 'tribe' (moral subculture) the bot is assigned to, although it is also worth acknowledging that this decision would lock the bot into an echo chamber. Yet, this may simply be the first step towards a more liberal, as in open-minded, bot that can think for itself. This is in contrast with remaining a strictly partisan bot, given that the approach proposed here takes some steps towards creating subjective beliefs regarding the trustworthiness of sources in the machine–a prerequisite to critical thinking.

Learning in General: This approach might be the first step in creating a *critical thinking* machine. The proposed model for learning for our chatbot relies on a relatively strong adherence to role models and strong scepticism of others. Humans develop their capacity to learn and think critically as they grow, expanding their ability to engage in nuanced negotiation of information uptake in increasingly complex social environments, that is to question the reliability of role models, defer to non-role models and decide when these decisions are appropriate in different scenarios. This may be achieved incrementally, mimicking the progression of human learning from childhood to adulthood.

Governance and trust:

'*Verification of social media sources and content is challenging. It is often difficult to determine the truth, accuracy or validity, both of sources providing textual content and content presented through other modalities (video, images or audio*' [41] p. 1).

The whole notion of a Journalists' Code and governance of what is published online has been further complicated by civic journalism. Paradoxically, news coming from the general public witnessing an event first hand is considered to be accurate, even though the content can be misleading, misamplifying and presented out of context [41]. Further research could contribute to defining patterns of trust in 'role models' that could identify trustworthiness in news regardless of its source.

6 | CONCLUSION

Here, we have attempted to outline a multidisciplinary and pragmatic approach to creating a troll-resistant update to a Tay-like moral conversational machine. Outlining the approaches to morality (and trustworthiness), taken from anthropology, psychology, philosophy and media studies, we have developed a model that we believe is reflective of the body of work in these areas.

Though this model has its obvious limitations by understanding morality in terms of 'trust relationships' we believe we have introduced a new approach to creating a moral machine and a broader understanding of trust networks. It has also opened the door to creating socially responsible chatbots in the future.

ACKNOWLEDGEMENTS

This paper was born out of several interesting and insightful discussions. We would like to thank Phillip Harvey for his philosophical rigour and for always taking an interest in what we are working on. We would also like to thank the Ethical AI reading group for clarifying various resources used in this paper. Finally, special thanks needs to be given to Joseph Corneli and Nigel Crook for painstakingly checking this work for factual/theoretical inaccuracies, and for giving, as ever, constructive critical feedback.

ORCID

Rebecca Raper https://orcid.org/0000-0001-8536-1291

REFERENCES

1. Wallach, W., Allen, C.: Moral Machines: Teaching Robots Right from Wrong. Oxford University Press (2008)
2. Anderson, M., Anderson, S.L.: Machine ethics: creating an ethical intelligent agent. AI. Mag., 28(4), 15 (2007)
3. Vanderelst, D., Winfield, A.: An architecture for ethical robots inspired by the simulation theory of cognition. Cognit Syst Res. 48, 56–66 (2018)
4. Killen, M., Smetana, J.G.: Handbook of Moral Development. Psychological Press, New York (2014)
5. Floridi, L., et al.: AI4People-An ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds. Mach. 28(4), 689–707 (2018)
6. Foley, M.J.: Microsoft launches AI chabot, Taya. ZDNet. (2016) Retrieved from: https://www.zdnet.com/article/microsoft-launches-ai-chat-bot-tay-ai/
7. Zemcik, T.: Failure of chatbot Tay was evil, ugliness and uselessness in its nature or do we judge it through cognitive shortcuts and biases? AI. Soc. (2020) https://doi.org/10.1007/s00146-020-01053-4
8. Lee, P.: Learning from Tay's introduction, Statement published in The Official Microsoft Blog. (2016) Retrieved from: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/
9. Suarez-Gonzalo, S., Mas-Manchón, L., Guerrero-Solé, F.: Tay is you: the attribution of responsibility in the algorithmic culture. Observatorio. 13(2), 1–14 (2019)
10. Stuart-Ulin, C.R.: Microsoft's politically correct chatbot is even worse that its racist one. Quartz. (2018) https://qz.com/1340990/microsofts-politically-correct-chat-bot-is-even-worse-than-its-racist-one/. Accessed: 14 April 2021
11. Heaven, W.D.: How to Make a Chatbot that isn't Racist or Sexist. MIT Technology Review. (2020) https://www.technologyreview.com/2020/10/23/1011116/chatbot-gpt3-openai-facebook-google-safety-fix-racist-sexist-language-ai/. Accessed: 14 April 2021
12. Wallach, W., Allen, C.: Android ethics: Bottom-up and top-down approaches for modeling human moral faculties. In: Proceedings of the 2005 CogSci Workshop: Toward Social Mechanisms of Android Science pp. 149–159. (2005)
13. Asimov, I.: Runaround. Astounding. Sci. fict., 29(1), pp. 94-103.(1942)
14. Alexander, L., Moore, M.: Deontological ethics. In: Zalta E.N. (eds.) The Stanford Encyclopedia of Philosophy. (2020) (Winter 2020 Edition), (forthcoming) https://plato.stanford.edu/archives/win2020/entries/ethics-deontological/
15. Brink, D.: Mill's moral and political philosophy. In: Zalta E.N. (eds.) The Stanford Encyclopedia of Philosophy, (2018) (Winter 2018 Edition), https://plato.stanford.edu/archives/win2018/entries/mill-moral-political/
16. Anderson, M., Anderson, S.L.: ETHEL: Toward a principled ethical eldercare robot. (2008). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.177.5971
17. Hardwig, J.: Epistemic dependence. J Philos. 82(7), 335–349 (1985)
18. Hutson, M.: Why You don't Really Know What You Know. MIT Technology Review. (2020). Available at: https://www.technologyreview.com/2020/10/21/1009445/the-unbearable-vicariousness-of-knowledge/. Accessed: 1 Nov 2020
19. Haidt, J.: The new synthesis in moral psychology. Science. 316, 998–1002 (2007)
20. Shackelford, T.K., Hansen, R.D. (eds.): The Evolution of Morality. Springer International Publishing (2016)
21. Sinnott-Armstrong, W. (eds.): Moral Psychology, vol. 1–3. MIT Press, Cambridge (2007)
22. Haidt, J.: The righteous mind: why good people are divided by politics and religion. Pantheon Books, New York (2012)
23. Kristjansson, K.: The Self and Its Emotions. Cambridge University Press, New York (2010)
24. Haidt, J.: The emotional dog and its rational tail: a social intuitionist approach to moral judgement. Psychol Rev. 108, 814–834 (2001)
25. Bowles, S., Gintis, H.: A Cooperative Species: Human Reciprocity and Its Evolution. Princeton University Press, UK (2011)
26. Bugental, D.B., Goodnow, J.G.: Socialisation processes. In: Damon, W. et al. (eds.) Handbook of Child Psychology: Vol. 3. Social, Emotional, and Personality Development, 5th ed, pp. 389–462. John Wiley & Sons, New York (1998)
27. Grusec, J.E., et al.: The development of moral behaviour from a socialisation perspective. In: M. Killen & J. G. Smetana (eds.) Handbook of Moral Development (2nd eds, pp. 113-134) New York: Psychological Press (2014)
28. Serafimova, S.: Whose morality? Which rationality? Challenging artificial intelligence as a remedy for the lack of moral enhancement. Nat. Humanit. Soc. Sci. Commun. 7, 199 (2020) Available at: https://doi.org/10.1057/s41599-020-00614-8. Accessed: 27 October 2020
29. Doebel, S., Koenig, M.A.: Children's use of moral behaviour in selective trust: discrimination versus learning. Dev Psychol. 49(3), 462–469 (2013)
30. Mascaro, O., Sperber, D.: The pragmatic role of trust in young children's interpretation of unfamiliar signals. PloS One. 14(10), e0224648, (2019) https://doi.org/10.1371/journal.pone.0224648
31. Smetana, J.G.: The role of parents in moral development: a social domain analysis. J. Moral. Educ. 28(3), 311–321 (1999)
32. Hooghe, M.: Social capital and diversity generalized trust, social cohesion and regimes of diversity. Can J Polit Sci. 40(3), 709–732 (2007)
33. Heintz, C., Karabegovic, M., Molnar, A.: The Co-evolution of honesty and strategic vigilance. Front Psychol. 7, 1503 (2016). https://doi.org/10.3389/fpsyg.2016.01503
34. Mercier, H.: People are less gullible than you think. Reason, (2020) March 2020 Issue. https://reason.com/2020/02/09/people-are-less-gullible-than-you-think/. Accessed: 7 October 2020
35. Curry, O.S.: Morality as cooperation: a problem-centred approach. In: The Evolution of Morality pp. 27-51. Springer, Cham (2016)
36. Curry, O.S., Mullins, D.A., Whitehouse, H.: Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. Curr Anthropol. 60(1), 47–69 (2019)
37. Dewey, J.: Democracy and Education. The Macmillan Company, New York (1916)
38. Rousseau, J.J.: Emile, or On Education. Translated by Barbara Foxley, London: J. M. Dent and Sons 1921 (1762) Available at: http://lf-oll.s3.amazonaws.com/titles/2256/Rousseau_1499_EBk_v6.0.pdf

39. Stenhouse, L.: An Introduction to Curriculum Research and Development. Heinemann Educational Books Ltd, Oxford (1975)
40. Tyler, R.W.: Basic Principles of Curriculum and Instruction. University of Chicago Press, Chicago (1949)
41. Brandtzaeg, P.B., et al.: Emerging journalistic verification practices concerning social media. J. Pract., 10, 323-342 (2015) Available at https://doi.org/10.1080/17512786.2015.1020331
42. Zi, Y., Ghanbar, A.: Journalism ethics development: a comparison of ethics code in USA, UK, AUS, Tunisia and China (2012)
43. Journalism Ethics and Standards: Wikipedia. (2020) https://en.wikipedia.org/wiki/Journalism_ethics_and_standards. Accessed: 20 October 2020
44. Newman, N., et al.: Digital News Report 2020 Pub. Reuters Institute for the study of Journalism, Oxford (2020) https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf
45. Ad Fontes Media: Media Bias Chart 6.0. (2020) https://www.adfontesmedia.com/gallery/. Accessed: 20 October 2020
46. Rajan, A.: Instagram will overtake twitter as a news source, BBC, (2020) https://www.bbc.co.uk/news/technology-53050959. Accessed 5 November 2020
47. Pasquinelli, M.: Google's PageRank algorithm: a diagram of the cognitive capitalism and the rentier of the common intellect. In: Becker, K., Stalder, F. (eds.) Deep Search: The Politics of Search Beyond Google. Transaction Publishers, London (2009)
48. Proudfoot, D.: Anthropomorphism and AI: turing's much misunderstood imitation game. Artif Intell. 175, 950–957 (2011)

How to cite this article: Bridge, O., et al.: Modelling a socialised chatbot using trust development in children: Lessons learnt from Tay. Cogn. Comput. Syst. 3(2), 100–108 (2021). https://doi.org/10.1049/ccs2.12019