*David Tan**

THE THOUGHT PROBLEM AND JUDICIAL REVIEW OF ADMINISTRATIVE ALGORITHMS

Abstract

The issue of whether algorithms can be characterised as thinking or having properties of thought has arisen in both judicial decisions like *Pintarich v Deputy Commissioner of Taxation* and scholarly discussion regarding issues like bias. This article refers to this issue as the 'thought problem' and introduces three principles for how to resolve it: (1) the manifestation principle; (2) the implementation principle; and (3) the equivalent treatment principle. The manifestation principle states that algorithmic outputs can be considered decisions where the manifestation of conduct of the agency supervising the algorithm would be understood to the outside world as a product of a thinking person. The implementation principle states that the humans in the executive who implemented the algorithm have responsibility for the algorithm. The equivalent treatment principle proposes to treat algorithms and humans who reason similarly as equivalent before the eyes of administrative law. The article does not try to conclusively resolve which principle is best, but rather suggests that the equivalent treatment principle is the most complete one for dealing with the thought problem.

I Introduction

A prominent feature of contemporary life is an increased use of automated processes in almost every aspect of our daily activities, including how we are governed. This has led to a worry among both administrative law academics and practitioners about a certain feature of automated decision-making, which this article calls the thought problem: that some aspects of judicial review — for example, determining whether a decision-maker is actually biased — assume

* Lecturer, Faculty of Law and Business, Deakin Law School. The author would like to thank attendees at the 2021 Australian Institute of Administrative Law National Administrative Law Conference and a Deakin Law School work-in-progress seminar for feedback on an earlier version of this article. A special debt is owed to Colin Campbell and Yee-Fui Ng who read through and commented on this article. The comments of anonymous reviewers also greatly strengthened the article. Any mistakes are of course my own.

that statutory powers will be exercised by thinking persons, but machine algorithms might not 'think'.[1]

Three principles will be considered for dealing with the thought problem in relation to administrative judicial review. The first is the manifestation principle, which asserts that if it appears that an agency has issued a decision, then the agency is taken as having made the decision.[2] The second is the implementation principle, where a human (in the government) implemented an automated system that is not compliant with some statute, responsibility should be attributed to the human. The third is the equivalent treatment principle, where a human (in the government) who implemented the machine algorithm should be treated as if that human personally followed all the rules of the machine algorithm. All three principles are differently motivated and so will be presented as independent alternatives, although the article does not exclude the possibility of a mixture of principles.

As a preliminary remark, it is important to separate the policy issue from the judicial review issue arising out of the use of machine algorithms. The policy issue is executive-facing: how should the government use machine algorithms?[3] The judicial review issue is judge-facing: if machine algorithms have been implemented, what is the role of the judge in reviewing them? This article only discusses the judicial review question and not the policy question. The judicial review issue can be further divided into a normative and a descriptive question. The normative question asks how judges should deal with machine algorithms supposing that new rules could be introduced into administrative law. The descriptive question queries how judges can review machine algorithms given the current rules of the legal

---

1 See generally: *Pintarich v Deputy Commissioner of Taxation* (2018) 262 FCR 41 ('*Pintarich*'); Yee-Fui Ng and Maria O'Sullivan, 'Deliberation and Automation: When Is a Decision a "Decision"?' (2019) 26(1) *Australian Journal of Administrative Law* 21; Sarah Lim, 'Re-Thinking Bias in the Age of Automation' (2019) 26(1) *Australian Journal of Administrative Law* 35; Justice Melissa Perry, 'iDecide: Administrative Decision-Making in the Digital World' (2017) 91(1) *Australian Law Journal* 29, 31; Katie Miller, 'The Application of Administrative Law Principles to Technology-Assisted Decision-Making' (2016) 86(1) *AIAL Forum* 20, 22; Lawrence B Solum, 'Legal Personhood for Artificial Intelligences' (1992) 70(4) *North Carolina Law Review* 1231, 1248, 1267; Will Bateman 'Algorithmic Decision-Making and Legality: Public Law Dimensions' (2020) 94(7) *Australian Law Journal* 520, 523–5; Anna Huggins, 'Addressing Disconnection: Automated Decision-Making, Administrative Law and Regulatory Reform' (2021) 44(3) *University of New South Wales Law Journal* 1048, 1061–70.

2 *Pintarich* (n 1) 49–50 [52]–[55] (Kerr J).

3 See, eg: Tania Sourdin, *Judges, Technology and Artificial Intelligence: The Artificial Judge* (Edward Elgar Publishing, 2021) ch 3; Yee-Fui Ng et al, 'Revitalising Public Law in a Technological Era: Rights, Transparency and Administrative Justice' (2020) 43(3) *University of New South Wales Law Journal* 1041; Bateman (n 1) 525; Miller (n 1) 31–2. Many of the chapters in the excellent edited collection from Janina Boughey and Katie Miller are also targeted at the policy side: Janina Boughey and Katie Miller (eds), *The Automated State: Implications, Challenges and Opportunities for Public Law* (Federation Press, 2021).

system. This article discusses the three principles as normative strategies and will only briefly comment on their descriptive fit in closing.[4]

There are five Parts to this article: Part II provides an overview of algorithms; Part III introduces the thought problem; Part IV then introduces the manifestation, implementation, and equivalent treatment principles; and finally, Part V shows how the principles can be used to resolve the thought problem.

II What Are Algorithms?

An algorithm, informally described, is a well-defined procedure for solving some problem or producing some output.[5] To illustrate with a basic example, consider a sorting procedure for arranging playing cards.[6] Suppose we have five cards of the same suit in five positions A–E:

Position A	**Position B**	**Position C**	**Position D**	**Position E**
Card 1	Card 2	Card 3	Card 4	Card 5

The problem is how to arrange these cards such that the card in position A has the lowest value out of the five, and that the card in position B has the second lowest value and so on. A simple sorting algorithm[7] would have a machine start with the card at B, compare that card to the one in A, rearrange the cards if the card in B has a lower value than the card in A, and then subsequently move to position C. If the card in B has a higher value than the card in A, then the machine would move to position C directly. Once at position C, it will compare the card in C with the card in B and swap the order if the card in C has a lower value. Then the machine will do the same with the cards in positions B and A again. Having done this, it will then move to position D and carry out the same process with comparing the cards in positions A, B and C. Finally, it will move to position E and make the same comparisons with the cards in positions A–D. The algorithm above will always output cards of the same suit in the right order no matter what the input (ie initial cards) is. In essence, the point is that all algorithms — from simple ones to complex machine learning algorithms — follow step-by-step procedures in a mechanical manner.

An important aspect of this definition is that an algorithm is not identical to a computer or machine — an algorithm is a sequence of rules executable by an entity.[8]

---

4 I do not necessarily invoke Ronald Dworkin's concept of 'fit': see generally Ronald Dworkin, *Law's Empire* (Hart Publishing, 1998) 230–1; but simply the idea of comparing the principles to their consistency or coherence with existing legal doctrine (where coherence is neutral between Dworkinian fit or any other descriptive view of law).

5 Thomas H Cormen et al, *Introduction to Algorithms* (MIT Press, 3rd ed, 2009) 5.

6 This example draws on the discussion in ibid 16–18.

7 Ibid 16–18.

8 See ibid 5.

Notice that the sorting algorithm above could have been executed by a human rather than a machine. In this article, the term *algorithm* will be used in this abstract sense of being a sequence of rules that could be executed by some entity whether it is a machine or, at least in theory, a human. The term *machine algorithm* will be used when specifically referring to an algorithm executed by a machine. This distinction becomes important for the equivalent treatment principle (although it is not pertinent for the other two principles discussed). Another implication of this distinction is that issues surrounding automated decision-making can also arise where humans blindly execute an algorithm set up by another without thinking. Hence, in my view, the difficulties associated with reviewing automated decision-making are really a special case of the more general problems associated with reviewing decentralised decision-making where one person sets the rules and another executes it.[9] The principles discussed in this article are thus potentially generalisable to other issues in administrative law that involve complex decentralisation (eg with outsourcing).

There are of course many types of machine algorithms ranging from the simple one mentioned above to sophisticated neural networks.[10] As a matter of taxonomical simplification, this article will only distinguish between predictable and unpredictable algorithms,[11] and further, will not consider so-called black box algorithms. The arguments in this article apply to all other types of algorithms aside from black boxes. These two terms will be defined below.

This article defines unpredictable algorithms as algorithms where one is uncertain how the algorithm will achieve the target output. For example, in 1997, a graduate-level artificial intelligence ('AI') class was tasked with a project: to program computers to play tic-tac-toe with each other 'on an infinitely large board'.[12] One entry used an evolutionary algorithm — an algorithm that tries out all possible

9 For similar views, see the comments on outsourcing in Matthew Groves, 'Fairness in Automated Decision-Making' in Janina Boughey and Katie Miller (eds), *The Automated State: Implications, Challenges and Opportunities for Public Law* (Federation Press, 2021) 14, 23.

10 For a summary of machine algorithms and some of their legal implications, see generally: Marc Cheong and Kobi Leins, 'Who Oversees the Government's Automated Decision-Making? Modernising Regulation and Review of Australian Automated Administrative Decision-Making' in Janina Boughey and Katie Miller (eds), *The Automated State: Implications, Challenges and Opportunities for Public Law* (Federation Press, 2021) 174; Christopher Markou and Simon Deakin, '*Ex Machina Lex*: Exploring the Limits of Legal Computability' in Simon Deakin and Christopher Markou (eds), *Is Law Computable? Critical Perspectives on Law and Artificial Intelligence* (Hart Publishing, 2020) 31.

11 It should be noted that the terms 'predictable algorithm' and 'unpredictable algorithm' as defined in this article are not technical terms used by computer scientists, but rather terms introduced here to aid with legal theorising.

12 Joel Lehman et al, 'The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities' (2020) 26(2) *Artificial Life* 274, 284.

strategies and continually uses the ones that resulted in wins.[13] The optimal strategy used by the machine algorithm was to immediately place their X or O a far enough distance away from the other program, such that the other program would crash when making its calculations.[14] Despite knowing the initial evolutionary algorithm, the programmers could not have predicted that the algorithm would have won by choosing the strategy of crashing the other program.

On the other hand, this article does not consider black box algorithms. Black boxes are described by computer scientist Cynthia Rudin and historian Joanna Radin as follows: '[i]n machine learning, these black box models are created directly from data by an algorithm, meaning that humans, *even those who design them*, cannot understand how variables are being combined to make predictions'.[15]As noted by Rudin and Radin, even from a designer's perspective, it is intrinsically difficult to explain the classifications or predictions of black boxes.[16] Black boxes tend to be found in machine learning algorithms — algorithms that aim to make classifications by extracting patterns from data.[17] The problem is that these machine learning models make classifications based on what fits existing data — given millions of datasets, the algorithm classifies an object as likely being a hot dog — rather than using a precise set of necessary and sufficient features for their classifications (eg the black box does not operate upon pre-defined features, such as whether a sausage is present, for when something is a hot dog or not a hot dog).[18]

An algorithm can be unpredictable even if it is not a black box. Take the example above of the evolutionary algorithm: repeat different strategies and continue using the most successful ones. That algorithm is transparent — we have just explained how it works and one can understand how we arrived at the final strategy — nonetheless, such an algorithm is not predictable since we do not know ahead of time which strategies are successful (if we did, we would not need to use such an algorithm). However, all black boxes are unpredictable. We have defined predictability as having a high confidence in the method by which the algorithm will achieve the target output (before the achievement of that result). However, if black boxes are opaque then they are not predictable by definition, since we do not know how the black box is making its classifications. Do note that predictability as defined here is not the same as the reliability of the black box, that is, how accurate the output of the black box is. If one

13 More generally, programmers of an evolutionary algorithm will determine a 'fitness function' that determines which machine algorithm gets selected: see ibid 277. In the tic-tac-toe case the fitness function would have been wins in the game.

14 Lehman et al (n 12) 284.

15 Cynthia Rudin and Joanna Radin, 'Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition' (2019) 1(2) *Harvard Data Science Review* (emphasis added).

16 Ibid. See also Cheong and Leins (n 10) 183–6.

17 Danilo Bzdok, Martin Krzywinski and Naomi Altman, 'Machine Learning: A Primer' (2017) 14(12) *Nature Methods* 1119, 1119. However, not all machine learning methods are black boxes. For example, classifying data using decision trees is a machine learning method, but it is typically not a black box since one can explain the decision trees.

18 See Cheong and Leins (n 10) 184–6.

had a black box that could classify things as a hot dog or as not a hot dog 95% of the time, this would be a reliable black box despite being unpredictable.

To give some concrete examples, consider the algorithm for recovering overpaid social security payments, which was commonly known as the 'Robodebt system'.[19] The Robodebt algorithm used an 'income averaging' system for identifying debt owed to the government.[20] This algorithm was not a black box model. One just needed to understand how to use the right averaging formula. The algorithm was also predictable, as the debt notice was issued via the identification of debt owed through an averaging function.[21] As for a black box algorithm, consider for instance a machine algorithm that tries to predict the type of crime being committed.[22] It is possible to create such a machine algorithm based solely on input details being the time and location of previous crimes — where no rules for specific weightings of features were implemented into the machine algorithm (eg the programmer did not input a rule that the probability of a violent crime was more likely at night).[23] Since it is unclear what features of the data played the most important roles for determining the probabilities of the type of crime, this would be a black box. Note that not all algorithms trying to make predictions are black boxes. Rudin and Radin discuss a machine learning model that ultimately generated the following rule for predicting whether someone would reoffend within two years — reoffending would occur if the person either: (1) has more than three prior crimes; or (2) 'is 18–20 years old and male'; or (3) 'is 21–23 years old and has two or three prior crimes'.[24] This machine algorithm is not a black box and is predictable given that we know how the rule operates and how it seems to have some rational link to rates of reoffending.

While black boxes themselves raise important conceptual problems for administrative lawyers, a separate article of its own would be required to satisfactorily address the issues raised.[25] There are several reasons why such issues can be left for a different time. First, even if black boxes could be made transparent, the thought problem still remains — judicial review principles still often assume thinking entities. Hence, the thought problem needs to be addressed regardless. Second,

19 For a general overview, see: Terry Carney, 'Robo-Debt Illegality: The Seven Veils of Failed Guarantees of the Rule of Law?' (2019) 44(1) *Alternative Law Journal* 4; Darren O'Donovan, 'Social Security Appeals and Access to Justice: Learning from the Robodebt Controversy' [2020] 158 *Precedent* 34.

20 *Prygodicz v Commonwealth [No 2]* (2021) 173 ALD 277, 287 [38] ('*Prygodicz*').

21 Ibid 287–8 [38]–[41].

22 Steven Walczak, 'Predicting Crime and Other Uses of Neural Networks in Police Decision Making' (2021) 12 *Frontiers in Psychology* 587943:1–11, 6.

23 Ibid 6. Although admittedly the reliability of this algorithm was better than randomly guessing: at 8.

24 Rudin and Radin (n 15).

25 See generally: Cheong and Leins (n 10); Ashley Deeks, 'The Judicial Demand for Explainable Artificial Intelligence' (2019) 119(7) *Columbia Law Review* 1829; Lim (n 1) 42; Bateman (n 1) 526–8; Janina Boughey, 'Outsourcing Automation: Locking the "Black Box" inside a Safe' in Janina Boughey and Katie Miller (eds), *The Automated State: Implications, Challenges and Opportunities for Public Law* (Federation Press, 2021) 136.

the issues with the thought problem arise even in non-black box cases. It would be expected that a significant amount of administrative automation does not need powerful machine learning techniques and thus could be done through non-black box algorithms (eg in *Pintarich v Deputy Commissioner of Taxation* ('*Pintarich*') as will be discussed below).[26] Thus, addressing the thought problem for such cases is still an important endeavour. Third, black box cases raise very different questions from the thought problem: what counts as a reasonable administrative decision, and are reliable black boxes that are nonetheless opaque considered as reasonable decisions or can such black boxes count as evidence for decisions? This investigation into reasons and justifications is a somewhat different issue than whether decision-makers need to form certain mental states when making decisions.[27]

III The Thought Problem

The crux of this problem has been well summarised by Will Bateman:

> ensuring that exercises of power are justified on social and democratic grounds is a prime objective of public law. … that objective … require[s] that statutory powers be exercised by agents who: have certain *cognitive capacities* …[28]

The thought problem occurs because certain aspects of judicial review seem to require thinking entities and yet it seems like machine algorithms do not have thoughts or mental states.[29]

---

26 *Pintarich* (n 1).

27 See generally: Maya Krishnan, 'Against Interpretability: A Critical Examination of the Interpretability Problem in Machine Learning' (2020) 33(3) *Philosophy and Technology* 487; Alejandro Barredo Arrieta et al, 'Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI' (2020) 58 (June) *Information Fusion* 82.

28 Bateman (n 1) 520 (emphasis in original) (citations omitted).

29 While there is some debate among philosophers of mind as to this proposition, this article assumes machine algorithms do not think (otherwise the thought problem trivially disappears). For views that algorithms might instantiate mental properties, consider the once-popular computational theory of mind that posits that mental states are computational states. To oversimplify, this theory posits that how the human mind operates is similar to how a machine executes an algorithm. On such a view it seems like even primitive machine algorithms instantiate primitive mental states. For an early framing of the computational theory of mind see Hilary Putnam, 'Psychological Predicates' in WH Capitan and DD Merrill (eds), *Art, Mind, and Religion: Proceedings of the 1965 Oberlin Colloquium in Philosophy* (University of Pittsburgh Press, 1967) 37. For a general overview see, Michael Rescorla, 'The Computational Theory of Mind' *Stanford Encyclopaedia of Philosophy* (Web Page, 24 July 2023) <https://plato.stanford.edu/entries/computational-mind/>.

It is also the case that, in administrative law, many do accept that machine algorithms do not think: see, eg: *Pintarich* (n 1) 67–8 [143]–[145] (Moshinsky and Derrington JJ); Lim (n 1) 38; Bateman (n 1) 526.

Two subcomponents of the thought problem have arisen in the literature and case law: (1) the attribution issue; and (2) the illegality issue. The attribution issue asks whether an action can be attributed to the executive when the action was a result of a machine algorithm. The illegality issue arises in relation to established grounds of administrative law which assume that certain decisions are illegal because of the presence or absence of certain mental states. The question is whether these grounds are still relevant in the context where a machine algorithm is used. These issues arise regardless of the level of discretion provided to the decision-maker; as will be seen below, existing administrative law doctrine assumes that decision-makers have cognitive capacities even in the most mundane of tasks. It is also noted that these problems are exacerbated where unpredictable algorithms are utilised.

A *Attribution*

The attribution issue was most prominently illustrated in *Pintarich*, which was an *Administrative Decisions (Judicial Review) Act 1977* (Cth) ('*AD(JR) Act*').[30] case. A formal letter had been sent to Pintarich stating that he owed a certain amount of money to the Australian Taxation Office.[31] The letter was produced when a delegate of the Deputy Commissioner inputted numbers into a computer program (the ensuing letter was not checked by the delegate).[32] Later, Pintarich was informed that the letter was sent incorrectly and that instead a different, larger amount of money was owed.[33] The Court ruled that the definition of a decision necessarily included 'the reaching of a conclusion as a result of a mental process'.[34] The court then found that since the human delegate had not turned their mind to the calculations in the letter no decision had been made,[35] a logical consequence of which is that the operation of the machine algorithm did not count as being a mental process leading to the letter being issued. Hence the initial letter was not a decision and thus had no legal effect. As Yee-Fui Ng and Maria O'Sullivan note, this decision potentially means that any automated finding (with no further human intervention) would not count as a decision.[36] This raises concerns whether such outputs can ever be reviewed under the *AD(JR) Act*, since it requires the existence of a decision or possible decision for the availability of review.[37]

Ng and O'Sullivan also note that the Administrative Appeals Tribunals in *Re Bowron and Secretary, Department of Social Security* ('*Re Bowron*')[38] and *Re Dimitrievski*

30 *Pintarich* (n 1) 61 [117].

31 See ibid 58 [101].

32 See ibid.

33 Ibid 60 [110].

34 Ibid 67–8 [143] (Moshinsky and Derrington JJ).

35 Ibid 68 [144]–[145].

36 Ng and O'Sullivan (n 1) 30.

37 See *Administrative Decisions (Judicial Review) Act 1977* (Cth) ss 3(1) (definition of 'decision to which this Act applies'), 5–7 ('*AD(JR) Act*').

38 (1990) 21 ALD 333, 336 ('*Re Bowron*'), cited in Ng and O'Sullivan (n 1) 30.

and Secretary, Department of Social Security ('*Re Dimistrievski*')[39] came to similar conclusions about decisions.[40] The issue from *Re Bowron* and *Re Dimitrievski* has now been resolved in the social security context since s 6A of the *Social Security (Administration) Act 1999* (Cth) ('*SSA Act*') deems outputs of computer programs as decisions of the Secretary.[41] However, Justice Melissa Perry has queried whether provisions like s 6A are conceptually coherent.[42] Her Honour notes that this is tantamount to 'delegating' authority to machine algorithms, but it is not quite clear whether delegation makes sense in the context of automation (eg who is the decision-maker and to whom has authority been delegated?).[43]

Even if we looked beyond the *AD(JR) Act*, similar issues might arise where applicants seek certiorari for quashing a 'decision' — which is an ancillary remedy for s 75(v) of the *Constitution*.[44] Certainly, *Pintarich* was an *AD(JR) Act* case, but it is not obvious that the characterisation of what a 'decision' is should be different across different types of review. It might be said that the appropriate target of certiorari is something like an 'exercise of power', but it is unclear how an exercise of power is substantively different from a decision. Even outside of the idea of decision-making, other jurisdictional issues arise. When considering the jurisdiction of the High Court under s 75(v), it is similarly unclear whether a non-thinking machine algorithm could be an 'officer of the Commonwealth'.[45] However, this problem is not as intractable since s 75(v) might be sidestepped by using s 75(iii), which does not have an 'officer of the Commonwealth' requirement and does not require any human involvement in the process.[46]

B *Illegality*

In certain cases, a decision may be taken to be illegal if certain mental states are absent or present in the decision-maker. Bateman notes that public law typically assumes that decision-makers reason in a linguistically sophisticated and environmentally sensitive way when exercising their powers under a statute.[47] As he further notes, no currently existing algorithm can reason in such a manner.[48] For example, take statutes which, whether explicitly or implicitly, place constraints on the

39 (1993) 31 ALD 140 ('*Re Dimistrievski*'), cited in Ng and O'Sullivan (n 1) 30.

40 Ng and O'Sullivan (n 1) 30.

41 See also ibid 30–1.

42 See Perry (n 1) 31.

43 Ibid.

44 See *Re Refugee Review Tribunal; Ex parte Aala* (2000) 204 CLR 82, 90–1 [14] (Gaudron and Gummow JJ).

45 Ng et al (n 3) 1058–9.

46 See also *Plaintiff M61/2010E v Commonwealth* (2010) 243 CLR 319, 345 [51].

47 Bateman (n 1) 523–4.

48 Ibid 525–6.

relevancy of considerations that decision-makers can take into account.[49] In *Tickner v Chapman,* the consideration of a factor was defined as an 'active intellectual' engagement.[50] This raises the question of how relevancy analysis can be applied if no entity has had an active intellectual engagement with the requirements of statute.

More generally, it can be said that in many cases statutes require that certain mental states must be present (eg relevant considerations) or excluded (eg irrelevant considerations) — thus giving rise to grounds of review that focus on mental states. The question is how to analyse such assumptions and grounds in the context of automation. To give a further example where grounds assume mental states, Sarah Lim has argued that actual bias as currently conceptualised has difficulties being applied to machine algorithms — a decision is unlawful if there is a 'pre-existing state of mind' which affects a proper consideration of the matter.[51] Nonetheless, since machine algorithms do not think or have states of minds, actual bias cannot, as it is currently understood, be attributed to machine algorithms — even if the machine algorithm seems to disproportionately produce certain outcomes over others in a way inconsistent with the purpose of a statute.[52]

It might be contended that, even if actual bias cannot be relied upon, apprehended bias might be used.[53] Since apprehended bias relies on what the 'lay observer' would perceive to be biased,[54] an output may be unlawful even if the source of the output has no mental state. This article does not try to contest this specific point since, as noted above, other grounds such as relevant and irrelevant considerations might also have mental element conditions. Nonetheless, an appeal to apprehended bias is not an easy solution. First, Lim notes that when deciding whether there is apprehended bias courts typically consider the 'prejudices, influences and frailties of human actors' and the extent to which humans can discard those prejudices (often allowing for some inevitable human frailty).[55] It is unclear how such considerations of ulterior interests and frailties could occur, or would be regarded as having occurred, with automated systems.[56] Second, if one tries to attribute these ulterior interests to

---

49 See generally *Minister for Aboriginal Affairs v Peko-Wallsend Ltd* (1986) 162 CLR 24, 39–42 (Mason J).

50 (1995) 57 FCR 451, 462 (Black CJ). See also Bateman (n 1) 523.

51 Lim (n 1) 37, quoting *Jia v Minister for Immigration and Multicultural Affairs* (1998) 84 FCR 87, 104. While Lim does spend most of her article on apprehended bias, she does argue that the rule against bias (as encompassing both actual and apprehended bias) in general faces difficulties since our doctrine refers to 'states of mind': at 37.

52 For some of the issues related to attributing bias to machine algorithms, see generally Lim (n 1) 38.

53 See generally *Ebner v Official Trustee in Bankruptcy* (2000) 205 CLR 337, 344–5 [6] (Gleeson CJ, McHugh, Gummow and Hayne JJ) ('*Ebner*').

54 Ibid.

55 Lim (n 1) 38 (emphasis omitted).

56 See also ibid 38–9.

programmers, Anna Huggins suggests that such an attribution may be difficult since they are far removed from the ultimate output of the machine algorithm.[57]

C *Predictability*

Both the attribution and illegality issues deepen when we consider unpredictable algorithms. If the output of a machine algorithm could be predicted by a human, who proceeded to use that algorithm, it could be plausibly argued that responsibility for the machine algorithm's outputs along with non-compliance should be attributed to them (as will be claimed under the implementation principle). For example, the 'biased' outcome of a machine algorithm could also be attributed to a human if that outcome was predictable when the human implemented the machine algorithm. However, where the machine algorithm is unpredictable, these claims seem difficult to defend since the human would not have known what the machine algorithm would do.

D *Existing Recommendations*

Given the increased use of machine algorithms, several organisations have made recommendations or practice guides for dealing with automated decision-making.[58] It should be noted that these proposals do not discuss the thought problem at length and, where it is canvassed, only the attribution issue is covered.

On the attribution issue, current recommendations are for legislative change to ensure the applicability of judicial review. The Law Council of Australia suggests that, in response to the judgment in *Pintarich*, there should be a 'comprehensive legislative response which ensures all [automated decision-making] is lawful and subject to judicial review'.[59] Other organisations suggest that this can be done through deeming provisions such as s 6A of the *SSA Act*. The Australian Human Rights Commission recommends that 'relevant legislation including s 25D of the *Acts Interpretation Act 1901* (Cth)' should be modified so that 'decisions' are interpreted as including 'decisions made using automation and other forms of artificial intelligence'.[60] Similarly, the Commonwealth Ombudsman suggests that the authority to make a decision using an automated process would 'only be beyond doubt' if specified by legislation.[61]

---

57 See Huggins (n 1) 1067.

58 See, eg: Australian Human Rights Commission, *Human Rights and Technology* (Final Report, 2021) chs 4–5, 7–8; Commonwealth Ombudsman, *Automated Decision-Making: Better Practice Guide* (Practice Guide, 2019); Law Council of Australia, Submission to Digital Technology Taskforce, Department of the Prime Minister and Cabinet, *Positioning Australia as a Leader in Digital Economy Regulation: Automated Decision Making and AI Regulation* (3 June 2022).

59 Law Council of Australia (n 58) 20–2 [70]–[83].

60 Australian Human Rights Commission (n 58) 62, recommendation 6.

61 Commonwealth Ombudsman (n 58) 9.

This article does not take issue with these recommendations but, as noted by Justice Perry above, there is little justification for why such deeming provisions are conceptually coherent.[62] The Australian Human Rights Commission recommends, for non-governmental entities, that there be a presumption that private corporations or legal persons are responsible for all actions regardless of whether a machine algorithm was used:[63]

> As a general rule, whoever is responsible for making a decision is responsible also for any errors or other problems that arise in the decision-making process. Where this person has relied on a third party in the process of decision making, and the third party caused the error to take place, the first person remains liable for the error.[64]

It might be argued that the same presumption should be applied to government entities. In essence, this is the implementation approach that will be discussed below. As will be explained, however, this is not a magic bullet. The implementation principle deals with the attribution issue but has difficulties with the illegality issue.

Sceptics might say that this level of theorising is over-analysis and that deeming provisions are either legal fictions or are a pragmatic approach with little downside. It is not obvious that these sceptical responses resolve the thought problem. If the legal fiction route is taken, then more needs to be said about how and when the government can create legal fictions. Further, this article shows that there are other ways to justify such deeming provisions without just assuming a fiction. The pragmatic view on the other hand finds a natural home or ally in the equivalent treatment principle discussed below; nonetheless, the equivalent treatment principle does not mean everything goes, and clear boundaries for how and when the principle applies are discussed. Further, it is unclear if these statutory approaches would be able to deal with constitutional judicial review — assuming the proposed worries about the attribution issue above do apply to the s 75(v) case.[65] If that is the case, one cannot declare outputs of machine algorithms as decisions of the executive by fiat; some theoretical justification for that approach must be provided.

Given the lack of discussion of the illegality issue and the only brief survey of the attribution issue in these recommendations, there is still room for the development of theoretical principles for dealing with the judicial review aspects of automated decision-making.

---

62 See above n 3 and accompanying text.

63 Australian Human Rights Commission (n 58) 78, recommendation 11.

64 Ibid 79.

65 I would like to thank an anonymous reviewer for highlighting this point.

IV Manifestation, Implementation and Equivalent Treatment

Three principles are introduced here as potential candidates for dealing with the thought problem: the manifestation, implementation, and equivalent treatment principles. It will be argued that the manifestation principle only deals with the attribution issue, and is a harder descriptive fit with existing administrative law. Hence, the focus of the article will primarily be on the implementation and equivalent treatment principles.

A *The Manifestation Principle*

In Kerr J's dissent in *Pintarich*, his Honour provided a method for dealing with the attribution issue:

> Its determination [of whether there is a decision] requires close assessment of whether the circumstances in which the conduct said to be, or not to be, a decision arose was within the normal practices of the agency and whether the *manifestation* of that conduct by an overt act would be understood by the world at large as being a decision.[66]

Call this the manifestation principle: if there is a manifestation of conduct that would appear to an ordinary person as a human decision, then it should be considered a decision. The idea is that it is undesirable if the executive could send official notices that appear to be finalised decisions but withdraw them any time after.

His Honour justified this principle as follows:

> It would undermine fundamental principles of administrative law if a decision-maker could renounce as 'not a decision' (and not even a purported decision) something he or she has manifested by an overt act taking the form of a decision simply by asserting there was a distinction between their mental processes and the expression of those mental processes in the overt act.[67]

Unfortunately, Kerr J did not spell out what fundamental principles would be undermined. Two normative suggestions are proposed here. One is a principle of legitimate expectations: a citizen having received a letter that appears to be a legal decision has formed a reasonable expectation that they can act on that legal decision. Another way to justify the manifestation principle is by rule of law-related reasons — eg stability and certainty. If citizens have to double-guess whether a decision was really made, this would reduce the predictability and stability of law.

Regardless of their attractiveness as normative justifications, it is doubtful that these justifications would fit with existing Australian law. First, the High Court has been fairly sceptical as to the doctrine of 'legitimate expectation'.[68] Second, as

66 *Pintarich* (n 1) 49 [52] (emphasis added).

67 Ibid 49 [55].

68 *Minister for Immigration and Border Protection v WZARH* (2015) 256 CLR 326, 334–6 [28]–[30] (Kiefel, Bell and Keane JJ).

Lisa Burton Crawford notes, while the High Court talks about the rule of law as an assumption of the *Australian Constitution*, no theoretically substantive notion of the rule of law — independent of statutory and constitutional text — has yet been used to invalidate governmental action.[69] It might be highlighted that the ground of apprehended bias does take appearances seriously; however, it is justified because the integrity of tribunals and courts is so important that 'even the appearance' of bias would harm it.[70] It is unclear how this justification would translate to non-tribunal executive decision-making cases.

Finally, the manifestation principle only deals with the attribution issue. Given the points above, this article will not focus on the manifestation principle. This is not a criticism of Kerr J; judges are only allowed to address the issue before the court and so it would have been inappropriate for his Honour to have extended it further.

B *The Implementation Principle*

Another way to deal with machine algorithms is what I call the implementation principle: even if machine algorithms do not think, a human decided to implement the machine algorithm, and thus responsibility lies with the human. The relevant humans here would be members of government, not programmers, since it is the government which chooses to implement machine algorithms.

The implementation principle draws inspiration from the command responsibility test in international humanitarian law, the control test in Canadian public law, and vicarious liability in tort law. The command responsibility test states that a commander is responsible for the acts of their subordinates where the commander either 'knew' or 'should have known' that their subordinates were going to commit war crimes.[71] In terms of the Canadian control test, the Supreme Court in *McKinney v University of Guelph* stated that an entity was a government actor for the purposes of the *Canadian Charter of Rights and Freedoms*[72] where it could be shown that the government exercised substantial control over that entity.[73] Further, in tort law, an

---

69 For the High Court's comments, see: *Australian Communist Party v Commonwealth* (1951) 83 CLR 1, 193 (Dixon J); Lisa Burton Crawford, *The Rule of Law and the Australian Constitution* (Federation Press, 2017) 80, 158.

70 *Ebner* (n 53) 345 [7].

71 *Rome Statute of the International Criminal Court*, opened for signature 17 July 1998, 2187 UNTS 3 (entered into force 1 July 2002) art 28(a)(i). See also SC Res 827, UN Doc S/RES/827 (25 May 1993), as amended by SC Res 1877, UN Doc S/RES/1877 (7 July 2009) art 7(3).

72 *Canada Act 1982* (UK) c 11, sch B pt I.

73 [1990] 3 SCR 229. Janina Boughey and Greg Weeks have argued that the control test might be a useful way to think about the 'officer of the Commonwealth' requirement in s 75(v) of the *Australian Constitution*: Janina Boughey and Greg Weeks, '"Officers of the Commonwealth" in the Private Sector: Can the High Court Review Outsourced Exercises of Power?' (2013) 36(1) *University of New South Wales Law Journal* 316, 354–6.

employer can be held responsible for the tortious actions of an employee where the employer themselves might not personally be at fault.[74] In particular, responsibility can be attributed to the employer where the tortious action of the employee was 'committed in the course or scope of employment'.[75]

We can further divide the implementation principle into two subcategories. First, the control subprinciple states that where there was a predictable machine algorithm within the human's control, responsibility is attributable to the human. Second, when it comes to unpredictable algorithms or where the executive did not fully understand the machine algorithm, we invoke a recklessness subprinciple. Even if the human did not predict the effects of the machine algorithm but a potential error was reasonably foreseeable, responsibility is still attributable to the human. It is important to note that on both subprinciples the human is held as responsible.

1 *Tools versus Agents*

There is an important point from which the implementation principle departs from the areas of law discussed above (international humanitarian law, Canadian public law and tort law). In those existing areas of law there is some kind of supervisor–subordinate or principal–agent relationship between the person who committed the illegality and the person who is ultimately held liable. This is an agency model of implementation. One can immediately see the awkwardness of using the agency model — machine algorithms are not really subordinates or agents of the executive.

In contrast, the version of implementation proposed here is a tool model. Under the tool model there is no claim that the machine algorithm is an authorised actor making decisions for the principal. Instead, it establishes a much simpler link of responsibility that if a human either has control over the tools used, or is reckless in using them, while leading to an unlawful act, the human should still be responsible for the use of those tools.

The agency and tool models have different implications for whether a recklessness subprinciple is used (as proposed in Part IV(B)(3)) or whether a strict liability subprinciple should be used instead. On the agency model, the strict liability subprinciple is more appropriate. The debates over strict liability in tort law centre around justifying the liability of a person who is not at fault.[76] Vicarious liability is a form of strict liability since the employers are not personally at fault.[77] Thus, if the

74 *Prince Alfred College Inc v ADC* (2016) 258 CLR 134, 148 [39] (French CJ, Kiefel, Bell, Keane and Nettle JJ) ('*Prince Alfred College*').

75 Ibid 148–9 [40].

76 See: Jules L Coleman, 'The Morality of Strict Tort Liability' (1976) 18(2) *William and Mary Law Review* 259, 269; Gregory C Keating, 'Strict Liability Wrongs' in John Oberdiek (ed), *Philosophical Foundations of the Law of Torts* (Oxford University Press, 2014) 292, 297.

77 See: *Prince Alfred College* (n 74) 148 [39]; Joachim Dietrich and Iain Field, 'Statute and Theories of Vicarious Liability' (2019) 43(2) *Melbourne University Law Review* 515, 516.

implementation principle is seen analogously to vicarious liability, then it should be a strict liability subprinciple rather than a recklessness subprinciple.

On the tool model, however, recklessness seems like a more appropriate standard. Suppose someone turned on a defective oven and had no idea that this would lead to the house burning down. It is less obvious that this person should be held to a strict liability standard. Since under the tool model we view a machine algorithm more as the defective oven than as an employee, recklessness seems to be the more suitable subprinciple.

This is a cursory glance at somewhat complicated theories of responsibility, so I am not committed to the tool model being incompatible with strict liability, but this article will assume that recklessness is the appropriate standard as on the surface it seems more coherent with the tool model.

2 *The Control Subprinciple*

Suppose that a human has control over, and can predict, what a specific machine algorithm will do. In such circumstances, the machine algorithm would be no different from a normal tool. A machine algorithm that is subject to a high degree of control is no different from calculators or Microsoft Excel sheets which operate by algorithm. No one would think that any thorny conceptual issues arise if a tax officer inputs the wrong number into a calculator or uses the wrong function in Excel. While such an error might be considered factual in nature, it would still likely enliven grounds of review with factual aspects — for example, jurisdictional fact errors and unreasonableness.[78] Where machine algorithms are not black boxes, we can detect illegality in many cases (although not in all, as will be discussed in Part V(B)) — ie we can tell how the rules of the algorithm depart from the statutory requirement. Since the decision-makers using them had full knowledge of the output type, we can also attribute responsibility straightforwardly.

Note that the implementation principle can still operate even if the majority in *Pintarich* was correct that decisions require a mental process.[79] The implementation principle attributes the decision-making to humans since humans choose to use the machine algorithm. Justice Kerr and others have criticised the majority in *Pintarich* because the delegate of the Deputy Commissioner had input the numbers into the machine algorithm to initiate the process of sending a letter — and thus the delegate did make a decision.[80] This is in essence a version of implementation; by choosing to use the machine algorithm to calculate some output, the delegate implemented the machine algorithm and should be held responsible.

78 On judicial review of factual errors, see generally Paul Daly, 'Facticity: Judicial Review of Factual Error in Comparative Perspective' in Peter Cane et al (eds), *The Oxford Handbook of Administrative Comparative Law* (Oxford University Press, 2020) 900, 906–7, 911–12.

79 See *Pintarich* (n 1) 67–8 [140]–[143] (Moshinsky and Derrington JJ).

80 See: ibid 49 [53] (Kerr J); Ng and O'Sullivan (n 1) 28.

The control subprinciple gives rise to two types of implementation. The first is system-implementation, where the human decides to set up an automated system and there is no further human intervention (everything else is automated). As an example of system-implementation, consider again the Robodebt system for collecting alleged overpaid social security payments.[81] In this system, the collection of data, requests for more information, and final notifications of debts were automated.[82] In *Prygodicz v Commonwealth [No 2]* ('*Prygodicz*'), there was no human intervention at the point when the final notice of debt was issued.[83] With system-implementation, it is the decision to implement the algorithmic system that is attributed to the government. Given that machine algorithms have no choice but to follow the process every single time, it is the implementer who is in control, even though they do not make every individual determination. Once the decision is attributed to the implementer at the systemic level, then whether something is unlawful is a matter of whether the algorithmic rule departs from the legal rule.

The second type of implementation is application-implementation, where the human uses the output of a machine algorithm in their decision, but it is still the human who makes the final decision. The implementation principle applies to application-implementation quite straightforwardly — the human chooses to use the output of the machine algorithm to assist in decision-making, just as how one might be assisted by a calculator. Hence, this is the human's decision and potential error.

3 *The Recklessness Subprinciple*

The recklessness subprinciple is required when dealing with unpredictable algorithms — if the decision-maker was not sure how the machine algorithm would operate, then they had little control over the machine algorithm. It is thus unclear if the output could be attributed back to those who implemented the machine algorithm. The recklessness subprinciple could also operate where there is a predictable machine algorithm, but the humans who implemented it did not take the time to understand how it would operate.

Under the recklessness subprinciple, addressing the use of unpredictable algorithms or addressing the ignorance of the executive regarding a predictable algorithm is unproblematic. This is because: (1) machine algorithms do not have minds of their own; (2) they were only set up because a human decided to do so; and (3) the human decided to ignore the fact that they were unpredictable. If a parent gave a child a box of matches, it is no excuse for the parent to say that they did not know what the child would have done. The same could be said about giving the ability to change

81 See above nn 19–21 and accompanying text.

82 See *Prygodicz* (n 20) 287–8 [38]–[41]. See also Order of Davies J in *Amato v Commonwealth* (Federal Court of Australia, VID611/2019, 27 November 2019) ('*Amato*').

83 *Prygodicz* (n 20) 287–8 [41]. The initial notice of a potential debt with a request for more information had no legal repercussions. It was the final debt notice under s 1229(1) of the *Social Security Act 1991* (Cth) ('*Social Security Act*') that had the legal effect: see *Prygodicz* (n 20) 287–8 [39]–[41].

or determine legal rights to an unpredictable algorithm. Hence, even in cases where members of the government did not know how a machine algorithm operates, or could not predict its outputs, the recklessness subprinciple still allows responsibility to be attributed where an error was reasonably foreseeable. Of course, not all cases are as simple as a child with a box of matches or cases of machine algorithms with clear risks of harm, and hence the proper application of the principle is to consider how likely harm is reasonably foreseeable in those harder cases.

4 *The Potential Problem with Illegality*

It is here foreshadowed, and discussed more carefully in Part V(B), that the implementation principle is limited in its ability to deal with the illegality issue. On traditional understandings of recklessness, one has to establish both that some event was reasonably foreseeable and that the event was unlawful or undesirable and yet the (liable) person went along with an action that causes the event anyway.[84] The implementation principle itself does not exactly give us a theory of when an act or omission should be a legal error; it only tells us that where such legal errors occur, they can be attributed to humans. This is the same even if one wanted to use a strict liability approach. The strict liability is on the human executive as a principal or supervisor, but it is the unlawful acts of the machine algorithm that are attributed to the human. Strict liability makes that attribution possible but does not provide resources for identifying what is unlawful with the machine algorithm's outputs.

C *The Equivalent Treatment Principle*

An early version of this principle was first stated by Lawrence B Solum in the context of a trustee relationship:

> The focus of the law's inquiry … [where an AI is a trustee] … ought to be on whether AIs can function as trustees. 'Can an AI do the job?' is the question the law should ask. 'Does the AI have an inner mental life?' is simply not a useful question in this context.[85]

The crucial slogan is as follows: what matters is what machines do, not whether they think.[86] Nonetheless, Solum does not develop this principle into a comprehensive theory applied to administrative law, which this article now attempts to do.

---

84 Peter Cane, 'Mens Rea in Tort Law' (2000) 20(4) *Oxford Journal of Legal Studies* 533, 535.

85 Solum (n 1) 1281.

86 This is thus a weaker form of AM Turing's claim that the question of whether machines think is 'too meaningless to deserve discussion': AM Turing, 'Computing Machinery and Intelligence' (1950) 59(236) *Mind* 433, 442. Instead, for Turing, you test what the machine does by observing if it can pass the imitation game (ie it does well on a task): see at 422. This article does not claim that the question of whether machines think is meaningless, simply that it is not useful for administrative law. With some slight modifications, equivalent treatment can also be seen as a weaker

To show this more concretely, suppose that there is a commander that uses drones to attack a certain city. If you lived in that city, would the appropriate response be to stop and ponder if the drone has 'murderous intent'? Would it make a difference if it were human pilots who personally flew planes to attack the city? The answer proposed is 'no' for both questions and the right response is to run away. The underlying logic is that, for the purposes of survival, it does not matter whether a drone really thinks or has murderous intent. Similarly, the gist of the equivalent treatment principle is that it is arbitrary, when pursuing certain goals, to treat machine algorithms differently from a human who acts in a similar manner — both should be treated equivalently.

1 *Attributional and Illegality Treatment*

In administrative law, the principle proposed is that a court should treat the actual human who used a machine algorithm in the same way as a human who had personally executed the rules of the algorithm (ie as if the algorithm was the reasoning of that human). The analogy in the drone case is that it would be arbitrary, for the purposes of survival, to distinguish between a human piloting a plane with a strategy to kill people, from a drone using that same strategy to kill people.

Recall that an algorithm was defined above in Part II in an abstract sense as a sequence of rules (call this 'R'), which in theory can be executed by a human. A machine algorithm refers to the subset of cases where R is executed by a machine. Whether R is executed by a machine or a human, it functions as the reasons for why an output is arrived at. The core argument is that in both cases where a human or a machine executes R, these are functionally equivalent states of affairs for administrative law since in both cases the same output (usually a change, or a refusal to change, legal rights, duties and interests) is produced on the basis of R. As will be further elaborated in Part IV(C)(2), whether one's theory of administrative law focuses on statutory compliance or rights satisfaction, the distinction between a human or a machine executing R is arbitrary. Hence, we should treat the use of R where personally executed by the human in the same way as R being implemented by the human. Note that for extremely complex algorithms, this requires imagining that the person has superhuman abilities to personally execute the algorithm. Part IV(C)(3) explains why this is not a problem for the equivalent treatment principle.

There are two aspects as to how similar treatment can be conceptualised here that correspond to the attribution and illegality issues:

form of functionalism in the philosophy of mind. Functionalism posits that mental states are functional states — that is, the state of being in some complex causal relationship: see generally John Heil, *Philosophy of Mind: A Contemporary Introduction* (Routledge, 1998) 89–104. Equivalent treatment can be seen as a weaker claim: we are not sure if a machine being in the same functional state as a human means that they are both thinking. However, if they are both in that functional state then for the purposes of administrative law there is no difference. For a functionalist account that AI could be capable of thinking, see Christian List, 'Group Agency and Artificial Intelligence' (2021) 34(4) *Philosophy and Technology* 1213.

- attributional treatment: if the human personally executing R made a decision, the human using a machine to execute R should be treated as having made a decision; and
- illegality treatment: if it is unlawful for a human to personally execute R, it should be unlawful for a human to implement R using a machine.

As an illustration of both attributional and illegality treatments, let us return to the Robodebt system. First, the system utilised an 'income averaging' machine algorithm to identify overpayments.[87] Once a person was identified, the system automatically sent a letter to notify the person to update their information and warned that failure to do so might result in a debt.[88] Where no further or sufficient information was provided, the Robodebt system assumed that there had been an overpayment and sent a notice, on behalf of the Secretary, stating that the addressee owed a debt pursuant to the *Social Security Act 1991* (Cth) ('*Social Security Act*').[89] For many people, the averaging machine algorithm did not correctly identify debt under the *Social Security Act*.[90]

As noted in *Amato v Commonwealth* ('*Amato*') the use of the wrong machine algorithm meant the debt notice would be invalid as 'there was no material before the decision-maker capable of supporting the conclusion that a debt had arisen'.[91] Since there was no probative material that indicated a debt was due, the inference that a debt had arisen was irrational which meant that there was no foundation for a penalty and no basis for the notice sent to the alleged debtor.[92]

Neither *Amato* nor *Prygodicz* questioned whether the Robodebt debt notice was a 'decision', presumably because s 6A of the *SSA Act* deems outputs of computer programs to be decisions of the Secretary.[93] However, suppose that s 6A did not exist. There would be three legal issues here. The first two stem from the attribution issue: whether there was a decision and whether it was the decision of the Secretary. The third stems from the illegality issue: whether the use of the machine algorithm was lawful.

---

87 *Prygodicz* (n 20) 287 [38].

88 Ibid 287 [39].

89 Ibid 287–8 [40]–[41]. The issuing of the notice was pursuant to s 1229(1) of the *Social Security Act* (n 83).

90 See *Prygodicz* (n 20) 281 [11]. A debt is owed for an overpayment as set out in s 1223 of the *Social Security Act* (n 83). See also *Prygodicz* (n 20) 281 [8].

91 *Amato* (n 82) 5–6 [8]–[10].

92 Ibid.

93 The debt notice was one of the 'decisions' being challenged in *Amato*: ibid 6 [10.2]. In an *AD(JR) Act* situation, s 1229 of the *Social Security Act* (n 83) is also the appropriate determination to review given that, once the notice is sent, then legal consequences follow: under s 1229(2), the 'debt is due' 28 days after the date of the notice, the debt is recoverable under s 1230C only once the 'debt is due'.

Applied to the Robodebt cases, attributional treatment tells us to imagine that the Secretary personally used the averaging function and personally sent out the letters. If that constituted a decision in this hypothetical case, then it would be arbitrary to treat the Robodebt situation differently. In such a case, the first two issues are answered affirmatively — we should treat it as if the Secretary had made the decision. This type of reasoning also provides the conceptual foundations for provisions like s 6A of the *SSA Act* which Justice Perry queried.[94] On the third issue as to unlawfulness, we compare a machine algorithm sending out debt notices based on a miscalculation and a hypothetical human who sent debt notices on that same miscalculation. Since such a debt notice sent by a human would lack evidence or be irrational, given *Amato*,[95] we should treat the machine notice the same — ie as being invalid as well (which is the illegality treatment).

Note that on both kinds of treatment it is assumed that there is a human actor and so there is some overlap with the implementation principle. Nevertheless, the equivalent treatment principle differs from the implementation principle in two important ways. First, the implementation principle requires a much broader notion of responsibility — it is premised in concepts of control or recklessness. By contrast, the principle of equivalent treatment mainly relies on very simple analogical[96] and (as will be elaborated in Part IV(C)(2)) practical reasoning. This is illustrated by the earlier drone thought experiment. The implementation theorist in that case has to attribute responsibility for drone attacks to the human commander's orders. However, from the equivalent treatment perspective, it is normatively much simpler: we compare whether being killed by a drone is analogous to being killed by a human pilot personally. Second, because the implementation principle is primarily a theory of responsibility rather than what counts as legal error, in some circumstances it will only be able to say that the outputs of machine algorithms are the responsibility of the executive — it cannot explain whether those outputs should be unlawful (this will be shown in Part V). The equivalent treatment principle, as noted above, does contain illegality treatment and so can deal with both the attribution and illegality issues.

It is also noted, speculatively, that the equivalent treatment principle acts as future-proofing for administrative law as well. This is supposing that we do ever reach a stage of development where we have artificial general intelligence ('AGI'), ie artificial intelligence systems able 'to solve a variety of complex problems in a variety of contexts, and to learn to solve new problems that they didnt [sic] know

94 See above nn 42–3 and accompanying text. Although it is noted that the need for a decision was not as crucial in *Amato* (n 82) as the *AD(JR) Act* was not used in that case and *Prygodicz* (n 20) was successfully run as an unjust enrichment case: at 279–80 [3].

95 See *Amato* (n 82) 6 [9]–[10].

96 See generally Cass R Sunstein, 'On Analogical Reasoning' (1993) 106(3) *Harvard Law Review* 741.

about at the time of their creation'.[97] The equivalent treatment principle potentially allows us to treat the AGI as if it were the decision-maker. At that level of sophistication, the AGI is functionally equivalent to a person. There is no need to consider complex questions of legal personality; if an AGI were to carry out illegal actions, those actions should be reviewed just as if that AGI were a human. For administrative law purposes it has done the same thing — changed legal rights and duties in a way inconsistent with law. Nonetheless, since we do not currently have AGIs, this potential situation is not covered in this article.

2 *Does Using a Machine Matter for Administrative Law Goals?*

The insight of the drone hypothetical is that, in certain circumstances, the appropriate kind of reasoning is not metaphysical — metaphysics being the reasoning about the nature of reality such as whether machines think — but rather practical reasoning, ie reasoning as to what actions should be taken in a given scenario.[98] A normative theory of judicial review more naturally falls under a theory of practical reasoning, since 'review' is a kind of institutional action. Seen in this light, as a species of practical or institutional reasoning, a normative theory of review concerns itself with administrative law goals.[99]

This does not mean that, outside of administrative law, it is never relevant to know whether it is a machine or human that is implementing an algorithm for practical reasoning. A crucial part of the drone example above was that the goal of survival made it arbitrary to distinguish between being attacked by a drone as opposed to a human pilot. Contextualised to a different goal, it might make sense to distinguish them (eg determining what mechanical parts need to be ordered for maintenance will depend on whether it is a plane or a drone being used). In order for equivalent treatment to be defended for administrative law, it must thus be shown that for administrative law goals it is arbitrary to distinguish between humans and machines executing certain algorithms. The argument here is that administrative law goals are very likely unaffected by the type of metaphysical questions surrounding machine algorithms — for example, the true nature of what 'thought' is, or what a 'decision' is.

Consider the contrast between rights and formalist approaches to judicial review.[100] While rights-based approaches focus on fundamental rights and liberties as the

---

97 Ben Goertzel and Cassio Pennachin, 'Preface' in Ben Goertzel and Cassio Pennachin (eds), *Artificial General Intelligence* (Springer, 2007) v, vi.

98 See Robert Audi, 'A Theory of Practical Reasoning' (1982) 19(1) *American Philosophical Quarterly* 25, 25.

99 For an example of a system of administrative law goals, see Paul Daly 'Administrative Law: A Values-Based Approach' in John Bell et al (eds), *Public Law Adjudication in Common Law Systems: Process and Substance* (Hart Publishing, 2016) 23.

100 See generally Thomas Poole, 'Between the Devil and the Deep Blue Sea: Administrative Law in an Age of Rights' in Linda Pearson, Carol Harlow and Michael Taggart (eds), *Administrative Law in a Changing State: Essays in Honour of Mark Aronson* (Hart Publishing, 2008) 15.

normative goal of administrative law (leading to general principles legitimate expectations),[101] Thomas Poole suggests that formalist approaches focus on rules rather than general principles, statutory construction, and the sidelining of international law.[102] The Australian version of formalism displays a particular focus on statutory interpretation and ensuring statutory compliance.[103] This emphasis on statutory compliance is presumably justified on democratic grounds: it is the Parliament that creates statutory law, including law conferring power on the executive, and so any identification of executive powers in this context must derive from that statute. In contrast, the rights theorists accept that the statute is important, but it is not the entire picture. Neither rights nor formalist theories take a normative view of the administrative state which requires deep metaphysical commitments.

On the rights theory, as noted above, it is the rights and liberties of those affected by executive decision-making that are crucial. Whether machine algorithms think or have personhood is not pertinent; it is the effects of machine algorithms on the rights and liberties of the populace which are important. Thus, it should not matter whether it is a human or machine algorithm that is affecting those rights and liberties.

Formalist theories are slightly more complex because on such views, a major function of administrative law is to ensure compliance with statutory rules (putting aside non-statutory executive powers).[104] Hence, it appears that if statutory rules do make metaphysical distinctions, administrative law requires compliance with those distinctions. However, we can distinguish between three types of statutory rules that are relevant to judicial review:

- eligibility rules: explain when a review or remedy is available, even if a legal error really has been made. For example, but not limited to, s 3 of the *AD(JR) Act* limiting review to decisions under an enactment;
- substantive power rules: provide the scope of the power of the executive. For example, s 116 of the *Migration Act 1958* (Cth) ('*Migration Act*') outlining the Minister's personal power to cancel visas; and
- methodological rules: explain the method — of judges — by which we can determine if a substantive power rule has been complied with (ie whether there is a legal error). For example, but not limited to, ss 5, 6 and 7 of the *AD(JR) Act*.

---

101 For one example, see David Dyzenhaus, Murray Hunt and Michael Taggart, 'The Principle of Legality in Administrative Law: Internationalisation as Constitutionalisation' (2001) 1(1) *Oxford University Commonwealth Law Journal* 5, 5–7, 10.

102 Ibid 25.

103 See: Elizabeth Fisher, '"Jurisdictional" Facts and "Hot" Facts: Legal Formalism, Legal Pluralism, and the Nature of Australian Administrative Law' (2015) 38(3) *Melbourne University Law Review* 968, 972; Matthew Groves, 'Substantive Legitimate Expectations in Australian Administrative Law' (2008) 32(2) *Melbourne University Law Review* 470, 500.

104 See Christopher Forsyth, 'Showing the Fly the Way Out of the Flybottle: The Value of Formalism and Conceptual Reasoning in Administrative Law' (2007) 66(2) *Cambridge Law Journal* 325, 328.

Using this tripartite distinction, statutory compliance for a formalist theorist requires compliance with substantive power rules. It does not mean that no one should ever change any rules about eligibility or methodology — if that were so, then statutes like the *AD(JR) Act* or the *Administrative Law Act 1978* (Vic) should never be amended.

Note that the rules above are statutory rules and not common law review rules since formalism as defined here is about statutory compliance. So, if the common law makes metaphysical distinctions, there are no logically tricky issues with a formalist saying they have to be changed (whereas they arise with statutory review rules since formalism is about statutory compliance). Additionally, it is accepted that there are of course borderline or hard cases between eligibility, substance, and methodology;[105] but such cases also abound in administrative law between other concepts like questions of fact and law, and between jurisdictional and non-jurisdictional error.[106] As noted by Murray Gleeson: '[t]wilight does not invalidate the distinction between night and day'.[107]

A formalist theory has nothing to say about the normative desirability of substantive power rules — eg whether or not Ministers should have wide-ranging personal visa cancellation powers is a matter for political philosophy or migration theory. Similarly, if substantive power rules do make distinctions about certain powers having to be exercised by humans, then a formalist will be committed to abiding by that. For example, suppose that s 116(1)(e) of the *Migration Act* was amended such that the Minister can only cancel a visa once they have consulted with a predictive machine algorithm for a risk assessment. A formalist will say that a judge reviewing s 116(1)(e) will require the Minister to have consulted such a machine algorithm as that is what the statute states. Whether or not that is the best way to cancel a visa is a policy question about immigration law — and not a judicial review question. In general, a formalist is unlikely to have anything to say about substantive power rules, as such rules generally fall under what the policy issues are surrounding executive use of machine algorithms.

On the other hand, a formalist theory would have a view on the desirability of eligibility and methodological rules — on the basis that they should maximise or satisfy compliance with substantive power rules. To give an example, suppose the *AD(JR) Act* was amended to add a ground of review such that any use of a machine algorithm would render a decision invalid unless explicitly allowed by the statute. It would be legitimate for a formalist to object to the addition of this ground as

105 For an example of some of the difficulties, see Solum's discussion of the difference between procedure and substance: Lawrence B Solum, 'Procedural Justice' (2004) 78(1) *Southern California Law Review* 181, 192–206.

106 Stephen Gageler, 'What is a Question of Law' (2014) 43(2) *Australian Tax Review* 68, 69.

107 Murray Gleeson, 'Judicial Legitimacy' (2000) 20(1) *Australian Bar Review* 4, 11.

a general rule — just as a formalist might object to liberal uses of the unreasonableness ground.[108] Methodologically, since formalists in Australia tend to think of judicial review in terms of statutory interpretation, it is really an examination of the statute on a case-by-case basis which would determine whether the statute would allow the use of a machine algorithm. For a formalist, a blanket ban would be inconsistent with proper rules of statutory construction.

Essentially, the normativity of eligibility and methodological rules for a formalist do not depend on deep metaphysical commitments about the nature of the mind. They depend on normative commitments to correct interpretive method and justiciability. Hence, illegality treatment is justified as long as the substantive power rule does not prescribe that the decision must be made only by certain entities. When it comes to eligibility, subject to complying with constitutional principles, a formalist would be concerned with satisfying or maximising statutory compliance while minimising legal error. As noted in Part III, the thought problem typically interferes with review, even where genuine legal errors may exist (having established that illegality treatment can assist to identify these errors). Consequently, for a formalist it should not matter if it is a machine or human executing an algorithm — if it leads to non-compliance, then review should be allowed. Accordingly, attributional treatment is also justified.

Of course, it is possible that legal systems will have a mix of formalist and rights-based elements.[109] However, even where combined there is unlikely to be a need for metaphysical theorising. The rights elements of the system will focus on promoting individual rights which will be balanced by formalism's emphasis on ensuring statutory compliance. The combination of the two does not seem to require an answer to questions in the philosophy of mind and consciousness.

3 *Can Humans Really Reason like Machines?*

Equivalent treatment asks us to imagine a human that can execute an algorithm R without machine assistance. It might be queried whether this is even possible. In cases of very complex algorithms, the answer is no. Hence, the comparisons in this article do sometimes require us to create a counterfactual with a superhuman who is immortal or can carry out the steps of an algorithm R in a much quicker time frame.

This does not undermine the comparisons made. The real point of the arguments in Part IV(C)(3) above is that it is irrelevant in administrative law whether a machine, individual human, group, alien or superhuman executes the algorithm.[110] The real inquiries are what rules R are being executed, regardless of the entity executing it, and what is R's function in administering the law. It is the function of the rules and

108 For an example of such an argument, see Timothy Endicott, 'Why Proportionality Is Not a General Ground of Judicial Review' (2020) 1(1) *Keele Law Review* 1, 4–6.

109 The *Charter of Human Rights and Responsibilities Act 2006* (Vic) ss 38 and 39 presents one such example.

110 See above n 86 regarding analogies with functionalism in philosophy of mind.

processes in place that is the primary object of inquiry. Nonetheless, for pragmatic reasons, this article uses the lawfulness of human action as the basis from which legality is assessed. Administrative courts and lawyers have long conceptualised what kinds of decision-making are lawful and unlawful. There is thus no need to reinvent administrative law.

More precisely, neither statutes nor the common law typically set out different administrative law rules based on the capabilities of those exercising statutory powers. For example, the availability of grounds of review depends on whether the decision-maker properly followed the statutory rules rather than the intelligence or physical strength of the decision-maker. Hence, a superhuman or genius employed by the Australian Public Service should follow the same administrative laws that a normal human would. Thus, comparing a machine algorithm with a superhuman does not distort the comparison since the superhuman plays the same role or function that a human would in decision-making (and in equivalent treatment it is their functions that are compared).

V Resolving the Thought Problem

The article now turns to consider how the implementation and equivalent treatment principles can apply to the thought problem. As noted above, the manifestation principle is limited in scope and so is not discussed further.

A *The Attribution Issue*

We can now illustrate how the attribution issue can be resolved by reference to *Pintarich*. Under the implementation principle, it is the human who decides to implement a machine algorithm. Thus, the thought problem creates no special difficulties with determining who makes a decision given that the humans implementing algorithms are capable of thought. With *Pintarich*, the letter was the result of a decision by the delegate to input the information into the computer.[111] Hence, the delegate made the decision.

Under the equivalent treatment principle, it can be argued that the automated letter in *Pintarich* was the decision of the human delegate using attributional treatment. Contrast what happened in *Pintarich* with a hypothetical where a human, without machine assistance but using the algorithmic rules of the machine algorithm, calculated the amount owed and sent the letter. Where a human had personally made calculations and sent the letter, there would be no doubt that a decision was made. As per attributional treatment, it would thus be arbitrary to treat it as if there were no decision in this case simply because the sending of the decision was automated.

111 See above nn 32, 79 and accompanying text. See also *Pintarich* (n 1) 50 [61]–[63].

B *The Illegality Issue: Mental Element Grounds*

As noted above, when examining the exercise of statutory power, it is often assumed that lawful exercises can only occur if certain mental elements are present or excluded. This can be seen in how certain grounds of review are only established once the presence or absence of such mental elements is shown. Four grounds will be considered here: (1) relevant and irrelevant considerations; (2) actual and apprehended bias; (3) improper purpose; and (4) bad faith. These grounds will be referred to as mental element grounds. The implementation and equivalent treatment principles discussed do not guarantee that all mental element grounds can be applied to machine algorithms, just as we should not be surprised that certain grounds do not operate in certain contexts. However, the approach below provides a principled way to distinguish which mental element grounds are suitable and which are not.

This article acknowledges that formalists would insist that the content of statutes determines legality.[112] Nonetheless, the claim that there is a category of mental element grounds would be consistent with such formalism; it is a claim that statutes can possess features that seemingly require certain mental states to be present or excluded when the power is lawfully exercised. For example, statutes may require certain considerations to be ignored or taken into account.[113] The question of whether mental element grounds are applicable to machine algorithms is thus a question of whether the typical features of statutes do necessarily require or exclude mental states when a power is exercised under those statutes. Where existing grounds can be applied to machine algorithms, this means certain types of exercises of power under certain types of statutes do not require the existence or exclusion of certain mental states to be lawful.

A general solution will be provided for the various mental state grounds mentioned above, but to illustrate let us first examine the specific case of grounds relating to irrelevant and relevant considerations. As indicated by Bateman, in *Tickner v Chapman*, 'consideration' is defined as an 'active intellectual process' which is a mental state (call this 'M').[114] Suppose that a law requires a work visa to be granted as long as the person has provided a skills assessment and there is a requirement that the criminality of the applicant must not be taken into account. Further suppose that the government intends to implement the following rule (call this 'R*') — if the candidate has provided a skills assessment and has no criminal record:

- then grant the work visa; or else
- deny the work visa.

R* above can be executed either by a human or machine. Notice that in R* there is an irrelevant step which is the condition that the applicant has no criminal record (the term 'step' is used to be neutral between being part of a mental process or a

112 See above n 104.

113 For an example of irrelevant considerations that are explicit, see the *Freedom of Information Act* (1982) (Cth) s 11B(4),

114 Bateman (n 1) 523; *Tickner v Chapman* (n 50) 462.

machine algorithm). The point of this hypothetical is that whether a machine or human executes R*, it will result in the same outcome (call this 'O') — work visas are denied because of an irrelevant step.

To resolve this issue on the implementation principle, consider the distinction between mental processes M and outcomes O. Where the executive is aware that the machine algorithm will result in visas being denied because of the irrelevant step, (outcome *O*), the implementation principle attributes that step to the executive since they implemented the machine algorithm with full knowledge of that outcome (the mental state of consideration *M* is attributed to the human). This constitutes an active intellectual process. If the executive is ignorant of the irrelevant step, however, the implementation principle stays silent. Recklessness does not quite help. As foreshadowed in Part IV, the foreseeable outcome that occurred must also be unlawful or undesirable in order for recklessness to apply. The implementation principle can tell us that the executive is responsible for a foreseeable outcome (ie that an irrelevant step is involved), but it does not tell us whether the presence or absence of an irrelevant step results in an unlawful outcome. Hence, the implementation principle does not provide resources to deal with cases of ignorance.

On the equivalent treatment principle, the question is whether the outcome of denying a visa due to an irrelevant step with no mental process is functionally different from the same outcome which is a result of an active mental process. Illegality treatment provides an answer here: imagine a human executing R* personally and whether this would be an unlawful decision. It is clear that a human who followed R* above would be making an irrelevant consideration error. Hence, it would be arbitrary, for the reasons discussed in Part IV(C), to treat the human who uses a machine that executes algorithm R* with the irrelevant step differently from a human who personally executes R* — both should be unlawful.

The above solutions can be generalised. For the implementation principle and grounds which require a mental element in order to lawfully result in some outcome, consider the machine algorithm equivalent that does not have M but results in the exact same O. If there is awareness by the human implementers that the machine algorithm will lead to O, then M can be attributed to the human implementer. On the equivalent treatment principle, we use illegality treatment; consider the algorithm R that is used by a machine as opposed to a human. If in the human case it would be illegal to use R to achieve O, then there are no normative reasons why we should not take it that it is illegal for a machine to use R to achieve O.

Let us consider how this can be utilised for actual and apprehended bias. With actual bias on the implementation principle, we might distinguish the actual mental state of being impartial M with the outcome of some kind of pattern of inequality O (it has been well-documented that machine algorithms can produce unequal outcomes).[115]

115 See, eg: Jon Kleinberg et al 'Discrimination in the Age of Algorithms' (2018) 10(1) *Journal of Legal Analysis* 113; Sourdin (n 3) 72–8. Although Kleinberg et al (n 115) also make the point that using machine algorithms can help to minimise discrimination since their rules are always discoverable: at 116.

On the implementation principle, where there is awareness that the machine algorithm would lead to this inequality O and the executive implemented the machine algorithm anyway, then arguably the executive was actually biased. With equivalent treatment, suppose some algorithm R leads to a pattern of inequality O. The question is whether this R, if personally executed by a human, would count as the state of bias. There can be cases where the answer would be in the affirmative. For example, consider an algorithm that never granted licences to people of a certain race (when race has nothing to do with the licence). Where the human executing this algorithm was aware that following the algorithm would lead to this outcome, functionally this is equivalent to a human directly trying to induce such discriminatory outcomes themselves. It is even arguable that if the human was not aware of the unequal O that this would still count as actual bias. As long as the human used the defective R which does not grant licences to certain racial ethnicities, this is functionally equivalent to a human who sat down and was actively discriminatory against those people (think of R as functionally playing the role of the mental state of the human). Hence, the decision of the human personally executing the algorithm would constitute actual bias and as such the human using a machine algorithm R should be similarly unlawful.

In Australia, apprehended bias occurs where it appears as if 'the decision-maker *might not* have brought an impartial mind to making the decision'.[116] It is unclear if this means the decision-maker must be capable of biased thought (ie in principle they are a thinking agent) even if they are not actually biased. If the law does not require an agent with a mind, then it turns out the apprehended bias rule is not a ground requiring a mental state to be present. If the law does require that the agent be capable of being biased, then the solution above can be applied. With the implementation principle, if the executive is aware that the machine algorithm's use would appear like it was the action of a biased agent and implemented it anyway, then the apprehended bias should be attributed to the executive. For equivalent treatment, we suppose a human used the same algorithm as the machine and consider whether it would breach apprehended bias. In cases where a human personally using the exact same algorithm would appear to be biased, then we should treat it as if the human using the machine algorithm is also acting unlawfully.

Lastly, consider grounds like improper purpose and bad faith. The implementation principle does allow for the grounds of improper purpose and bad faith where the executive is aware of how the machine algorithm operates. If the executive implemented the system for an improper purpose, there was an improper purpose for the output of that system. With bad faith, suppose a machine algorithm with bad programming and arbitrary outcomes was used and the executive was aware of these faults in setting up this machine algorithm. It can be argued that the executive here acted in bad faith. Equivalent treatment allows for a similar strategy. If the human intentionally uses a machine algorithm for an improper purpose or uses it maliciously, that is functionally equivalent to a human who personally implemented

116 *Hot Holdings Pty Ltd v Creasy* (2002) 210 CLR 438, 459 [68] (McHugh J) (emphasis added). See also Lim (n 1) 37–8.

the algorithm for an improper purpose or in bad faith. Hence, it is unlawful in both cases. It appears, however, that in cases of both implementation and equivalent treatment, that if the human is unaware of the outcomes of the machine algorithm (including with unpredictable algorithms), then it is unlikely improper purpose or bad faith can be established.

VI Concluding Remarks: Fruitfulness and Descriptive Fit

In this article, three ways of thinking about the judicial review of administrative machine algorithms have been provided: (1) the manifestation principle; (2) the implementation principle; and (3) the equivalent treatment principle. The manifestation principle is limited only to resolving the attribution issue. The implementation principle is perhaps the simplest and most intuitive to apply but has limitations when it comes to the illegality issue — it does not attribute responsibility in cases of ignorance (see above Part V(B)). The equivalent treatment principle is somewhat more complicated but has several benefits. First, it does not suffer from the aforementioned incompleteness of the manifestation and implementation principles. Second, the equivalent treatment principle does not require a thick theory of legal or moral responsibility,[117] unlike the implementation principle. While this article does not take a hard line on the appropriate principle, it is proposed that these benefits put the equivalent treatment principle in front.

This article has introduced these principles as normative ones and has largely left the question of their descriptive fit with existing laws untouched. At the very least, even if they do not quite fit current administrative law doctrines, they present a principled method for modifying the law of judicial review. Nonetheless it is arguable that some aspects of the implementation and equivalent treatment principles can be applied even now (as noted in Part IV(A), it is harder to argue this for the manifestation principle). At common law, differentiation is a fairly orthodox method: rules are not necessarily applied the same way with new factual situations. The use of machine algorithms presents a very different factual situation from the old administrative system and so differentiation might possibly permit some of the principles here. The implementation principle attributes responsibility to humans and so does not change concepts radically (except perhaps with the introduction of the recklessness subprinciple). Similarly, the equivalent treatment principle introduces a focus on administrative functions, but the point is to apply human doctrines to functionally equivalent uses of machine algorithms. Hence, it does not require drastic changes either since the frame of reference is always existing human doctrines. With the *AD(JR) Act* and constitutional review, both implementation and manifestation principles attribute either responsibility of the machine algorithm or its rules to humans and so some of the review obstacles might be surmountable as a human is still involved in the decision-making process. Lastly, as noted above in Part III(A),

117 We might say that one theory is thicker than another where the theory has more propositions to justify or explain the kind of moral or legal phenomena it is concerned with.

s 75(v) of the *Australian Constitution* requires an 'officer of the Commonwealth' and while this may prevent constitutional review, s 75(iii) is a potential alternative as it does not require humans in the process.

This is a rather limited comment on the descriptive fit of the implementation and equivalent treatment principles, but it gives some reason to think they are not completely foreign to Australian administrative law.