Public Trust in Digital Governance:
The Role of Artificial Moral Agents in Delivering Ethical Public Services

Amanda Dahl
Leverhulme Centre for the Future of Intelligence
University of Cambridge

March 2025

Abstract

What Artificial Moral Agents, if any, should we be deploying into the UK Civil Service? In this paper, I consider the potential of AI to enhance public service delivery amid rising resource constraints and bureaucratic complexity. I explore both the promise and risk of adopting AMAs in public service, especially in contexts where empathy and contextual understanding are crucial for administrative fairness. In the first section, I define the parameters of artificial moral agency and identify the spectrum of AMAs, with focus on those that pose considerable risks when applied in the public sector. In the second section, I develop a theory that existing civil service principles should guide ethical considerations, help us to constrain the negative implications of AMA deployment and create a balance between human discretion and artificial agency in public services. Finally, I respond to the question of what AMAs should be deployed in the UK Civil Service by proposing a phased approach to deployment that prioritises low-risk implementations before progressing to more ambitious applications, with a call to foster a collaborative human-AI partnership that upholds public trust in government services.

The total word count of this document is 4,928 words, excluding the abstract, tables, and references.

Introduction

This paper considers the development and deployment of artificial moral agents (AMAs), systems capable of moral reasoning and ethical decision-making, within the UK civil service. Faced with resource constraints and complex bureaucracies, UK public services are increasingly looking to artificial intelligence (AI) to improve efficiency and address systemic issues (Mason, 2025). It examines and categorises the inherent risks involved in ceding aspects of human judgement to machines within public service delivery. The analysis aims to determine the appropriate balance between human discretion and artificial agency in AI-enhanced public services and advocates for a gradual, risk-informed approach to operationalisation, which is grounded in civil service principles. Ultimately, it calls for a new type of collaborative partnership between human civil servants and their digital counterparts in order to maintain public trust.

While the existing conversation around AMAs is broadly philosophical, this paper aims to make several contributions by applying elements of philosophy, sociology and public administration in a multidisciplinary approach to the specific domain of artificial moral agency in public service delivery within the UK civil service. In the first half, I look at the technology itself, with the goal of narrowing down the spectrum of AMAs to those with the riskiest implications if deployed. For the second half, I consider the socio-technical impact of the deployment of AMAs in public service, examining existing bureaucratic ethical obligations in higher-risk areas where administrative fairness and accountability are critical. Finally, I propose there is a need to cultivate a new form of relationship between human civil servants and digital "colleagues", fostering competent and respectful collaboration in service of the citizen. The paper concludes with a call for a systematic approach to determine the appropriate balance between human and artificial moral agency in the UK civil service.

Context of AI in Public Service

In this essay, a few terms are used interchangeably, especially when it comes to describing “public service”, “digital public service” and “civil service”. Within the context of the digital public services referred to, it should be assumed to include services which are provided at all levels of central and local government, as well as arms length bodies, and to encompass the “the major dimensions of public policy –

levels of benefits, categories of eligibility, nature of rules, regulations and services" (Lipsky, 1980, p.14).

Digital public services in the UK include the online delivery of government services and information by local councils, government departments, and public sector organisations through digital channels such as websites, mobile apps, and online portals. Such services cover areas like tax filing, healthcare, education, and social welfare benefits, and have been an iteratively improving part of the UK public landscape since the late 1990s (Eaves & Goldberg, 2017). These services use digital technologies to facilitate interaction with citizens and businesses, providing access to information, applications, and transactions and are generally built with the intention to offer improved citizen experience through process automation, increased transparency, improved accessibility and greater inclusivity.

It is important to note the context and gravity of discussing AI in UK public service, given the social and political environment in which it is rapidly evolving. Public servants face many challenges in delivering digital services to citizens, including limited resources, outdated technology, obsolete policies, and cumbersome bureaucracy. In response, AI is often viewed as a potential silver bullet, capable of transforming government functions, magically erasing complexity and cruft, employing data-driven logic and reasoning to create a higher functioning society. This optimistic, yet techno-deterministic view is echoed by politicians who advocate for integrating AI into government operations to enhance efficiency (Booth, 2025), although historically, a gap often exists between policy intent and implementation (Hudson, Hunter & Peckham, 2019).

In order to achieve the best possible future for the deployment of AI in government, it is essential to consider the different types of AI which are being deployed in various contexts across public services, to understand which applications of the technologies carry far greater risk than others if granted moral agency and to reflect on the impact on human lives and livelihoods which may result from such replacement. Because the political zeal for AI deployment is so strong, there is urgency for the UK civil service to determine an AMA adoption strategy that accounts for varying degrees of risk and adheres to core civil service principles.

Section 1: Definition and Spectrum of Artificial Moral Agency

Having established the political context of enthusiasm for AI-enhanced public services, it is important to define what we mean by artificial moral agency, as this foundational understanding will guide our exploration of its application. First, agency: Joanna Bryson defines an agent along classic scientific lines, situating it as an artificial entity that is an “actor, living in and changing the world.” (Bryson, 2009, p. 1). And next, moral agency: Haskar (1998), in his philosophical explorations, defines a moral agent as an individual who possesses the capacity to make ethical decisions and is accountable for their actions.

Taking the definition further, Formosa and Ryan (2021) specify that an artificial moral agent is “a bot that can take in environmental inputs (interactivity), make ethical judgments on its own (autonomy), and act on those ethical judgments in response to complex and novel situations (adaptability) without real-time human input” (p.2). And when we speak about alignment, Cave, et al. (2019) define ethically aligned AMAs as “those whose behaviour adequately preserves, and ideally furthers, the interests and values of the relevant stakeholders in a given context” (p.3).

The latter definitions would best be ascribed to an AMA at the height of its potential capacity, with full capabilities to understand context, situation and the overall human condition. However, not all AIs are equal in their capacities, and as yet, Artificial General Intelligence (AGI) has not been achieved, meaning that Formosa and Ryan’s definition of an AMA with sufficient complexity to act entirely autonomously in ethical situations is not yet possible beyond the theoretical. This doesn’t mean we should exclude the concept of AGIs from our consideration of artificial moral agency. Rather, as Peter Asaro suggests, artificial moral agency is better thought of as a continuum, a pluralistic moral space where a “range of different systems, with different levels of sophistication” occupy a variety of agentic roles (Asaro, 2006, p.11), merely meaning that artificial moral agency resists binary classification and is rapidly evolving.

A spectrum of artificial moral agency

This section explores five distinct levels of artificial moral agency, from 'amoral machines' to fully autonomous agents, to uncover implications for the deployment of AI in civil service roles. Artificial moral agency spans a broad spectrum, and James Moor (2006) classified AMAs into levels of ability, based on a combination of capabilities. Asaro, Formosa and Ryan suggest five levels of AMA, with “amoral” machines, such as a Roomba, at Level 1. Machines which are programmed in a top-down, rule-based manner for implicit ethical behaviour, such as ATM machines, are at Level 2 and are considered “robots with moral significance” (i.e. only dispense money to the owner of the card). At Level 3 there is a split, where artificial “moral intelligence”, narrow and likely domain-specific, yet explicitly ethical, is categorised as Level 3a. Agents with “dynamic moral intelligence”, capable of broad, explicitly ethical and deep moral agency are placed at Level 3b. The fully autonomous artificial moral agents of Formosa and Ryan fall into Level 4 (Asaro, 2006, Formosa and Ryan, 2021). For ease of conceptualisation, Table 1 is included with a breakdown of the levels and classifications of AMAs.

Table 1: Levels of Artificial Moral Agency
This table is based on the works of Asaro (2008), Cave et al. (2019), Formosa and Ryan (2021), and Moor (2006), outlining the classifications and descriptions of different levels of artificial moral agency relevant to near-term civil service applications.

Level	Classification	Agency	Description
1	Amoral Machines	Limited or none	Examples include machines like a Roomba or industrial robot, which operate under pre-set conditions without any ethical considerations.
2	Machines with Moral Significance	Implicit	Machines programmed with top-down, deontological ethical behaviour (e.g., ATM machines, emergency response systems that prioritise distribution of resources and aid).
3a	Narrow Artificial	Explicit, Narrow, Partial	AI with domain-specific moral reasoning capacity, and yet which is explicitly ethical; these agents can make basic moral decisions

	Moral Intelligence		within a limited context such as judicial sentencing decisions or smart traffic management systems
3b	Dynamic Moral Intelligence	Explicit, Broad, Partial	Capable of broad, explicitly ethical, and deep moral agency; these agents can adapt and apply ethical reasoning more flexibly. This includes the new generation of reasoning models such as OpenAi's o1 and o3 series. This could include systems to support government decision-makers by analysing potential policies through an ethical lens, considering public welfare, justice, and sustainability in real-time.
4	Fully Autonomous Artificial Moral Agents	Explicit, Broad, Full	Agents that can operate independently with full moral agency; not yet existing in a meaningful way (in March 2025). Examples of future systems might include systems for autonomous warfare or AI for social service, with agents that autonomously implement and adapt behaviour, considering the ethical implications of prioritising resources and adjusting to changing needs.

Narrowing the scope of the spectrum

Both the upper and lower boundaries of the spectrum of agency can largely be excluded from our focus on the consequences of ceding moral agency to AMAs, but for different reasons. On the low end of the spectrum, decisions have already been taken to deploy the technology. On the high end, the technology still may be a few years away.

Level 1 and 2 AMAs (lower level or narrow artificial intelligences) are already being deployed in UK public service. For example, Redbox is a large language model (LLM) which assists with generating text such as Ministerial briefings, providing suggestions based on context and handling Official Sensitive classified documents in a secure environment (Government Digital Service, 2024). Parlex is an AI tool which

forecasts for Ministers and policy makers the reactions of Parliament to proposed policies, and the Consult tool parses the vast datasets gathered during public consultations to quickly provide insights for more effective policymaking (Government Digital Service, 2024). While these tools operate under specific algorithms and cannot independently make moral decisions, they could still be classified as Level 2, having implicit moral agency and “moral significance”. This level involves systems that have some autonomy in decision-making processes but continue to rely heavily on human guidance and ethical frameworks. Redbox and other LLM-based technologies are being deployed across government with a broadly positive reception, in part because they can be seen as mere tools to augment human capabilities and in no way attempt to usurp agency from their human users (Elgot & Booth, 2025).

Conversely, AMAs at Levels 3b and 4 are not yet in existence in any meaningful way, and thus are not relevant to discussions of near-term operationalisation of AMAs. I will, however, briefly return to this topic when I touch on the future of AMAs in public service later in this paper. It is therefore the Level 3a AMAs which I would like to focus on in this discussion.

Categorising the Domain

The next section will categorise the various AI systems currently under development in the UK civil service, recognising the differing characteristics of each. Based on my experience within the UK government, there are over 150 different AI solutions currently in various stages of the development pipeline, many of which may come close to Level 3a agency, in that they are domain-specific, explicit moral agents, capable of making autonomous decisions if granted the remit to do so. These include automated threat and anomaly detection for police forces, AI for university assessments and marking, and an AI-powered solution that aims to transcribe probation interviews, identifying emotional cues (e.g., aggression, sadness, joy) to highlight them to human probation officers (DSIT, 2025). It is this category of AMA which forces the question of how we might make the transition from amoral AI which is straightforward to classify as a tool (Redbox, Parlex, Consult), to something which

is more complicated to classify, and begs questions as to how much autonomy is appropriate for an artificial moral agent in the context of the civil service.

A recent review by the think tank Reform identified that AI is being deployed in the UK government across five categories: (1) business planning and demand prediction (such as A&E waiting times), (2) assessment streamlining (determining eligibility for Universal Credit claims), (3) process automation (checking evidence documents), (4) chatbots and automated call centres, (5) translation/transcription and coding co-pilots (Hill & Eke 2024). Each of these categories, by their nature, carries a different level of complexity and risk due to the degree of moral agency required of the AI system in that particular context. At the moment, across all categories there is a consistent level of agency, at Level 2. However, there are indications of Level 3a in the development pipeline (DSIT, 2025).

Table 2: Types of AI Operationalised
Table 2 illustrates the classification of various types of AI systems currently deployed within the UK government, correlating them with their respective levels of moral agency, as per Table 1. Building on the classifications identified by Hill & Eke (2024), it highlights that all listed AI categories are aligned with Level 2: Machines with Moral Significance, indicating that they operate under predefined ethical guidelines and require significant human oversight in decision-making processes.

AI Type in Development	**Level of Moral Agency**	**Description**
Business planning and demand prediction (e.g., A&E waiting times)	Level 2: Machines with Moral Significance	Systems that forecast based on ethical considerations but rely on human judgement.
Assessment streamlining (eligibility for Universal Credit claims)	Level 2: Machines with Moral Significance	Similar to ATM machines; follow programmed ethical guidelines predominantly based on human input.
Process automation (checking evidence documents)	Level 2: Machines with Moral Significance	Execute processes based on algorithms without moral decision-making capabilities.

Chatbots and automated call centres	Level 2: Machines with Moral Significance	Provide information and assistance based on programmed ethical frameworks but lack autonomous agency.
Translation/transcription and coding co-pilots	Level 2: Machines with Moral Significance	These tools assist users but operate within pre-set ethical parameters and guidelines.

Section 2: Human Principles and Ethical Guidelines

In this section we move from categorising the technology itself to understanding its significance to existing human systems. Now that we have narrowed the scope of this discussion to the middle Level 3a AMAs, how do we determine which of the Reform categories of AI will carry the greatest complexity once at Level 3a and, therefore, risk to public trust? It is helpful to consider that there are normative structures already in place to manage the complexity of *human* moral agency required to function successfully across those five categories of public service, and that these structures not only provide us with a framework for possible deployment of AMAs, but actually (perhaps usefully) constrain this deployment as well, by nature of the longstanding history and well-established principles of UK government bureaucracy.

Many of the categories identified by Reform include systems where experienced civil servants hold a position of responsibility, accountability and moral agency for decision making in complex and unpredictable circumstances, situations in which a citizen's well-being hangs in the balance. The civil servants involved in these decisions are expected to follow a code of practice, based on principles articulated by Lord Nolan, namely: honesty, integrity, objectivity, accountability, selflessness, openness and leadership (Committee on Standards in Public Life,1995). They are character traits which civil servants are expected to cultivate, conducting themselves in a neutral manner, with the intention of ensuring public expectations are met. In 2005, these principles were condensed into the Civil Service Code, setting out core values of "integrity, honesty, objectivity and impartiality" (Constitutional Reform and Governance Act, 2010, p.2). Perhaps understandably, the characteristic of empathy

is not included in the code itself. However, while not explicitly defined, the values and principles enshrined in the Code presuppose an understanding of empathy as a foundational element in fostering effective and ethical public service.

From a philosophical perspective, the Civil Service Code is an interesting example of an attempt to merge modern public administration with classical ethical frameworks. On the surface, it appears to be a top-down, rule-based deontological system, with an emphasis on execution of duties, where the actions of civil servants are judged by neutral adherence to policy. However, the initial appearance of the system as deontological, with clear rules and duties is actually only part of the equation. The other aim of the Civil Service Code is to operationalise and institutionalise Aristotelian virtue ethics within this deontological framework, creating a hybrid approach.

Thick vs Thin rules

This hybrid approach can be examined by looking at the hierarchy of the civil service. The Nolan principles came about for a reason, in reaction to scandals involving bias and corruption, universal issues in human hierarchical systems, which are not unique to the civil service. In many ways, the challenges of integrating artificial moral agents into public service reflect historical dynamics around the division of labour. The history of rules can be a helpful way into understanding the distribution of moral agency across human systems. As historian Lorraine Daston puts it, “it is an ancient philosophical problem, the mismatch between universals and particulars. Usually we have to tweak a rule in order to make it fit the case at hand precisely” (Taylor, 2022, 16:00). Daston examines the distinction between "thick" and "thin" rules, highlighting their different applications and implications in scientific and ethical practices.

Thick rules are those that contain a high level of contextual understanding and moral considerations, where a “thick rule is one which foresees the variability under which rules will be applied” (Taylor, 2022, 16:07). These rules are characterised by their adaptability and recognition of the complexities inherent in human behaviour and scientific inquiry. They are not prescriptive directives but instead guidelines that take into account the intricate realities of specific situations. This allows for a more

nuanced approach to decision-making, where contextual factors can significantly shape outcomes.

Conversely, thin rules are described as straightforward, general, and easily applicable across many situations. Such rules provide clear guidance with minimal complexity, facilitating uniformity and predictability in practice. While they promote efficiency and consistency, their simplicity can be a limitation in contexts that require a deeper understanding of moral or contextual nuances.

In addition to calling out the distinction between rules suited for universals and particulars, Daston also calls our attention to Gaspard de Prony's calculation project during the French Revolution (Daston, 2017). Inspired by Adam Smith's observations on pin factories, Prony organised a strict pyramidal hierarchy to calculate logarithm tables, with leading mathematicians at the top developing formulas, skilled "algebraists" translating them into numerical procedures, and unskilled workers at the bottom performing the actual calculations. This pyramid structure anticipated debates around automating intellectual labour, foreshadowed the eventual development of computing machines, and can be seen in modern bureaucracies like the civil service.

The UK civil service hierarchy mirrors Prony's pyramid structure, with roles and responsibilities divided across different grade levels. Theoretically, lower grades operate under "thinner" or less flexible rules, while senior civil servants are expected to have capacity to operate in a world of “thicker” rules with greater ambiguity. However, consider the core values of the Civil Service Code: integrity, honesty, objectivity and impartiality. These principles focus on character rather than rules. There is emphasis on professional application of *phronesis* or practical wisdom as a key component of administrative fairness (Kinsella & Pitman, 2012). This means a key distinction from Prony's calculators lies in the nature of work performed by frontline staff like police officers, nurses, and social workers. Unlike rote formula execution, their roles necessitate substantial discretionary judgment to navigate nuanced, complex situations that demand empathy and sound decision-making.

While embodying Prony's layered approach, the civil service crucially diverges by embracing human discretion at its foundational levels. This shapes a system valuing

not just rule adherence but thoughtful application of rules informed by practical wisdom. The civil service recognises that effective public service delivery, especially in roles dealing with vulnerable individuals, cannot be reduced to rigid formulas. It requires civil servants to apply codified guidelines with discernment, accounting for the intricacies of each unique circumstance.

The role of discretion in administrative fairness

If one considers the Civil Service Code as the rules by which civil servants conduct their decision making, it will follow that the categories of AI deployment which pose the greatest risk to public trust will be the ones where human discretion has always played a key role, often this means frontline public service work. Frontline public servants play a vital role in policy implementation through their discretionary judgments and actions on the ground. Even where policies aim for standardisation, successful implementation still heavily depends on local context and the "messy engagement of multiple players with diverse knowledge" (Davies et al., 2008, p.190). This echoes the concept of "street-level bureaucrats" (Lipsky, 1980), which highlights how frontline workers' discretionary decisions can prove instrumental in determining a policy's success or failure.

Empathy in decision making

Exercising discretion with empathy and understanding of human contexts is crucial for effective public service delivery, especially in complex situations impacting vulnerable individuals. Frontline civil servants are expected to uphold ethical principles from the Civil Service Code, however, equally important is applying these principles with discernment, taking into account the nuances of each situation. For example, the discretionary nature of frontline services can be seen in emergency assistance schemes that aim to provide crisis support to low-income families. In one particular council, the Welfare Assistance Scheme requires applicants to fit into certain "categories of vulnerability" in order to receive a Daily Living Expenses grant, such as having a dependent child at immediate risk, or having a learning disability (Perry et al., 2014). However, no published criteria are provided for how such "vulnerability" is assessed, allowing frontline civil servants administering the scheme to exercise discretion in evaluating each applicant's situation and level of urgent

need. This shows how the design of many localised benefit schemes delegates significant discretionary power to street-level bureaucrats in deciding who qualifies as sufficiently "vulnerable" or "in crisis" to receive emergency financial support, based on their own judgments (Lipsky, 1980). An empathetic civil servant can use practical wisdom to ensure administrative fairness by exercising judicious discretion. While rules and regulations aim to constrain discretion, Lipsky argues that street-level bureaucrats often view efforts to dictate service norms as illegitimate incursions on their professional judgment. They will exploit provisions for exceptions and loopholes in rules to circumvent reforms intended to limit their discretion (Lipsky, 1980). Such examples demonstrate the importance of studying frontline decision-making in the implementation of AMAs across public service because no top-down rule can account for such nuance.

Risks of lack of empathy

Administrative fairness and impartiality have long been foundational to the civil service, but it is worth noting that the application of discretion in a top-down bureaucracy has produced myriad flaws and failures. To wit, “indicators of perceived unfairness, inconsistency, or sloppy administration” have plagued the quality of UK public services for at least the past three decades (Hood & Dixon, 2015). For example, in areas like assessment streamlining to determine health service eligibility, human civil servants currently apply codified principles to complex situations requiring empathy and contextual understanding. A prominent utilitarian approach based on decision theory is called the quality-adjusted life year (QALY) which is defined as “the time in good health equivalent to a year of ill health” (Parmigiani, Inoue & Lopez, 2009, Section 4.3.3). The QALY is a heuristic which can help with decision making when it comes to the prioritisation of National Health Service (NHS) funds, but it is an imperfect algorithm, with which research has identified issues with equity and measurement of true preference (Whitehead & Ali, 2010).

In fact, the Health Service Ombudsman saw a fivefold increase in complaints about frontline NHS decision making between 1990-2010, suggesting declining fairness and consistency as managerialism and use of QALYs increased (Hood & Dixon, 2015). This means there is room for improvement, and a problem space in which AMAs could help, but it is at the same time essential to understand what we are

proposing is replaced by artificial decision making, especially at the risk of unfair treatment, lack of accountability, and absence of empathy for vulnerable individuals.

Section 3: Prioritisation of AI deployment in public service

As the above sections show, there are inherent risks if moral agency is fully ceded to AMAs in areas of thick rules. As Lorraine Daston states, "human beings can improvise. This is a very difficult challenge for programmes which expect the world to be steady as she goes, stable and predictable…We should be using AI for the parts of our world which really are predictable" (Mackereth & Drage, 2022, 30:45).

What exactly are the parts of the public service world “which really are predictable”? We can say more clearly where they are not. Applications like automated decision-making, customer-facing roles and chatbots pose pronounced risks related to fairness, bias, accountability and misunderstanding sensitive issues if human discretion is removed. Given these risks, in the near term, prioritisation of AMA implementation should focus first on lower-risk categories operating under strict ethical guidelines with limited autonomy, like business planning, demand prediction and process automation with human oversight. Only once competence in AI governance matures should higher-risk applications be considered that necessitate significant moral reasoning and ceding of human discretion.

Practically speaking, the standard approach of evaluating a benefits case, as outlined in the Green Book (HM Treasury, 2020), should incorporate a comprehensive risk assessment. This means fully understanding the levels of artificial moral agency required, the categories of AI use in public service, rule implementation approaches, and most importantly - the specific risks to public trust, fairness and accountability for each proposed application.

By taking an incremental, risk-based approach as illustrated in Table 3, continuously evaluating benefits alongside ethical principles, public trust can be maintained as AI is judiciously incorporated into areas of government where it enhances rather than replaces human discretion and context.

Table 3: Risk-assessed roadmap
Table 3 offers a possible roadmap for the deployment of AMAs in UK public service, which allows for iterative improvement and gradual build up of AI governance as technologies mature.

Priority	Risk level	Categories	AMA Level
1	Low	Business planning and demand prediction	Level 1, 2
2	Medium	Assessment streamlining, Process automation	Level 2, 3a
3	High	High Chatbots/call centers impacting vulnerable individuals	Level 3b

Transforming Priorities: The AGI Paradigm Shift

While I have advocated for a measured, evolutionary approach to AMA integration in public service, critics might find this perspective overly cautious given predictions of an imminent intelligence explosion that could rapidly deliver Level 4 AGI capabilities. Some circles believe that humanity is on the cusp of an explosion in intelligence, with human-level AGI on the very near horizon. Moorhouse and MacAskill (2025) suggest that we will soon experience a “century in a decade” of technological progress. Should this level of AI arrive, we need to have already spent time considering the philosophical frameworks best suited for shaping these technologies, and will also likely need to envision an entirely new framework for human-machine collaboration within civil service roles.

Hybrid Model of Agency

Most Level 2 AI has been built in a top-down manner, which is much like a Kantian, deontological design, where thin rules descend from the top and are percolated through the system. This sort of AI has been created in the expert systems tradition, where it is believed that a certain set of rules will suffice to govern system behaviour. Wallach and Allen highlight a key limitation of such systems: their rigidity, as they fail to account for the complexities and nuances found in real-world situations (Wallach &

Allen, 2009). As we have seen, top-down rules are not strong enough to replace human agency and discretion in crucial frontline roles.

Conversely, a bottom-up approach allows for the emergence of behaviour from the interaction of simpler components, which could foster a more responsive AI which could be adaptive to thick rules. This perspective aligns with the more sophisticated capabilities we see at Levels 3a and 3b of moral agency as described by Moor, Formosa, and Ryan, where AI systems potentially exhibit explicit moral reasoning and can adapt to context in a meaningful way. This is a bit more suitable to public service purposes, in that it approaches something closer to Aristotelian virtue ethics, as emergent characteristics of an AI. However, there is then less of an incentive for the AMA to act under prescribed rules, and a propensity for it to make mistakes whilst learning.

Wallach and Allen (2009) suggest it is unwise to assume "sophisticated capacity for moral judgment will just emerge from bottom-up engineering" (p.116). By harnessing the strengths of both top-down and bottom-up methodologies, a hybrid approach emerges as the most promising path forward. This hybrid model can react in a more flexible fashion to contexts where both thick and thin rules apply. However, there is a missing component, which, as Wallach and Vallor (2020) argue, may never be possible to produce in an AMA, and that is emotion, including crucially the empathy required for compassionate discretion in frontline public service.[1]

A New Human-AI Collegiate Relationship

A conventional master-servant dynamic between humans and machines, advocated by thinkers like Bryson (2009), seems ill-suited for integrating Level 4 AMAs because of the (theoretical) advanced relational capabilities of the technology. Instead, I propose cultivating a new type of relationship akin to the bond between a service dog and its human handler, along the lines of that proposed by Donna Haraway, who

[1] The emotional dimensions of AMAs represent a fascinating area for research - one I hope to address in future work. For this paper, I chose to focus on near-term planning for Level 3a AMAs in public service because it has the most practical impact on my immediate work in government.

states that “all ethical relating, within or between species, is knit from the silk-strong thread of ongoing alertness to otherness-in-relation.” (Haraway, 2016, p.141). A partnership which eschews the master-servant dynamic and instead favours equal partnership of capable adults of different species is founded on mutual training, trust, and respect, with distinct responsibilities for each party that ultimately service a unified mission.

In a public service context, this could translate to Level 4 AMAs being assigned specific duties and job descriptions aligned with their capabilities, much like their human counterparts. They may be given humanlike names and occupy particular grade levels commensurate with the complexity of their roles (Brown, 2025). However, a key distinction would be preserving ultimate human discretion on decisions carrying profound moral weight or impacting vulnerable individuals.

The Level 4 digital civil servant built in the hybrid top-down/bottom-up model would augment rather than replace human faculties like reasoned judgment. At the same time, they could be held accountable through performance evaluations (much like their human peers) and be expected to provide transparent rationalisation for their actions, responding to freedom of information requests and Prime Minister’s questions in the same way that human civil servants do.

This collaborative paradigm presents a balanced path forward, allowing AMAs to enhance operational efficiency and consistency, while retaining uniquely human capacities like empathy and practical wisdom. Considerable further exploration is required into developing AMAs with sufficient emotional aptitude and moral discernment to potentially participate as full ethical agents in the future.

For now, the focus should remain on formalising responsible frameworks for Level 2/3ab AMA deployment across carefully risk-stratified public service domains. But this grand moral vision of digital civil servants illustrates the need to proactively cultivate partnership models that respect human moral primacy while harnessing the potential of artificially intelligent colleagues. Only through such an evolutionary approach can we shape a future where technology and ethical governance remain constructively intertwined.

Conclusion

Delivering trustworthy public services hinges on understanding how human agency interacts with artificial agency in contexts of administrative fairness. While AI technologies can offer significant improvements in efficiency, frontline civil servants play an essential role in making discretionary decisions that directly influence the implementation of policies. Ceding too much discretion to artificial agents risks compromising the fundamental principles of the civil service and eroding public trust.

The key recommendations of this paper include prioritising low-risk AI implementations, establishing oversight and governance of AI deployment, and incrementally expanding AMA use as governance frameworks mature. Practical challenges such as managing risks related to bias, accountability, and empathy must be proactively addressed to ensure equitable service delivery. Future research should explore the emotional capabilities of public service AI, and an interdisciplinary approach will be crucial for advancing both the technology and ethical standards that define public services.

The future of AI in the UK civil service should emphasise human-machine collaboration, where AMAs serve as supportive colleagues rather than replacements for human judgement. Such a responsible integration approach, coupled with risk-aware prioritisation of adoption, will enable a more efficient and fair public service, ensuring that both the technological advances and the core human values that underpin civil service are upheld.

References

Asaro, P. M. (2006). What should we want from a robot ethic? *International Review of Information Ethics*, 6, 9–16.

Booth, R. (2025). ‘Mainlined into UK’s veins’: Labour announces huge public rollout of AI. *The Guardian*. https://www.theguardian.com/politics/2025/jan/12/mainlined-into-uks-veins-labour-announces-huge-public-rollout-of-ai

Brown, R. (2025). *Synthetic workers and the enterprise: A comprehensive introduction to how the enterprise thinks about autonomous AI staff* (1st ed.). Mission Control AI PBC. Retrieved from https://usemissioncontrol.com

Bryson, J. J. (2009). *Robots should be slaves*. In *Artificial models of natural intelligence*. University of Bath.1.

Committee on Standards in Public Life. (1995). *First report of the Committee on Standards in Public Life* (Cm 2850). HMSO. 14.

Daston, L. (2017). *Calculation and the division of labour, 1750-1950: 31st annual lecture of the German Historical Institute, Washington, DC, November 9, 2017*. Max Planck Institute for the History of Science.

Daston, L. (2020). *Rules: A short history of what we live by*. Princeton University Press.

Davies, H., Nutley, S., & Walter, I. (2008). Why "knowledge transfer" is misconceived for applied social research. *Journal of Health Services Research & Policy*, 13(3), 188–190. https://doi.org/10.1258/jhsrp.2008.008055

Eaves, D., & Goldberg, D. (2017). *UK government digital service: Moving beyond a website (HKS Case Draft)*. Harvard Kennedy School.

Elgot, J., & Booth, R. (2025). AI tool can give ministers 'vibe check' on whether MPs will like policies. *The Guardian*. https://www.theguardian.com/technology/2025/jan/20/ai-tool-can-give-ministers-vibe-check-on-whether-mps-will-like-policies

Formosa, P., & Ryan, M. (2021). Making moral machines: why we need artificial moral agents. *AI & Society*, *36*(3), 839–851. https://doi.org/10.1007/s00146-020-01089-6

Gabriel, I., et al. (2024). The Ethics of Advanced AI Assistants, Ch. 5 & 6, “Value Alignment” and “Well-being”. Google DeepMind. 30.

Government Digital Service. (2024). *Redbox: AI for civil servants*. Retrieved 21 February 2025, from https://ai.gov.uk/projects/redbox/

Haraway, D. J. (2016). Manifestly Haraway. In *The Companion Species Manifesto*. 141.

Haksar, V. (1998). Moral agents. In *The Routledge Encyclopedia of Philosophy*. Taylor and Francis. Retrieved 1 March 2025, from https://www.rep.routledge.com/articles/thematic/moral-agents/v-1. https://doi.org/10.4324/9780415249126-L049-1

Hood, C., & Dixon, R. (2015). Consistency and fairness in administration: Formal complaints and legal challenges. In *A government that worked better and cost less? Evaluating three decades of reform and change in UK central government*(online edn, Oxford Academic, 21 May 2015). https://doi-org.ezp.lib.cam.ac.uk/10.1093/acprof:oso/9780199687022.003.0006

HM Treasury. (2020). *The Green Book: Appraisal and evaluation in central government*. UK Government. Retrieved from https://www.gov.uk/government/publications/the-green-book-appraisal-and-evaluation-in-central-government/the-green-book-2020

Hudson, B., Hunter, D., & Peckham, S. (2019). Policy failure and the policy-implementation gap: Can policy support programs help? *Policy Design and Practice*, 2(1), 1–14. https://doi.org/10.1080/25741292.2018.1540378

Kinsella, E. A., & Pitman, A. (2012). *Phronesis as professional knowledge: Practical wisdom in the professions*. Brill.

Lipsky, M. (1980). *Street-level bureaucracy : dilemmas of the individual in public service.* Russell Sage Foundation.

Mackereth, K., & Drage, E. (Hosts). (2022). *The exorcism of emotion in rational science (and AI) with Lorraine Daston* [Audio podcast episode]. The Good Robot Podcast. https://www.thegoodrobot.co.uk/post/lorraine-daston-on-the-exorcism-of-emotion-in-rational-science-and-ai

Mason, R. (2025, March 12). AI should replace some work of civil servants, Starmer to announce. *The Guardian*. https://www.theguardian.com/technology/2025/mar/12/ai-should-replace-some-work-of-civil-servants-under-new-rules-keir-starmer-to-announce

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4), 18-21. https://doi.org/10.1109/MIS.2006.80

Moorhouse, F., MacAskill, W. (2025). *Preparing for the Intelligence Explosion*. Retrieved 13 March 2025 from https://www.forethought.org/research/preparing-for-the-intelligence-explosion

Parmigiani, G., Inoue, L. Y. T., & Lopez, H. F. (2009). *Decision theory: Principles and approaches* (1st ed.). John Wiley & Sons.

Perry, J., Williams, M., Sefton, T., & Haddad, M. (2014). *Emergency use only: Understanding and reducing the use of food banks in the UK*. 119.

Taylor, L. (Host). (2022, September 28). Rules and order [Audio podcast episode]. *Thinking Allowed*. Retrieved from https://podcasts.apple.com/us/podcast/thinking-allowed/id261548752?i=1000580939153&r=960

UK Department for Science, Innovation & Technology (DSIT). (2025). *AI opportunities action plan*. Retrieved 1 March 2025 from https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan

UK Public General Acts. (2010). *Constitutional Reform and Governance Act 2010* (c. 25). Retrieved 9 March 2025 from https://www.legislation.gov.uk/ukpga/2010/25/section/5

Vallor, S. (2022). Carnegie Council podcasts: "That wasn't my intent": Reenvisioning ethics in the information age, with Shannon Vallor. Retrieved 24 February 2025 fromhttps://podcasts.apple.com/us/podcast/carnegie-council-podcasts/id13006 2462?i=1000547007737&r=3218

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford Academic. 116.

Wallach, W., & Vallor, S. (2020). *'Moral Machines: From Value Alignment to Embodied Virtue'*, in *Ethics of Artificial Intelligence*. Oxford Academic.

Whitehead, S. J., & Ali, S. (2010). Health outcomes in economic evaluation: The QALY and utilities. *British Medical Bulletin*, 96, 5-21. https://doi.org/10.1093/bmb/ldq033