

AI as a constituted system: accountability lessons from an LLM experiment

Kelsie Nabben 

Max Weber Fellow, Robert Schuman Centre for Advanced Studies, European University Institute, Florence, Italy
Email: kelsie.nabben@eui.eu

Received: 20 October 2023; **Revised:** 18 April 2024; **Accepted:** 15 May 2024

Keywords: AI; ethnography; governance; LLM; organisation

Abstract

This study focuses on the practicalities of establishing and maintaining AI infrastructure, as well as the considerations for responsible governance by investigating the integration of a pre-trained large language model (LLM) with an organisation's knowledge management system via a chat interface. The research adopts the concept of "AI as a constituted system" to emphasise the social, technical, and institutional factors that contribute to AI's governance and accountability. Through an ethnographic approach, this article details the iterative processes of negotiation, decision-making, and reflection among organisational stakeholders as they develop, implement, and manage the AI system. The findings indicate that LLMs can be effectively governed and held accountable to stakeholder interests within specific contexts, specifically, when clear institutional boundaries facilitate innovation while navigating the risks related to data privacy and AI misbehaviour. Effective constitution and use can be attributed to distinct policy creation processes to guide AI's operation, clear lines of responsibility, and localised feedback loops to ensure accountability for actions taken. This research provides a foundational perspective to better understand algorithmic accountability and governance within organisational contexts. It also envisions a future where AI is not universally scaled but consists of localised, customised LLMs tailored to stakeholder interests.

Policy Significance Statement

This paper provides a crucial empirical contribution to literature and practice on AI governance, and specifically, accountability. The findings emphasise that with clear context mapping and accountability measures put in place, AI can be governed for greater normative and mechanistic accountability within bounded, local contexts. The findings also challenge universal, homogenous visions of AI futures, offering more local, context specific possibilities. For policymakers, this work underscores the importance of crafting adaptable regulations, some of which are embedded in the design of the technological architecture itself, and some of which are externally imposed, monitored, and enforced through institutional measures, to ensure AI technologies are both innovative and accountable. Understanding the role of AI and people's behaviour in relation to AI in real-world settings enables practical, implementable AI policy development that safeguards public interests while promoting technological advancements.

1. Introduction

This paper examines the considerations involved in establishing, deploying, and governing AI as a system constituted within the context of a specific group of stakeholders to respond to the research question: “how can large language models (LLMs) be made more accountable to their users?.” Using participatory ethnographic research methods, I investigate accountability as a key dimension of AI governance through a case study at BlockScience, an engineering services firm (where I also do contract work), as they engage in the design, deployment, and governance of a LLM to shed light on accountability as a key dimension of AI governance. Specifically, BlockScience’s experimental project focuses on the integration and development of the internal knowledge management system (KMS) into a chat interface powered by an LLM, referred to as “KMS-general-purpose technology (GPT).” Their aim was to create a “deployable, governable, interoperable, organisation level KMS.” To undertake this study, I adopt a novel approach towards understanding “AI as a constituted system.” AI as a constituted system requires consideration of the social, economic, legal, institutional, and technical components and processes required to codify, institutionalise, and deploy AI as a product.

The concept of “constitutional AI” arises from the field of AI safety research which aims to align AI behaviour with human goals, preferences, or ethical principles (Yudkowsky, 2016). This concept of provisioning a text-based set of rules to be narrowly interpreted by an AI agent for training and self-evaluation emerged as an AI safety approach in relation to LLMs; which are a subset of AI agent created from Natural Language Processing techniques (Bai et al., 2022; Durmus et al., 2023). AI governance refers to the ecosystem of norms, markets, and institutions that shape how AI is built and deployed, as well as the policy and research required to maintain it, in line with human interests (Bullock et al., 2024; Dafoe, 2024). However, the constitutional approach to AI governance is limiting, as it draws on understandings of a constitution as a closed system of norms or unifying ideals that are expressed in plain language and employed as a governance mechanism. In practice, the governance of LLMs involves a range of considerations that extend beyond simply responding to natural language prompts. I refer to this approach as “AI as a constituted system” (see Table 1). This approach involves acknowledging AI as a constituted system of social and technical processes and decisions, in order to establish and maintain accountability

Table 1. Mapping AI as a constituted system

Consideration	Case study: KMS-GPT
<i>Social:</i>	
Purpose	To provide an easily accessible interactive knowledge source to augment existing work practices.
Stakeholders (first order—direct contact with system, inputs, and/or outputs)	<ul style="list-style-type: none"> - BlockScience (private organisation) - BlockScience team members
Stakeholders (second order)	N/A as outputs were not used for public materials or client projects at this stage of the project.
Values (what is and is not normatively acceptable)	<p>According to the LLM/GPT Policy:</p> <ul style="list-style-type: none"> - Safe and ethical use of LLMs - Data privacy of training data, outputs, and no input or use of sensitive or confidential information. - Transparency (i.e., internal disclosure of LLM use if not clear; citation if outputs are publicly used) - Efficiency and productivity - A key value of BlockScience is “our moral and ethical obligation to display intellectual integrity in our research and to design and maintain safe systems in our engineering work.”

(Continued)

Table 1. Continued

Consideration	Case study: KMS-GPT
	<ul style="list-style-type: none"> - Adherence to copyright laws - Prohibition from using LLMs to engage in activities that are illegal or unethical, or to generate content that is inappropriate, offensive, or harmful - Human review of all outputs before being shared with others - Adherence to the policy whenever using LLMs for company purposes
<i>Technical:</i>	
Underlying model	Open AI GPT-4
Training approach	Reinforcement Learning from Human Feedback (RLHF)
Data (inputs)	All non-sensitive company archive data, included open Slack message channels, research publications, blogs, and research notes. Exclusion of sensitive data.
Data (outputs)	Various. For internal testing only
Access to model (design)	Project technical team
Access to model (implementation)	Project technical team
Access to model (use)	BlockScience team members that choose to join the “KMS-GPT” Slack channel
Access to outputs from model	For internal testing only
<i>Economic:</i>	
Funding source	Internally funded R&D project by BlockScience
Capital and labour costs (internal)	<ul style="list-style-type: none"> - Team member compensation - Operational overhead costs - Opportunity cost of internal R&D
Capital and labour costs (external)	<ul style="list-style-type: none"> - Open AI account - Other cloud infrastructure
Related business model/value (if any)	<ul style="list-style-type: none"> - Capabilities enhancement (tools and skills) - External facing R&D
<i>Institutional (internal):</i>	
Monitoring responsibilities	<ul style="list-style-type: none"> - Project technical team + Operations Council - LLM policy
Monitoring intervals	<ul style="list-style-type: none"> - Weekly technical team meeting - Monthly wider team meeting
Consequences for misuse of model	Ops Council emergency response + situation report and team member review meeting if required
Consequences for misbehaviour of model	Technical team review and action (e.g., change data access, server settings, or switch off the model, depending on situation)
Enforcement measures	Powers extending within BlockScience as a private company.
<i>Legal (external):</i>	
Relevant legal considerations	<ul style="list-style-type: none"> - US law, with consideration for some team members in other jurisdictions - Respecting NDAs, commercial, and confidential data - Respecting IP law - Respecting internal team member data, including personal information or personal interactions.

within the interests of the stakeholders responsible for, and engaging with, an LLM. This includes establishing written policies among a set of stakeholders to guide its operation in a specific context, the institutional structures that provide the hierarchies of answerability and responsibility and governance processes to oversee and adapt the system, as well as the technical building blocks of the underlying model, database construction, data access, and maintenance of the necessary operating infrastructure, such as servers.

1.1. *Accountability in AI*

Researchers emphasise that accountability is the key to AI governance, little is known about how to ensure AI agents are accountable to the interests of their constituents in practice (Schwartz et al., 2023). This is in large part because AI operates in an “accountability gap” between its technical features and the social contexts in which it is deployed, making it a challenge to verify compliance with normative principles that are being accounted for (Mittelstadt et al., 2016; Lechterman, 2024, p. 164). Accountability in the narrow, concrete sense of “giving account” refers to the mechanisms by which individuals, organisations, or institutions are held responsible for their decisions and actions in relation to one another (Bovens, 2007). However, accountability in the use of AI models in local, multi-stakeholder environments is more than a “relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences” (Bovens, 2007, p. 450), as an algorithm as an actor cannot face consequences in the same way as a human or organisation. Wieringa (2020) points out that algorithmic accountability requires appreciation of the entanglement between algorithmic systems, how they are enacted, and by who. Thus, *algorithmic* accountability is the analysis of the actions taken between an actor (e.g., a developer, a firm, a contracted organisation, or the deployer of a “human-out-of-the-loop” system), a forum of stakeholders, the relationship between the two, and the consequences of certain actions (Wieringa, 2020). Building on this conceptualisation with a qualitative contribution and viewing AI as a constituted system, the definition of accountability adopted in this article is the ability for those that constitute an algorithm to accept responsibility for the actions of an algorithm and be held to account. One way to increase accountability in AI systems is to help designers identify the social, technical, and institutional context of a system (Suchman, 2002; Wieringa, 2020). Indeed, Holmström and Hällgren (2022) hypothesise that informed constituting of AI in organisations requires an awareness of AI’s context-dependent nature, along with a clearly defined purpose for use and an appropriate level of management and transparency (2022). Thus, algorithmic accountable needs to move beyond mere technical compliance or normative ethical guidelines to a more comprehensive, multi-scale, system-wide approach that encompasses social, institutional, technical, economic, organisational, and legal considerations of an algorithmic system, how decisions are made, and what consequences can be made enforceable upon which actors. Yet, few guidelines or frameworks exist that show how to practically do this.

This article ethnographically traces an LLM experiment to consider the socio-technical composition of an LLM infrastructure to detail the processes and experiences of establishing and maintaining AI accountability within a specific, organisational context. Thus, rather than pursuing AI accountability at scale by trying to aggregate the diverse preferences from the general public into written principles as a governance tool, this approach defines and maintains rules outside the system, whilst iteratively imbuing them in the technical architecture of the system and the institutional settings surrounding the system. This combination of technical settings, social norms, and institutional policies for monitoring, enforcement, and consequences pursues accountability for both designers and users of an LLM in an organisational setting. The argument presented, that accountability is a core pillar of AI governance and must be considered across the social, technical, and institutional dimensions of AI as a constituted assemblage, provides an important contribution to the relevant literature and discourse on AI governance and accountability.

1.2. Methods

To explore these key themes of accountability and processes of architecting and governing AI as a constituted system, this article adopts the theoretical lens of infrastructure studies. The field of infrastructure studies acknowledges technological systems as socio-technical architectures, composed on inextricably linked social and technical components, including people, processes, software, hardware, resources, maintenance, and so on (Star, 1999; Golden, 2013). AI has already been documented in general as a material artefact composed of social and technical components, including raw resources, human labour, data, state legislation, and more (Crawford, 2021).

This study employs digital ethnographic methods to contribute empirical evidence to this notion of establishing the goals and boundaries of an AI system in an organisational context in a governable and accountable manner. AI practice is interdisciplinary by nature and benefits from not only technical but *socio-technical* perspectives that acknowledge the inextricable links between the social and technical dynamics of any complex system (Sartori and Theodorou, 2022). Ethnographic methods, as a subset of anthropology, focus on elucidating human (and non-human) interactions that are situated in social and technical contexts. Digital ethnography specifically seeks to understand the dynamics of a culture, organisation, or conventions in a digital world by embracing engagement with, and mediation through, digital tools (Pink et al., 2016). For this study, research data were collected from digital platforms and was mediated through digital means (e.g., video calls and Slack message interactions), rather than being physically co-located with research participants in one field site. Of significance for guiding ethnographic inquiry in algorithmic domains is the question of “how nonhuman entities achieve a delegated agency within socio-technical networks” (Hess, 2001, p. 236). Orlikowski demonstrates the value of ethnographic approaches for revealing the co-constitutive entanglements of social and material dimensions of technical arrangements in organisational settings (2007). Ethnography in the context of technology design provides a methodological lens by which to consider the arrangement of human and non-human elements into stable organisations and artefacts, towards enduring alignment of such elements (Suchman, 2000).

The anthropology of AI is a longstanding tradition in the anticipation of AI-augmented practice. Anthropologists were original participants in conferences on cybernetics, “the science of controlling the flow of information in biological, mechanical, cognitive, or social systems with feedback loops” (Mead, 1968), from which the field of AI stems. In 1996, researchers mused about what “a digital computer or robot-controlled device to perform tasks commonly associated with the higher intellectual processes characteristic of humans” could do to augment human processes, including data analysis and new logics to help anthropologists understand human reasoning (Chablo, 1996). Anthropologists also have a key role in expressing ethical concerns about AI’s potential to perpetuate bias based on sex, race, class, and ability (Richardson, 2020). For example, anthropologist Diana Forsythe focused her work on how computer scientists and engineers create systems and the implications of those creations on society at large (2001). Understanding the cultural, social, and cognitive contexts in which technologists work, and how their perspectives, biases, and assumptions are embedded into the technologies they create, aids us to understand and influence how society understands itself (Forsythe, 2001). In this case, the ethnography of an AI that BlockScience trained and deployed within the organisation engages in an effort to understand the considerations of constituting AI as a system within a given context, as a basis for thinking about AI accountability in various contexts. I focus on the current state of AI, rather than focusing on a future state of “Artificial General Intelligence” (or “Superintelligence”). Nonetheless, this study raises important questions and further research directions regarding the considerations of AI deployment, constitutionalisation in complex socio-technical ecosystems, and accountability.

The few ethnographies of AI that have been conducted so far tend to focus on either innovation in ethnographic methods and AI (Sumathi et al., 2023), the ethnography of software engineers (Forsythe, 2001; Blackwell, 2021), or the ethnography of AI in the broader materiality and social structures within which it is created (Crawford and Joler, 2018; Crawford, 2021). Anthropologists argue that even if we cannot explain how some of our machine learning models work, we can still illuminate them through anthropological methods (Munk et al., 2022). Beyond focusing on the algorithms themselves, Marda and

Narayan assert that to understand the societal impacts of AI, we need to employ qualitative methods like ethnography alongside quantitative methods to “shed light on the actors and institutions that wield power through the use of these technologies” (Marda and Narayan, 2021, p. 187).

BlockScience’s instance of “KMS-GPT,” explained in detail throughout the remainder of this article (see Figure 1), was a suitable field site to respond to the research question for multiple reasons. First, it represents an important area for the investigation of LLM accountability, as BlockScience is a professional services firm that decided to experiment with deploying an LLM as an interface to translate and communicate their KMS. There has been much conjecture about how LLM’s will disrupt professional services industries (Dell’Acqua et al., 2023), as well as how AI will change the nature of how organisational knowledge is constructed, transmitted, and maintained in the first place. Second, it was an organisational scale experiment with bounded locality by which to observe the constitution of an AI system. This also presented a clear context for which I was able to gain access and ethics approval to conduct data collection. To conduct this research, I collected digital ethnographic data under university ethics approval from the organisation and research participants. Research data included call recordings, internal “Slack” chat organisational messaging app conversation threads, and direct interactions with KMS-GPT, as well as unstructured ethnographic interviews with BlockScience team members, personal participation in meetings and conversations, and reflections. In my role as “the ethnographer,” I became the research instrument to identify, articulate, and, where appropriate, translate the principles and values that were present (Forsythe, 2001). My position in relation to the research field of BlockScience was one of an insider-outside. As a member of the organisation, I was an insider, with access to KMS and KMS-GPT, as well as a participant in Slack conversations and research calls as they unfolded. Simultaneously, as a social scientist and “non-technical” team member in an engineering firm, I was an outsider with participant-observer status to study the experiment. BlockScience’s KMS-GPT experiment is the object of this ethnographic study.

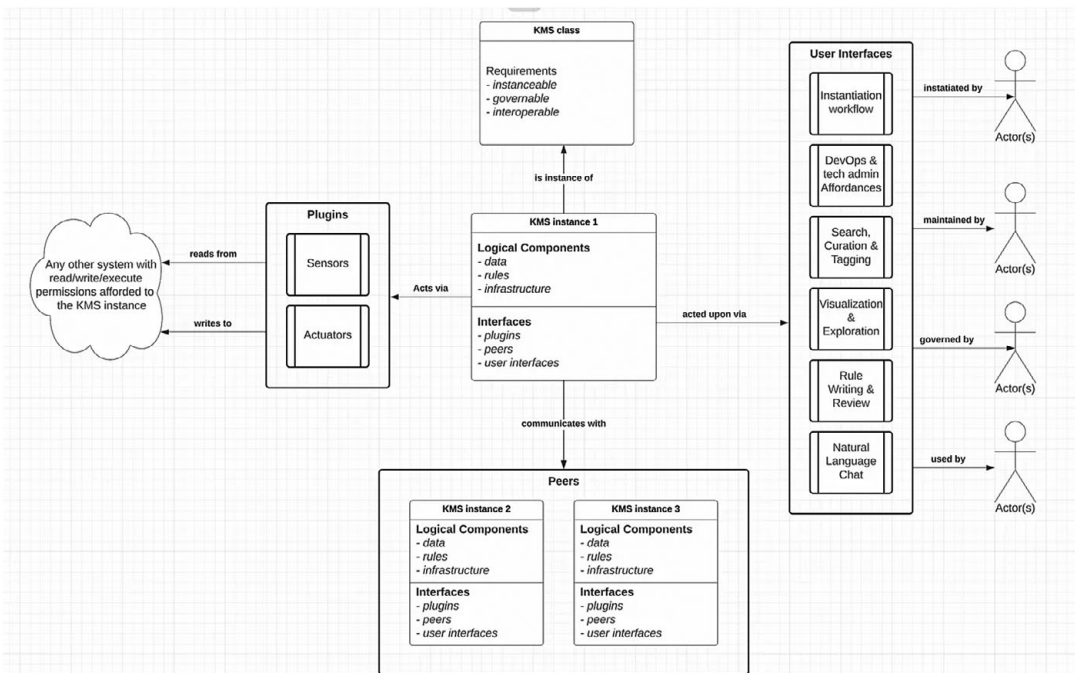


Figure 1. “High-level architecture for KMS as protocol” (Zargham, 2023).

No raw or sensitive research data were input directly into ChatGPT, and direct outputs from GPT have only been used where explicitly mentioned.

The remainder of this piece is structured as follows: In Part 1, I outline the considerations that have been navigated in constituting KMS-GPT as a knowledge management LLM in an organisational context. In Part 2, I lay out the governance considerations that deployment of KMS-GPT raises. I then map the accountability relationships at play between individual stakeholders, forums of stakeholders, and decision-making authorities in relation to AI in an organisational context. Finally, in Part 3, I analyse how social-organisational contexts provide institutional infrastructures, as the technical, legal, and social structures that help bind a complex field together and govern field interactions to enable and constrain governance activities (Hinings et al., 2017), and how these can be leveraged to create and reinforce AI accountability. Thus, this study provides a novel interpretation of constitutionalisation in relation to AI, to improve understanding of how algorithms can be governed and held accountable within organisational contexts. Additionally, it explores the potential for alternative AI futures that diverges from the prevalent model of uniform, globally scaled AI, and instead emphasises the need for localised, customised LLMs that cater to the needs and interests of particular stakeholder groups.

2. Part 1: KMS-GPT: a “well constituted knowledge object”

Knowledge is not only about what people know, but also what they do with machines and related production processes to create more knowledge. Technology is directly related to knowledge (Risse, 2022). AI’s ability to impact knowledge production—for example, how information is known, interpreted, and disseminated at individual, organisational, and societal levels—represents a crucial facet of AI disruption, as well as its potential for unintended consequences. Whilst much AI research posits that AI will disrupt key industries, including education, healthcare, professional services, and others (Dwivedi et al., 2021), Zargham, the Founder of BlockScience and one of the masterminds of the project as it unfolded, focuses on the generative ability of AI to transform his sector for the better. Specifically, Zargham envisions the possibility of creating inter-organisational linked KMS’s as “knowledge networks” (Zargham, 2023), as the basis of a new mode of pluralistic intra- and inter-organisational ways of knowing.

The KMS research project at BlockScience arose out of an internal experiment. Thus, the purpose of KMS-GPT was clear: to provide an easily accessible interactive knowledge source to augment existing work practices. The project involved re-imagining how data, and thus organisational knowledge, was organised internally, in the type of way that you may be able to imagine a group of systems theorists, engineers, and philosophers might find interesting and important. Having taken non-confidential data from BlockScience’s organisational drive, team recordings, and organisational messaging services and making it into a searchable database accessible via an (internal only) web URL, the question facing the project became, what’s next?

“The problem is, I don’t actually use it” I stated fairly directly to Orion Reed, my colleague, and the researcher leading the Knowledge Management project internally. “If I need answers to something like travel policies or recording locations, I ask the right team member for it, or someone that might know.” To me, navigating knowledge, especially in a digital-first, distributed organisation, was about relationships.

Orion had joined the weekly “Governance Research Pod” call to gather feedback and scope further requirements on the internal KMS. Having stayed awake until 12 am to join across time zones, he was probably regretting it by this point.

After almost an hour of discussing the features and potential directions of the KMS project, I said in passing, “it would be cool if we could interact with it in a different way, like through a Chat-GPT interface.”

“We could do that!” responded Orion. And to the credit of him and a small internal team, consisting of a database engineer (David Sisson), a control theorist (Michael Zargham), and a software developer (Luke Miller), we could.

Two days later, “KMS-GPT” was live. It was deployed in the company’s Slack channel to make it accessible through the Slack chat interface in the same way as a chat with any other colleague. It could not

answer questions yet, but it existed. The old, Web-search bar KMS platform interface had shifted to the new, exciting interactive AI chat interface that numerous team members were engaged in.

The multi-month experiment that followed (and remains ongoing) touches on the themes of knowledge, governance, interfaces, identity, ethics, attention, and accountability. The following sections detail considerations pertaining to some of these themes, with a particular focus on accountability—the mechanisms and/or processes through which individuals, organisations, or institutions are held responsible for their decisions and actions in relation to others (Bovens, 2007).

2.1. Creating a KMS AI

The first dimension of constituting an AI was technical. At their core, LLMs are reasoning engines that work by predicting the logical sequence of words to formulate a response to prompts. On the back end, they are a collection of “tokens” (coded data), and the algorithmic model predicts what token (in the output of a natural language word) is most likely to come next based on a substantial amount of pre-training, fine-tuning, and reinforcement learning. According to current training methods, getting to this point requires pre-training, data tokenisation, and fine-tuning. Pre-training involves feeding a model a load of data that needs to be curated (or “cleaned” through categorisation) and tokenised. Data tokenisation refers to turning (potentially sensitive) data into a distinctive identifier or “token” while maintaining its value and link to the original data. In fine-tuning, a smaller and more specific dataset is used to adjust weights (which are numerical values that start as random coefficients and are adjusted during training to minimise loss—the error between the output of an algorithm and a given target value) to bias towards certain responses for more relevant and coherent text for specific types of use cases. Reinforcement learning from human feedback is then used to optimise model performance by having humans rate model outputs. This is part of mapping an LLM appropriately to a given context and environment, including guardrails for political correctness and harm reduction (Bakker et al., 2022). A model is then deployed and hosted on a cloud or other server infrastructure, as well as monitored and maintained.

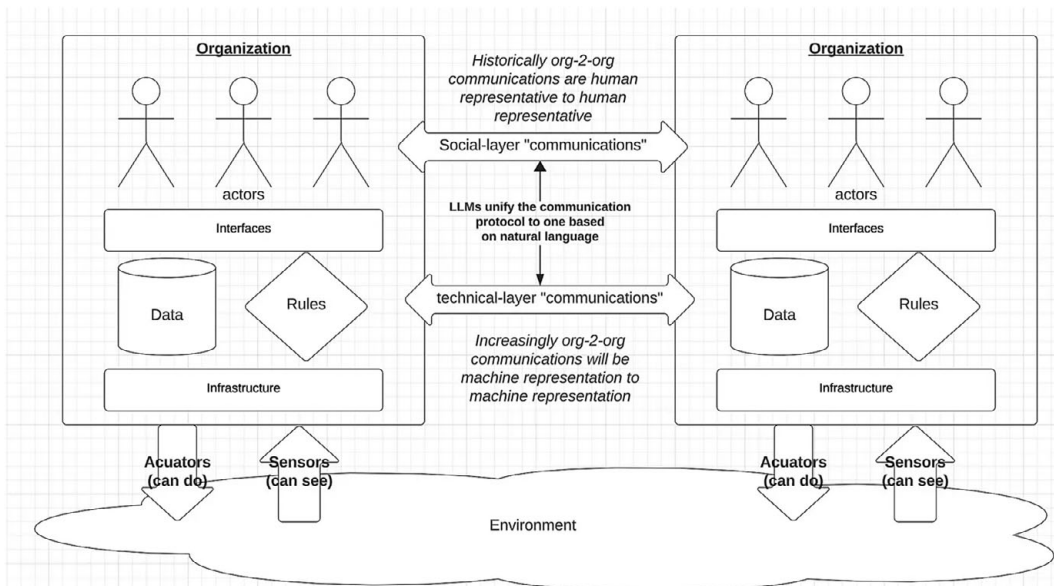


Figure 2. “Communication among Cyborg organisations can be human-to-human, human-to-machine, machine-to-human, or even machine-to-machine. With LLMs, natural language may be used for all of these interactions” (Zargham, 2023).

Building on this, KMS-GPT is based on the LLM model “Chat-GPT” by the company Open AI. This meant the team was augmenting a Pre-Trained Transformer Model, rather than training a model from scratch (which is incredibly expensive and resource intensive to do in terms of data and computational power to do). BlockScience was able to customise Chat-GPT but the underlying model remained closed-source, proprietary software code. The technical team that built the KMS system took organisational information (such as blogs, internal chat discussions, and policy documents) which they called “knowledge objects,” then “vectorised” this information into a database (to create a store of numbers that correspond to information), so it could then be queried by the LLM to generate a response and serve them to stakeholders via the Slack integrated chat interface. By taking the existing model, deploying a private instance on our own servers, and training it on our own data, we created a customised, local Chat-GPT agent that we could interact with through our organisation’s Slack chat (meaning it had a separate database and interface to Open AI). While this article focuses on constituting an LLM AI agent within the boundaries of an organisation, part of this context is that anyone can constitute their own AI as a system, for any purpose (Zargham et al., 2024). This process has become more user friendly since the time of data collection for this article, as highlighted by Open AI’s release of a “GPT store” for people to publicly or privately share their own customised GPT agents.

KMS-GPT was based on an AI framework for improving the quality of generic LLM-generated responses by “grounding” the model on external sources of knowledge to supplement the LLMs’ internal representation of information called retrieval-augmented generation (RAG). Grounding refers to the process of using LLMs with information that is use-case specific and not available as part of the trained knowledge set (Intellectronica, 2023). RAG means that the model references this as an authoritative knowledge base outside of the generic Chat-GPT training data sources before generating a response to enhance the quality, accuracy, and relevance of the output generated. Implementing RAG in an LLM-based question and answer system increases the probability that the model has access to more current and reliable facts, and that users have access to the model’s sources to check its claims for accuracy (Cai et al., 2022). In contrast to “fine-tuning,” which adapts an LLM’s internal knowledge by training it further on domain-specific data, RAG provides external information. According to Orion, “it would be a very expensive and minimal gain to try and fine-tune the model for our use case,” as RAG also reduces the need for users to continuously train the model on new data and update its parameters, which can lower computational and financial costs.

Through a multi-month-long series of iterative interactions with both the AI and each other about the governance and operational mechanics of KMS-GPT, we were constituting our own KMS. To constitute a knowledge system means that stakeholders of the organisation were able to participate in the design, implementation, operation, maintenance, and governance of knowledge infrastructure, data, prompt inputs, generated outputs, accountability measures, and recourse. This included decisions about the underlying foundational LLM model, data it had access to, servers, local integration into Slack, and access. BlockScience team members referred to this process of constituting our own system as “self-infrastructure” (defined below) (Nabben, 2023).

A common way team members tested KMS-GPT’s efficacy and relevance to our specific organisational context was to ask the same question to KMS-GPT and Open AI’s generic Chat-GPT, such as “what is self-infrastructure?” When Chat-GPT’s model was asked the same question, it responded: “As of my last training cut-off in September 2021, the term “self-infrastructure” isn’t well-defined or widely used in any particular field.”

When a team member then asked KMS-GPT what self-infrastructure is after more information was added to the dataset, KMS-GPT responded: “Self-infrastructure is a concept related to Web3 and digital infrastructure, where individuals and communities regain control over their governance surface by actively participating in the creation, organisation, and shared ownership of digital infrastructure. It involves a more decentralised and participatory mode of digital self-organizing, allowing communities to define their own values and rules. This approach aims to create a more equitable and inclusive digital landscape by empowering users to take control of their digital environment and directly contribute to the infrastructure they rely on.” (Although KMS did not include it, the correct reference for this would be Nabben, 2023.) Due to special attention from the BlockScience team on verifying source data, KMS-

GPT's ability to cite the data that was used to generate a response came later, when it was able to quote numbers that correlated to data (e.g., "1, 16, 22").

In this way, KMS-GPT shifted the nature of knowledge management within the organisation on at least two dimensions. First, KMS-GPT facilitated mediated knowledge to be produced. That is, the existing knowledge of the organisation was able to be leveraged to train the model to reason outputs in response to prompts. These outputs represent mediated forms of knowledge. Second, KMS-GPT facilitated the co-constitution of knowledge. The process of asking and receiving information from an LLM is collaborative and time effective. We pose questions or topics, and the model responds based on its training, and then we iterate. The understanding derived from this exchange is co-constructed between human and machine.

3. Part 2: governance considerations

3.1. *GPT as a constituted system and AI policymaking*

This section of this article considers the interplay between the technical considerations and the social inputs that shape AI as a constituted system. Deploying a GPT model that was trained on organisational data required substantive considerations to ensure that private data regarding client projects or team members was not accidentally leaked by the model in its responses. The team at BlockScience had already been thinking about constitutionalisation as boundary defining rule sets in digital contexts for research purposes (Zargham et al., 2023a). In the context of KMS-GPT, "constituting an AI" refers to the technical, social, and institutional processes of embedding the AI agent with boundary rules, principles, and norms of behaviour, throughout its training and fine-tuning process, in line with our preferences and priorities as an organisation. It's a "bundle of knowledge objects with familiar user experience," making it a "knowledge object," stated Zargham, in an AI transcribed conversation about "books" with database engineer David Sisson and Orion. The environmental context in which KMS-GPT was operating was within the boundaries of BlockScience as a private organisation. The stakeholders were members of that organisation. In an effort to explore the possibilities of AI-augmented knowledge management, we tested KMS-GPT out by asking it to justify its own existence.

"Based on the insights... please construct a short blog post (3 to 4 paragraphs) explaining why LLMs will be of benefit rather than harm to forward looking professional service firms that adopt them early and efficaciously."

After 5 minutes and 33 seconds, KMS-GPT responded: "Getting ready to dance..."

Defending his creation, Orion added: "...it's hitting some timeouts too. I think there might be a bottleneck with the async rewrite as others have encountered a similar issue with async openai calls."

This may have been due to hitting token limits, which are restrictions on the number of tokens (words) that an LLM can process in a single interaction, when using the pay-to-play OpenAI model. Based on some "napkin math," Orion figured out that the model could be faster on third-party software and hardware (such as Google Cloud and a Graphic Processing Unit (a specialised electronic circuit that can process many pieces of data simultaneously, making them useful for machine learning)). We knew that it would make sense in future to run our own model if usage of the model was high. This would be beneficial cost-wise and would also allow us to manage the technology stack—including embeddings (representations of text in numbers to capture the meaning of words), vector store (a database of embeddings that can be queried and retrieved by an LLM model), and the LLM model itself, which could be based on open-source software.

Upon re-prompting with the same text, KMS-GPT responded: "Making it all up..." After "restarting again with 4x the timeout limit..." the next response was not bad at all. The blog post included: "Title: Embracing LLMs: A New Era of Efficiency and Growth in Professional Services... In this post, we'll explore how LLMs can enhance the professional services industry by streamlining workflows, improving resource allocation, and fostering the growth of future leaders."

These responses showed that knowledge disruption via AI still had a way to go. The system required steering, both in terms of user prompting, and back-end architecture and hosting. KMS-GPT was not

without its hallucinations. Orion (and others) would continually test the model by asking it context-specific questions that the general Open AI GPT model could not answer as specifically, if at all.

Orion summoned KMS directly in a Slack channel entitled “#bsci-llm-research,” stating: “@KMS what is computer-aided governance?” (a central and novel concept in BlockScience’s research).

After some processing time, KMS-GPT responded: “Assistant: Computer-Aided Governance is a decision-support process that leverages technology to evaluate the potential results of policies and interventions, leading to inclusive and informed decision making. It is a way to use technology to help make decisions that are based on data and analysis, rather than relying solely on intuition or opinion.”

Orion: “@KMS could you expand on that?”

KMS-GPT responded much more enthusiastically than Orion was polite: “Of course! My previous response was regarding my capabilities as a language model. As a language model, I am able to process and understand large amounts of text, and can use this knowledge to provide accurate and informative responses to a wide range of questions. Additionally, I am constantly learning and improving, which means that my capabilities are constantly evolving. Is there anything specific you would like me to elaborate on?”

Orion wrote, more to the general channel of research participants and observers now: “^ this is still lacking... not sure why it does this yet.” He then wrote: “@KMS I was referring to computer-aided governance.”

Demonstrating its context-specific capabilities based on the custom training data it was fed, KMS-GPT responded accurately this time: “Computer-aided governance involves the use of technology to aid in the decision-making processes of government entities. This can include the use of data analytics, machine learning, and other tools to help identify patterns and make predictions about the effectiveness of different policies and programs. It can also involve the use of digital platforms to engage citizens and stakeholders in the decision-making process. Overall, computer-aided governance has the potential to improve the efficiency, transparency, and accountability of government, but it also raises important ethical and legal considerations that must be carefully addressed.”

Still, KMS-GPT was not without its idiosyncrasies. At times, people would feel inspired to query “@KMS” in Slack with a question, only to receive a response from Orion: “^ on a train so it was down for a bit,” as the model was initially being run off his laptop, meaning when it was closed and he was offline, no one was able to access it. KMS-GPT was also architected as a direct plug-in to Slack. This meant that people could summon it from within a Slack chat. It also meant that KMS-GPT observes its own conversations. Everything the LLM puts into Slack becomes part of the local RAG dataset, which includes public channels on the BlockScience Slack. This made it more difficult to test, as KMS-GPT could repeat previous responses to general questions, such as “What is KMS?,” instead of generating responses by reasoning about them, based on disparate datasets. Orion and Zargham were already discussing interactive mechanisms to counteract this. For example, “changing how distance is calculated for embedded piece of text” by getting team members to give a predetermined positive or negative emoji symbol (such as a red circle or a green circle) that act like a weight, and move inaccurate responses further away. This ability to evaluate KMS-GPT responses became a way for stakeholders to express their preferences over the behaviour of the model.

It turns out we were touching the frontier of AI governance research. The goal was to align KMS-GPT to the subjective, group-level interests of BlockScience as an organisation, and its purpose as an interactive interface to the KMS. This concept of creating systems that adhere to human values by emphasising ethically grounded rules aligns with established principles of steering systems in cybernetics (known as the “ethical regulator theorem”) (Ashby, 2020). Key purpose-oriented artefacts became digital objects within the KMS training dataset, and the basis of which KMS-GPT reasoned. This included the mission and purpose of BlockScience, to “bridge the digital and physical worlds through innovative and sustainable engineering,” guiding the development and governance of safe, ethical, and resilient socio-technical systems, and leveraging emerging technologies to solve complex problems through the values of ethical design and innovation, diversity of perspective, integrity and autonomy, and positive-sum impact (Zargham and Ben-Meir, 2023). This core principle of aligning KMS-GPT as a system to its

purpose within its operating context—BlockScience as an organisation—guided the multiple ways that team members interacted with and governed the AI. The sections that follow show how accountability measures evolved and were implemented across not just technical and social but also institutional dimensions of the KMS-GPT experiment.

3.2. *AI policymaking*

While everyone online in the Slack channel was excitedly getting to know what KMS-GPT was capable of and prompting it with various questions like their newest friend, the organisational function responsible for overseeing Operations (known as “Ops Council”) stepped in. In a truly responsible manner, their suggestion was to ask KMS to “draft our policy around how to safely use LLMs,” GPT models, and similar AI tools.

In practice, ChatGPT was leveraged to draft the initial structure of the LLM policy. For full transparency, the BlockScience LLM/GPT Policy included its own disclosure: “ChatGPT (OpenAI’s GPT-3.5) was used to provide some structure and content of this policy document. While this policy document is internal only (so not an externally published material) and attribution is therefore not required by the policy, this note has been included for full transparency” (Republished in part, with permission).

The draft policy document stated: “The purpose of this policy is to establish guidelines for the safe and ethical use....” “Team members are permitted and encouraged to use LLMs for work-related purposes where applicable and where these tools are expected to improve efficiency, productivity, or a similar metric.” This was then augmented by inserting organisation-specific values and details. A key value of BlockScience is “our moral and ethical obligation to display intellectual integrity in our research and to design and maintain safe systems in our engineering work.” In alignment with this value, team members were instructed to use LLMs responsibly, ethically, and transparently. This included consideration for things like protection of confidential information (in training data and in interacting with a model), citation when used, and accountability through questions (and/or invoking the Operations Council) to address any uncertainties or potential violations.

Ops Team Member Daniel “Furf” stated: “Hi all—in working through our draft policy for LLMs, one of the open questions is around how to properly cite content provided by LLMs. @danlessa [Danilo] suggested that ‘we should have a standardised way of citing/referencing LLMs. It should include some basic information so that we can trace back the information, like a versioning identifier for both the source and trained model and optionally the query (or a pointer to the inputs that were used)’.”

Based on these guardrails, one feature that we did add is a link back to sources, to understand what sources of data KMS-GPT was drawing on to produce its response.

Furf then shared: “Here is a proposed draft for our BSci policy on the use of LLMs. This is definitely still a work in progress and input is most welcome. A couple of questions not yet addressed here which probably warrant further discussion include:

- What should our policy be for use of LLMs/GPT for internal communications? Or is it the same as for external/public materials? An example here is this policy document itself ... I used ChatGPT to help build the framework of the policy and then adapted to our own needs. Do I need to cite it as a source?
- The policy as written focuses on generated text. What should our policy be for generated code?
- It feels like there is a distinction between GPT providing language versus generating code; one feels more “creative” and the other is something more akin to using a calculator or other tool, but is this an accurate representation?
- Is the answer different depending on whether the code is used internally or is used for a client project, placed in a public repo, or similar?
- The policy as written covers all LLMs. Is there any difference if we are using KMS-GPT (which is informed based on our own materials) versus ChatGPT or another external LLM?”

Setting parameters such as the policy document were part of “constituting the algorithm.” In constituting KMS-GPT, we were determining the social, technical, and institutional boundaries of the system. This included determining which people are (and by extension, are not) subject to BlockScience’s governance and by extension KMS-GPT’s operation. It also included mapping the environment that BlockScience operates within as a private organisation, defining BlockScience’s purpose as an organisation, and thus clarifying the purpose of KMS-GPT, in order to motivate the ways that people interact with the AI agent. The policymaking process also considered accountability in terms of the hard technical limitations on the model, and consequences for the responsible human actors if things went wrong. For example, KMS-GPT was not to be used for external content generation or client projects while it was in this early phase of the experiment, as the organisation would ultimately be liable for misuse. However, part of the roadmap for development of the system, and permission to conduct this ethnography, was to figure out the parameters and limitations of the tool, and whether or not it was reliable enough to be used beyond a test-case environment.

3.3. Data management

KMS-GPT’s application within BlockScience required further training on relevant data within this specific organisational context. This raises all kinds of questions about data management and access. As Zargham stated to the developer team during an internal presentation, “the kind of protocol you’re creating is an ontology protocol...in particular, that allows ontologies to interoperate, even if they’re not consistent with one another.” This characterisation of AI is consistent with what researcher Nick Bostrom defines as an “oracle” model, a rational agent that is constrained to only act through answering questions (Bostrom, 2014). Researchers argue that even passive oracles, such as KMS-GPT, require goals and thus governance to infer how to reason in ways that are friendly to human values (Yudkowsky, 2008). KMS-GPT as an oracle was in no way ready to monitor its own alignment with the normative principles of its stakeholders in a trustworthy way.

A key turning point in the constitution of the system to reflect the values and organisational responsibilities of privacy and confidentiality occurred when the experiment came to the attention of the Head of Operations. “This is all amazing, and happening so quickly,” posted Nick Hirannet, Chief Operating Officer (COO) at BlockScience. “Putting a quick brake check on this...can you confirm...what data we’ve fed into it? I think one primary thing we need to get squared away is if we are in danger of violating any confidentiality/NDA stuff.” As raised by the COO, creating an LLM like KMS-GPT required being clear about what data went into training the model and thus what could be extracted out of it, and whether this information could be shared beyond “BlockScience’s walls.” In order to maintain a high level of confidentiality, only non-confidential and non-client specific information had been utilised in training the model. Specifically, the dataset included non-project specific data on the organisation’s drive, all sharable internal meeting recordings, all Slack channels that were public within the organisation, open GitHub repositories, as well as everything we had ever written and published in blogs and research papers. These were uploaded in accordance with existing data access control infrastructure and practices. Maintaining such settings was also dependent on Open AI’s policies not changing about how queries were sent to the model, and what information was stored for training.

In testing some of these data concerns and the limitations of KMS-GPT for personal privacy and censorship of certain information, team members asked the model questions, such as “@KMS who works for BlockScience?” Its responses were sound, perhaps due to limitations in the training data itself. KMS-GPT: “I don’t have specific information about the individual employees working at BlockScience. However, the team generally consists of experts specialised in data science, computational methods, engineering, research, and analytics for designing socio-technical economic systems and working on algorithm design problems with complex human behaviour implications.”

Still, these types of concerns prompted other features to be added to improve the legibility of the model.

For example, Orion announced to the team in Slack: “New KMS bot is up with some changes:

- New embeddings with better data, fast vector database, better semantic search technique, ~100× faster embedding time, improved embedding model, ~10× faster initialisation, vastly improved embedding parameters
- New memory model, should now be **much** better at being conversational, memory is per-thread
- Better UX, the bot now rewrites old messages so no more leftover “Processing...” messages.
- First implementation of context citations, the bot will show up to five sources that contributed to the answer (by including a line: “Generated in 10s from sources 1, 2, 3, 4.”)
- As an initial impression, I think we can improve this quite a bit, it is currently just the raw output from the semantic/vector search so can be quite noisy or seemingly irrelevant “_ (ヾ) _.”

3.4. *Localised negotiations*

The process of constituting KMS-GPT came through iterative interactions and negotiations from multiple stakeholders, and technical, social, and institutional formalisation of these. The governance of KMS was bounded by the local scale and context of the organisation. The actors within BlockScience, along with the institution itself as a relatively small, private organisation, were able to be held accountable for the operational and governance considerations surrounding KMS-GPT. As a result, constituting an AI according to its purpose (or regulating the model in line with the ethics of its purpose, stakeholders, and environment) was effective under the ownership and within the bounds of BlockScience as the given decision-making authority. It was possible to elicit preferences from local stakeholders, negotiate the trade-offs and prioritise between these, adjust the model and/or data infrastructure, and maintain oversight of its operation. This was because there was a degree of localisation to the scale that the AI model was deployed and governed at (otherwise known as the principle of subsidiarity, or governance at the most local level, whereby people have domain-specific knowledge that is applicable to the operating context (Drew and Grant, 2017)).

3.5. *Mapping how AI systems are constituted: key considerations and insights*

The following table seeks to present some considerations for mapping how an AI system is constituted (such as the context, values, and architecture), as well as the accountability norms and mechanisms of an LLM that is deployed among a defined set of stakeholders. Ethnographic insights are presented in relation to key considerations during the development and deployment of KMS-GPT, and the emergent approach of AI as a constituted system of technical, social, and institutional components and processes.

Viewing AI as a constituted system expands the scope of accountability beyond mere technical compliance or ethical guidelines to a more comprehensive, system-wide approach that encompasses social, institutional, technical, economic, organisational, and legal components. The framework provided in [Table 1](#) extends this literature by mapping a more detailed and systematic way to work through some of these considerations. Technical accountability refers to the creation of transparent, interpretable, and auditable AI systems. Developers must employ methodologies that allow for the tracing of decisions made by AI systems back to understandable and justifiable processes, ensuring that AI operations can be scrutinised and evaluated against established norms and standards. Social accountability in AI necessitates that AI systems are developed and deployed in ways that are sensitive to the norms and values of stakeholders involved in their use and outputs. This requires mechanisms for community engagement and participatory design that allow for the voices of diverse stakeholder groups to be heard and integrated into the AI development process. Meanwhile, institutional accountability refers to the role of internal governance within entities that develop and deploy AI, including the adoption of policy guidelines and the implementation of practices that ensure responsible AI use. Externally, organisational accountability requires the establishment and enforcement of policies and standards that govern AI development and use, necessitating a collaborative effort between organisations, policy makers, and regulators across multiple scales to create frameworks that ensure AI systems are developed and used responsibly, and in adherence

with the law. Thus, accountability is not just about holding an AI model to account but understanding the broader governance and relational networks in which it is entangled, how these are imbued in the technology itself, and then what specific technical or social-institutional aspects of the system as a whole can be modified in order to govern and align behaviour. It is within this context, that who is responsible and accountable, and how this can be verified and enforced becomes possible.

A large part of what made it possible to map the social and technical context of KMS-GPT and to trace accountability relationships for the behaviour of an algorithmic model is that it was bounded to the environment of a specific organisation. The architecture of KMS-GPT allowed for gradual scale-up of the governance boundaries of the AI to effectively negotiate, monitor, and iterate on accountability measures. The approach of AI as a constituted system localises accountability measures to traceable feedback loops between roles, responsibilities, and behaviours. Accountability occurred at the institutional level because there were clear boundaries within a private organisation in terms of stakeholder principles, clear authority for monitoring AI compliance with these principles, and responsive feedback loops.

Part of the initial success of KMS-GPT was that it was bounded to the values, norms, and governance of a small organisation. Governing AI involves social and institutional dynamics that extend beyond the technical limitations of an AI model and cannot be delegated to an AI. For BlockScience, this web of accountabilities was relatively manageable due to its localised nature, defined set of stakeholders, and clear overarching organisational objectives. Thus, accountable AI governance may be more complicated beyond the context of a relatively small organisation. The ability to understand how a model is constructed, to verify that it is operating in line with normative principles, or intervene to change it becomes much more complicated for general purpose AI at a larger scale, as imagined in the “constitutional AI” approach of natural language instructions that are compiled by large numbers of diverse people groups (such as “Americans” in an experiment run by consumer LLM company Anthropic which polled 1000 people (Anthropic, 2023)). The real challenges surface when these AIs operate in less bounded, global contexts where a multiplicity of actors and goals makes it harder to delineate responsibilities, monitor behaviours, and enforce consequences, furthering the argument for purpose and context specific, locally governed AI models. If KMS-GPT/s was to operate across multiple organisations, at multiple scales, then understanding the specific requirements of the context, exacting the preferences of heterogeneous constituents, constituting the algorithm, monitoring its behaviour, and enforcing consequences (such as policy changes) in response to misbehaviour becomes far more complex. With multiple decision-making authorities and stakeholders, and more indirect actor-forum relationships and communication channels, determining the purpose of a model, constituting it according to the preferences of multiple people and organisations, querying its design, data training practices, and behaviours, and enforcing policies on it would be significantly more difficult (which is why some have proposed blockchain technology as an institutional scaffolding to ensure economic alignment between an AI agent and human interests (Berg et al., 2023)), introducing a whole new set of design challenges around resources, incentives, implementation, and enforcement when things go wrong).

3.6. From “constitutional AI” to “AI as a constituted system”

The approach of conceptualising AI as a constituted system presented in this article provides a much broader scope for AI governance discussions, as opposed to the approach of instructing an AI model with a natural language constitution to align it with human values. For example, the approach of “constitutional AI” was coined by the team developing the Open AI competitor “Anthropic” to describe the practice of provisioning a list of rules or principles as a way to train AI agents to operate in line with the preferences of their deployer, and self-monitor their behaviour against these rules (Bai et al., 2022; Durmus et al., 2023). The obvious question here is how does a group of people, such as an entire democratic country, collectively express, articulate, and determine their preferences to be algorithmically constituted? These are also questions that the likes of Open AI’s “superintelligence alignment” team is thinking about. Co-lead of Open AI’s Superalignment team, Jan Leike, proposes that society’s values can be “imported” into language models to determine how they navigate subjective questions of “value” (in terms of personal

values, or ethics) and align LLMs to group values and preferences through deliberative democracy simulations or “mini publics” to write AI constitutions (Leike, 2023). While this offers some normative accountability, it does not provide concrete accountability mechanisms, or actors to hold accountable, for more trustworthy AI. In contrast, AI-as-a-constituted-system provides a contrast to some of the experiments being conducted by major AI developers for broad scale, public, and “democratic” inputs to AI as a GPT (Anthropic, 2023; Zaremba et al., 2023), to something more domain, task, and constituent specific.

Furthermore, Big Tech governance approaches assume the widespread adoption of universal, general-purpose AI models. In terms of aligning AI with human interests, this then warrants large-scale expression of semi-homogenous social values, and these principles are being expressed in natural language for agents to interpret and abide by. In contrast, approaching AI as a smaller scale, custom purpose, constituted system allows for far greater flexibility in terms of how local models can be constituted by a certain group of stakeholders, and for a specific purpose. This allows for the trade-offs between heterogeneous social values and accountability measures for action in accordance with these values to be collectively negotiated by human stakeholders, or at least deliberately and explicitly deferred to an AI model. Thus, when AI is approached as a constituted system, the boundaries, measures, and mechanisms of accountability can be more comprehensive than what can be expressed through a natural language constitution alone for alignment. AI as a constituted system enables designers to encompass social, technical, and institutional boundaries (such as bespoke training models, data privacy policies, data access restrictions, permissions, written rule sets, server infrastructure, and so on). Further research directions include at what scales various modalities of accountability apply (technical, social, institutional), in what combinations these mechanisms apply at various scales, and what dimensions remain to be explored in more depth (such as economic, legal, material, and environmental). Scope could also include deeper analysis of the advantages and trade-offs in explainability and control between various commercially provided LLM foundation models, and how this ultimately may impact accountability.

4. Part 3: Resurrection

It turns out that AI operates wherever people decide to have engineers build and deploy it. “AI is always a system with inputs, outputs, sensors, and actuators,” stated Zargham on a team “Governance Research Pod” call. It is always tactical (meaning able to pursue goals); it does not have strategic autonomy (as the ability to set goals) (Zargham et al., 2023b). AI always has encoded objectives and constraints that it has been trained and instructed on.” In other words, according to anthropologists “all knowledge work is situated practice” (Elish and Boyd, 2018).

It is only by locating AI as a social and technical assemblage in its constitutive context of people, purpose, and environmental context that we can begin to think about accountability relationships, accountability mechanisms, and enforceable lines of responsibility. Within the institutional context and organisational bounds of BlockScience, KMS-GPT had a clear purpose, constituency that it was accountable to, and governing authority. These social and organisational contexts provided boundaries in terms of how the AI model subjectively “should” act, as well as when institutional policies, expressed in infrastructure, were required to create and reinforce accountability. Constituting the AI was heavily reliant on the social norms inherent in the organisational culture at a local level, including commitment to an iterative process of experimentation. KMS-GPT relied on human processes to collectively determine and govern subjective goals and values. Furthermore, operational practices around data management infrastructure required ongoing care and attention to maintain accountability within the local context. Oftentimes the Ops Council, in representations of the broader forum of BlockScience, would query Orion as the project lead and a proxy to explain KMS-GPT’s behaviour. This shows how AI (in its current state) is dependent on its deployers to create, instruct, maintain, and modify it. Thus, accountability still occurs between human stakeholders, with AI agents as an intermediary but not a substitute for human explanation.

4.1. “Project Goldfish”

At the informal conclusion of the internal experiment, KMS-GPT was considered one of the most exciting projects across the entire team. Following a number of internal conversations about goals, future directions, and possible next steps, the team decided to dedicate the internal resources to revive it. They re-named the model from “KMS-GPT” to “KMS-LLM,” as other foundation models (such as “Mistral”) had been explored, and the technicians wanted to avoid lock-in and maintain flexibility in the underlying technical architecture, rather than being tied to GPT.

They also called it code-name “Project Goldfish” (because every successful project needs a code-name and an emoji 🐟). The story goes that the name was to invoke the revocation of the experiment that took place and transform it into a full-blown project. David explained, “In the United States, goldfish are a very common thing, but they have a precarious nature in terms of a tendency to die. They are pets. It’s fun. It swims around. And then you lose it. And you want another one” (which sounds like a childhood trauma). This is a parallel description of what happened with the first LLM integration with KMS. Moving forward, KMS-GPT would be allocated internal research and development resources and scaffolded with a proper project management plan and infrastructural set up (including getting the model off Orion’s laptop and onto a persistent cloud server). The goldfish was expected to change in colour and size in terms of delivering on knowledge management and collective cognition but not disappear from 1 day to the next. Once the team knows they can get it “from 0 to 1,” the next phase is to try to connect it with the knowledge systems of other organisations, or in Zargham’s words, “to build a bigger fishbowl.”

To me, the goldfish was about more than this. The goldfish was not KMS-GPT, the goldfish was me. It was my memory, my brain capacity, and my experience that simultaneously reflects the augmentation of human cognitive capabilities through machine assistance, and the desire to do this in responsible ways. Instead of needing to remember where the travel reimbursement was filed in the team drive or needing to submit a proofread request to the communications team 2 weeks in advance of needing it (never happens), GPT was my mental aid. It was not independent of me.

BlockScience had reconstructed knowledge management throughout the organisation, changing the fundamental ways we perceive the world and relate to one another as a group of people and an organisation. The goal of this project moving forward was not to try to stop GPT from disappearing, it was for us to try to understand how our systems and roles were being replaced and restructured, how to determine what “shared” aims are, and how to steer this in constructive ways towards these aims. Project Goldfish also comes with a tagline: “developing a deployable, governable, interoperable [between organisations], organisation level, KMS.” AI becomes “governable” (or not) in the context of underlying (and not new) data management practices. As stated by Zargham on a semi-public research presentation on the KMS-GPT experiment, the focus becomes “what data are we collecting, how we are governing it, using it, and what is coming out through it. It’s not the LLM, it’s the whole assemblage that we are deciding, that we are governing.” Orion added “LLMs as an interface are only as good as the underlying data structure.”

Extrapolating the learning from KMS-GPT to other contexts would be very difficult in unbounded, global forums to create and maintain AI accountability, given that localised context and iterative experimentation were key features of the constitutional process. Yet, the next step in the experiment was for another organisation, an open research collective, to deploy and govern their own instance of a knowledge management GPT (code named “Knowledge Organisation Infrastructure” or “KOI pond” as a pun on Koi fish), and to allow the models to interact to create a “local knowledge network” (Zargham et al., 2024).

5. Conclusion

This article has ethnographically investigated how a customised Chat-GPT LLM is established within the social, technical, economic, organisational, and legal dimensions of an organisation to create accountability. It argues for an approach to LLM governance that focuses on AI as a constituted system—a

dynamic assemblage shaped by ongoing, iterative negotiations and decisions among stakeholders that establish and uphold accountability norms and mechanisms. The research extends beyond conventional AI governance discussions, which often emphasise universal technical solutions, to examine the intricate networks that connect humans, machines, and institutions to create and maintain accountability. The ethnographic approach adopted provides a nuanced view of the interactions between technical and social elements within a specific organisational context, challenging one-size-fits-all approaches to AI alignment with human values and governance. A practical framework is introduced for practitioners, institutions, and policymakers to map the social, technical, and institutional dimensions of LLMs as constituted systems. This framework helps to articulate context-specific operations and underlying values and connect these to tangible social and technical accountability processes and mechanisms. Thus, this paper emphasises the importance of understanding AI as a system shaped by diverse social and technical inputs, where accountability is continuously negotiated among various actors, processes, and maintenance practices, allowing for clearer identification of relationships and accountability mechanisms within AI systems. By taking a context-specific and multidimensional approach to configuring algorithmic systems, this study shifts the focus from universally applicable AI governance solutions to more tailored, context-specific strategies that clearly define stakeholders and accountability measures. As such, this paper invites further research into the qualitative and quantitative aspects of AI as a constituted system, including the investigation into the applicability of this approach across various scales and contexts as new LLM models and technical advancements emerge.

Acknowledgements. The author is grateful to the team at BlockScience for research access and feedback, especially Michael Zargham, Orion Reed, David Sisson, Luke Miller, Furf, and Jakob Hackel. Thank you also to Sebastian Benthall for feedback.

Author contribution. Writing: Original draft—K.N.; Writing: Review and editing—K.N.

Provenance statement. This article is part of the Data for Policy 2024 Proceedings and was accepted in Data & Policy on the strength of the Conference's review process.

Competing interest. The author declares none.

References

- Anthropic** (2023) "Collective Constitutional AI: Aligning a Language Model with Public Input". Anthropic (blog). October 17. Available at <https://www.anthropic.com/index/collective-constitutional-ai-aligning-a-language-model-with-public-input> (accessed 12 January 2024).
- Ashby M** (2020) Ethical regulators and super-ethical systems. *Systems* 8(4), 53.
- Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, Chen A, Goldie A, Mirhoseini A, McKinnon C, Chen C, Olsson C, Olah C, Hernandez D, Drain D, Ganguli D, Li D, Tran-Johnson E, Perez E, Kerr J, Mueller J, Ladish K, Landau J, Ndousse K, Lukosuit K, Lovitt L, Sellitto M, Elgag N, Schiefer N, Mercado N, DasSarma N, Lasenby R, Larson R, Ringer S, Johnston S, Kraven S, El Showk S, Fort S, Lanham T, Telleen Lawton T, Conerly T, Henighan T, Hume T, Bowman SR, Hatfield-Dodds Z, Mann B, Amodei D, Joseph N, McCandlish S, Brown T and Kaplan J** (2022) Constitutional AI: Harmlessness from AI feedback. [arXiv:2212.08073](https://arxiv.org/abs/2212.08073); <https://doi.org/10.48550/arXiv.2212.08073>.
- Bakker M, Chadwick M, Sheahan H, Tessler M, Campbell-Gillingham L, Balaguer J, McAleese N, Glaese A, Aslanides J, Botvinick M and Summerfield C** (2022) Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems* 35, 38176–38189.
- Berg C, Davidson S and Potts J** (2023) "Institutions to Constrain Chaotic Robots: Why Generative AI Needs Blockchain" Available at <https://ssrn.com/abstract=4650157> (accessed 16 December 2023).
- Blackwell AF** (2021) Ethnographic artificial intelligence. *Interdisciplinary Science Reviews* 46(1–2), 198–211.
- Bostrom N** (2014) *Superintelligence: Paths, Dangers, Strategies*. United Kingdom: Oxford University Press.
- Bovens M** (2007) Analysing and assessing accountability: A conceptual framework. *European Law Journal* 13(4), 447–468.
- Bullock JB, Chen YC, Himmelreich J, Hudson VM, Korinek A, Young MM and Zhang B** (2024) Introduction. In Bullock JB, Chen Y-C and Himmelreich J (eds.), *The Oxford Handbook of AI Governance*. United Kingdom: Oxford University Press, pp. 1–20.
- Cai D, Wang Y, Liu L and Shi S** (2022) Recent advances in retrieval-augmented text generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*, New York, NY, USA: Association for Computing Machinery, pp. 3417–3419. <https://doi.org/10.1145/3477495.3532682>.

- Chablo A** (1996) What can artificial intelligence do for anthropology? *Current Anthropology* 37(3), 553–555. <https://doi.org/10.1086/204518>.
- Crawford K** (2021) *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.
- Crawford K and Joler V** (2018) *Anatomy of an AI System: The Amazon Echo As An Anatomical Map of Human Labor, Data and Planetary Resources*. AI Now Institute and Share Lab, (September 7, 2018). Available at <https://anatomyof.ai> (accessed 23 September 2023).
- Dafoe A** (2024) AI governance: Overview and theoretical lenses. In Bullock JB, Chen Y-C and Himmelreich J (eds.), *The Oxford Handbook of AI Governance*. United Kingdom: Oxford University Press, pp. 21–44.
- Dell’Acqua F, McFowland E, Mollick ER, Lifshitz-Assaf H, Kellogg K, Rajendran S, Krayner L, Candelon F and Lakhani KR** (2023) Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. Harvard Business School Technology & Operations Mgt. Unit Working Paper No. 24–013. Available at <https://ssrn.com/abstract=4573321> (accessed 22 December 2023).
- Drew J and Grant B** (2017) Subsidiarity: More than a principle of decentralization—A view from local government. *Publius: The Journal of Federalism* 47(4), 522–545. <https://doi.org/10.1093/publius/pjx039>.
- Durmus E, Nyugen K, Liao TI, Schiefer N, Askell A, Bakhtin A, Chen C, Hatfield-Dodds Z, Hernandez D, Joseph N, Lovitt L, McCandlish S, Sikder O, Tamkin A, Thamkul J, Kaplan J, Clark J and Ganguli D** (2023) “Towards Measuring the Representation of Subjective Global Opinions in Language Models.” <https://doi.org/10.48550/arXiv.2306.16388>.
- Dwivedi YK, Hughes L, Ismagilova E, Aarts G, Coombs C, Crick T, Duan Y, Dwivedi R, Edwards J, Eirug A, Galanos V, Iavarasan PV, Janssen M, Jones P, Kar AK, Kizgin H, Kronemann B, Lal B, Lucini B and Williams MD** (2021) Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management* 57, 101994. <https://doi.org/10.1016/j.ijinfomgt.2019.08.002>.
- Elish MC and Boyd D** (2018) Situating methods in the magic of big data and AI. *Communication Monographs* 85(1), 57–80. <https://doi.org/10.1080/03637751.2017.1375130>.
- Forsythe D** (2001) *Studying those Who Study us: An Anthropologist in the World of Artificial Intelligence*. Indiana: Stanford University Press.
- Golden TD** (2013) Sociotechnical theory. In Kessler EH (ed.), *Encyclopaedia of Management Theory*. Indiana SAGE Publications, Ltd, pp. 752–755. <https://doi.org/10.4135/9781452276090>.
- Hess DJ** (2001) Ethnography and the development of science and technology studies. In Atkinson P, Coffey A, Delamont S, Lofland J and Lofland L (eds.), *Sage Handbook of Ethnography*. Thousand Oaks, CA: SAGE Publications, pp. 234–245.
- Hinings CR, Logue D and Zietsma C**. (2017) Fields, institutional infrastructure, and governance. In Greenwood R, Oliver C, Lawrence T and Meyer R (eds.), *The SAGE Handbook of Organisational Institutionalism*. Thousand Oaks, CA: Sage Publications Inc, pp. 163–189.
- Holmström J and Hällgren M** (2022) AI management beyond the hype: Exploring the co-constitution of AI and organizational context. In Atkinson P, Coffey A, Delamont S, Lofland J and Lofland L (eds.), *The Sage Handbook of Ethnography*. Thousand Oaks, CA: Sage Publications, pp. 1575–1585.
- Intellectronica** (2023) “Grounding LLMs”. [TECHCOMMUNITY.MICROSOFT.COM](https://techcommunity.microsoft.com/t5/fasttrack-for-azure/grounding-llms/ba-p/3843857) (blog). June, 9, 2023. Available at <https://techcommunity.microsoft.com/t5/fasttrack-for-azure/grounding-llms/ba-p/3843857> (accessed 29 August 2023).
- Lechterman TM** (2024) “The concept of accountability in AI ethics and governance. In Bullock JB, Chen Y and Himmelreich J (eds.), *The Oxford Handbook of AI Governance*. United Kingdom: Oxford University Press, pp. 164–182.
- Leike J** (2023) “A proposal for importing society’s values. *Musings on the Alignment Problem*.” *Substack* (blog). March 9. Available at <https://aligned.substack.com/p/a-proposal-for-importing-societys-values> (accessed 2 December 2023).
- Marda V and Narayan S** (2021) On the importance of ethnographic methods in AI research. *Nature Machine Intelligence* 3(3), 187–189.
- Mead M** (1968) *Cybernetics of cybernetics*. In von Foerster H, White J, Peterson L and Russell J (eds.), *Purposive Systems*. New York: Spartan Books.
- Mittelstadt BD, Allo P, Taddeo M, Wachter S and Floridi L** (2016) The ethics of algorithms: Mapping the debate. *Big Data & Society* 3(2), 1–21. <https://doi.org/10.1177/2053951716679679>.
- Munk AK, Olesen AG and Jacomy M** (2022) The thick machine: Anthropological AI between explanation and explication. *Big Data & Society* 9(1), 1–14. <https://doi.org/10.1177/20539517211069891>.
- Nabben K** (2023) Web3 as ‘self-infrastructuring’: The challenge is how. *Big Data & Society* 10(1), 1–6. <https://doi.org/10.1177/20539517231159002>.
- Orlikowski WJ** (2007) Sociomaterial practices: Exploring technology at work. *Organization Studies* 28(9), 1435–1448. <https://doi.org/10.1177/0170840607081138>.
- Pink S, Horst H, Postill J, Hjorth L, Lewis T and Tacchi J** (2016) *Digital Ethnography. Principles and Practice*. Thousand Oaks, CA: SAGE Publications Ltd.
- Richardson K** (2020) The complexity of otherness: Anthropological contributions to robots and AI. In Dubber MD, Pasquale F and Das S (eds.), *The Oxford Handbook of Ethics of AI*. United Kingdom: Oxford Academic. <https://doi.org/10.1093/oxfordhb/9780190067397.013.36> (accessed 2 September 2023).

- Risse M** (2022) Artificial intelligence and the past, present, and future of democracy. In S. Voeneke, Kellmeyer P, Mueller O and Burgard W (eds.), *The Cambridge Handbook of Responsible Artificial Intelligence*. Cambridge: Cambridge University Press, pp. 85–103.
- Sartori L and Theodorou A** (2022) A sociotechnical perspective for the future of AI: Narratives, inequalities, and human control. *Ethics and Information Technology* 24(1), 4. <https://doi.org/10.1007/s10676-022-09624-3>.
- Schwartz S, Yaeli A and Shlomov S** (2023) Enhancing Trust in LLM-Based AI Automation Agents: New Considerations and Future Challenges. <https://doi.org/10.48550/arXiv.2308.05391>.
- Star SL** (1999) The ethnography of infrastructure. *American Behavioral Scientist* 43(3), 377–391. <https://doi.org/10.1177/00027649921955326>.
- Suchman L** (2000) Organizing alignment: A case of bridge-building. *Organization* 7(2), 311–327.
- Suchman L** (2002) Located accountabilities in technology production. *Scandinavian Journal of Information Systems* 14(2), 91–105. Available at <https://aisel.aisnet.org/sjiss/vol14/iss2/7>.
- Sumathi S, Manjubarkavi S and Gunanithi P** (2023) Ethnography and artificial intelligence. In Sahoo M, Jeyavelu S and Jurane A (eds.), *Ethnographic Research in the Social Sciences*, 1st Edn. Routledge: India, pp. 145–157.
- Wieringa M** (2020) What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. New York, NY: Association for Computing Machinery, pp. 1–18. <https://doi.org/10.1145/3351095.3372833>.
- Yudkowsky E** (2008) *Dreams of Friendliness*. *Less Wrong (blog)*. Available at <https://www.lesswrong.com/posts/wKnwejJGriTS9QxxL/dreams-of-friendliness> (accessed 9 February 2024).
- Yudkowsky E** (2016) “The AI alignment problem: why it is hard, and where to start”. *Intelligence.org* (Transcript). Available at <https://intelligence.org/files/AlignmentHardStart.pdf> (accessed 3 December 2023).
- Zaremba W, Dhar A, Ahmad L, Eloundou T, Santurkar S, Agarwal S and Leung J** (2023) “Democratic inputs to AI”. *OpenAI*. March 25. Available at <https://openai.com/blog/democratic-inputs-to-ai> (accessed 12 January 2023).
- Zargham M** (2023) *A Language for Knowledge Networks*. *September 1*. *BlockScience (blog)*. Available at <https://blog.block.science/a-language-for-knowledge-networks/> (accessed 1 September 2023).
- Zargham M, Alston E, Nabben K, Ben-Meir I** (2023a) “What Constitutes a Constitution?” *April 6*. *BlockScience (blog)*. Available at <https://blog.block.science/what-constitutes-a-constitution/> (accessed 1 September 2023).
- Zargham M and Ben-Meir I** (2023) The Animating Purpose of BlockScience. June 14. *BlockScience (blog)*. Available at <https://blog.block.science/the-animating-purpose-of-blockscience/> (accessed 1 September 2023).
- Zargham M, Ben-Meir I and Nabben K** (2024) “Knowledge Networks and the Politics of Protocols”. *BlockScience (blog)*. Available at <https://blog.block.science/knowledge-networks-and-the-politics-of-protocols/> (accessed 1 September 2023).
- Zargham M, Zartler J, Nabben K, Goldberg R and Emmett J** (2023b) “Disambiguating Autonomy: Ceding Control in favor of Coordination in Cybernetic Organizing.” *BlockScience Medium (blog)*. Available at <https://medium.com/block-science/disambiguating-autonomy-ca84ac87a0bf> (accessed 1 September 2023).