Data & Policy (2024), 6: e62
doi:10.1017/dap.2024.74

DATA FOR POLICY PROCEEDINGS PAPER

Signalling and rich trustworthiness in data-driven healthcare: an interdisciplinary approach

Jonathan R. Goodman[1] and Richard Milne[2,3]

[1]Leverhulme Centre for Human Evolutionary Studies, University of Cambridge, Cambridge, UK
[2]Kavli Centre for Ethics, Science, and the Public, Faculty of Education, University of Cambridge, Cambridge, UK
[3]Wellcome Connecting Science, Cambridge, UK
Corresponding author: Jonathan R. Goodman; Email: jrg74@cam.ac.uk

Received: 27 November 2023; **Revised:** 13 March 2024; **Accepted:** 08 April 2024

Keywords: AI; health data; signallingtrust; trustworthiness

Abstract

Discussions of the development and governance of data-driven systems have, of late, come to revolve around questions of trust and trustworthiness. However, the connections between them remain relatively understudied and, more importantly, the conditions under which the latter quality of trustworthiness might reliably lead to the placing of 'well-directed' trust. In this paper, we argue that this challenge for the creation of 'rich' trustworthiness, which we term the Trustworthiness Recognition Problem, can usefully be approached as a problem of effective signalling, and suggest that its resolution can be informed by a multidisciplinary approach that relies on insights from economics and behavioural ecology. We suggest, overall, that the domain specificity inherent to the signalling theory paradigm offers an effective solution to the TRP, which we believe will be foundational to whether and how rapidly improving technologies are integrated in the healthcare space. We suggest that solving the TRP will not be possible without taking an interdisciplinary approach and suggest further avenues of inquiry that we believe will be fruitful.

Policy Significance Statement

Public trust in the governance of health data systems and AI tools should be placed in institutions and organisations that are trustworthy. However, we have no model for understanding how trustworthy institutions can demonstrate themselves to be so. In this paper, we draw on interdisciplinary studies of social trust and behavioural ecology to propose an approach to this 'trustworthiness recognition problem' that examines the costs associated with effective signalling of the characteristics of trustworthiness.

1. Introduction

Discussions of the development and governance of data-driven systems have, of late, come to revolve around questions of trust and trustworthiness. Each of these terms has received substantial recent attention in the policy and academic literature, particularly in the context of the relationship between data science, data policy and the public. However, the connections between them remain relatively understudied and, more importantly, the conditions under which the latter quality of trustworthiness might reliably lead to the placing of 'well-directed' trust. In this paper we argue that this challenge, which we term the

https://doi.org/10.1017/dap.2024.74 Published online by Cambridge University Press

'trustworthiness recognition problem' can usefully be approached as a problem of effective signalling, and suggest that its resolution can be informed by taking an interdisciplinary approach that relies on work in economics and behavioural ecology.

In the development of AI the European High-Level Expert Group, among others, has set out conditions for 'trustworthy AI'—placing trust at the centre of AI development, despite debates about the nature and locus of this trust (Braun et al., 2021; Gille et al., 2020). The recent development of systems for health data research increasingly places 'trusted' research environments at their heart (although see Graham et al., 2022), while the UK government's agreement with the US company Palantir on the NHS federated data platform has prompted concerns about the consequences for 'patient trust' (Armstrong, 2023).

The UK government's 2022 *Data Saves Lives* white paper makes "Improving trust in the health and care system's use of data" its first goal – partly in response to the lack of trust identified as a core failing of the UK's abortive care.data plans in 2014 (Carter et al., 2015).

In such contexts, trust plays an important role because of the inability of individuals to provide specific consent to individual uses of data. Instead, they are asked to allow others to make decisions on their behalf. Empirical research with members of the public has repeatedly shown that while many are willing for their health data to be used in research, a lack of trust can be a significant barrier to enabling this, and that the willingness to place trust is not equally distributed across individuals, social groups or countries (Kalkman et al., 2022; Milne et al., 2019; Wellcome Global Monitor, 2018).

More recently, attention has shifted from questions of public trust in health data systems to the *trustworthiness* of these systems. This shift in emphasis recognises that trustworthiness is a quality of an individual or an institution and relatedly that trust is a relational construct between two parties. There are three integral parts to the relationship, commonly conceptualised: the trustor, the trustee and the domain in which they are trusted. Whereas studies of 'public trust' emphasise the former of these, it is often the 'trustee' who is both concerned about public trust and who, crucially, has the ability to act on the nature of this three-part relationship. Recognising the multiple parts of this relationship—including the nature of the relationship itself—is, we argue, central to understanding the conditions for 'well-directed trust' in which trust is placed and refused intelligently (O'Neill, 2018).

In this paper, we present an approach to trustworthiness in the governance of health data. Our aim is to support the development of systems in which trust is 'well-directed', in which trustworthy governance of data and artificial intelligence systems is accompanied by the ability to effectively signal that it is trustworthy in terms of its commitments and responsiveness. Approaching this from the perspective of patients and the public, we set out what we term the 'trustworthiness recognition problem' (TRP): the challenge of distinguishing institutions and people that are trustworthy from those that are not, in circumstances in which the former may not be able to demonstrate that quality while the latter may signal that they are (Figure 1).

We draw on existing literature in two areas. First, in section 1, we elaborate the TRP and the importance of 'rich trustworthiness' for a system of well-directed trust in which individuals and organisations are able to cooperate effectively. In sections 2–4, we consider how an understanding of effective signals of trustworthiness can be informed by a substantive body of work in economics and behavioural ecology. Finally, in sections 5 and 6, we draw out the implications of this for AI and health data governance, arguing for the need to contextualise what we call the costs of trustworthiness in the healthcare space.

2. Trust and trustworthiness

The question of trust has long been important for the social sciences, recognised as central to social cohesion and the effective cooperative functioning of complex societies in which we have less opportunity to directly interact with those upon whom we rely (Giddens, 1991; Putnam et al., 1993). Trust relationships are particularly relevant in situations of uncertainty and vulnerability, when we need others to act in a specific domain in ways that we hope — but cannot guarantee — that they will. In these situations, trust reduces or minimises risk — enabling the 'leap of faith' (Möllering, 2001) or 'leap in the dark' (Hawley, 2019). For Giddens (Giddens, 1991), trust is essentially confidence in the reliability of a

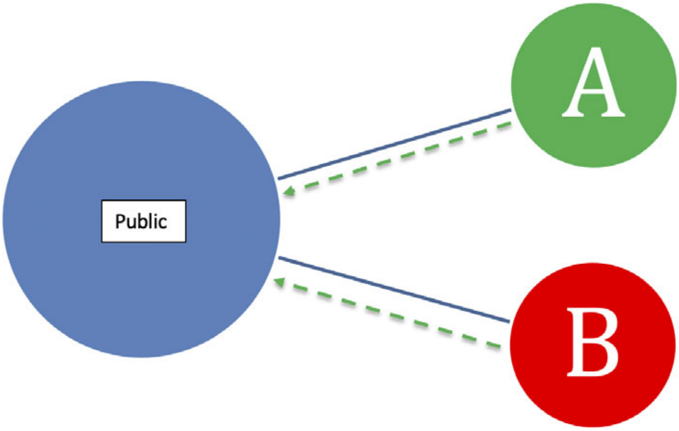

Figure 1. *The TRP. Institutions A and B (A is trustworthy, B is not) signal the same information to the public (dashed green line), which aims to demonstrate trustworthiness. The public must choose which institution to trust; in the absence of the ability to differentiate between A and B's signals, it cannot effectively recognise which institution is trustworthy.*

person or system, regarding a given set of outcomes or events, based on faith in their honour or love, or in the correctness of the technical knowledge that governs expert systems.

However, what distinguishes trust from other types of relationships we have, such as reliance or confidence, has been a matter of continued debate. Scholars have pointed, inter alia, to the role of commitments, imputed motives, emotions, goodwill or encapsulated interests in explaining what differentiates a trust relationship from one of simple reliance (Baier, 1986; Hardin, 2002; Hawley, 2019; Jones, 1996). These debates have direct consequences for practice in domains such as healthcare and health data systems in which trust relationships are fundamental (Gille et al., 2015).

However, the placement of trust is only one part of a trust relationship, recognised in the growing research and policy interest in trustworthiness. As Gille et al. (2020) have pointed out that in the case of the EU High-Level Expert Group (HLEG)'s guidance on trustworthy AI, policy interventions jump from establishing the importance of trust to describing the qualities of trustworthy systems without examining how trusting relationships might be established, or on what basis. The example of the HLEG captures a shift to focus on what it is that makes the trustee an appropriate recipient of trust. Trusting involves behaving as if the recipient of trust were trustworthy, but as O'Neill has pointed out,

> 'Trust is valuable when placed in trustworthy agents and activities, but damaging or costly when (mis)placed in untrustworthy agents and activities. So it is puzzling that much contemporary work on trust – such as that based on polling evidence – studies generic attitudes of trust in types of agent, institution or activity in complete abstraction from any account of trustworthiness'. (O'Neill, 2018).

We suggest that trustworthiness, in its broadest possible sense, refers to a directly unobservable quality of an individual or agent that determines their acting in a way that favours the interest of a given agent in a particular domain.[1] For example, when someone gives permission for their health data to be shared between their GP surgery and a hospital system for their care but opts-out of allowing their data to be used for research or planning, the NHS is trustworthy insofar as it does not allow the data to be used for these purposes, sell the patient's data to a third party or e-mail the patient's private information to that person's

[1] Relatedly, the fact that trustworthiness is unobservable directly renders it at risk of mimicry. See Goodman (2023) for an analysis of this issue in the setting of human cooperation.

https://doi.org/10.1017/dap.2024.74 Published online by Cambridge University Press

employer. The patient trusts the service to use the data as agreed; the service is trustworthy insofar as it does not betray this trust.

A focus on trustworthiness leads to a valuable discussion of the desirable qualities and characteristics of individuals or institutions, and again what it is that distinguishes this from merely being 'reliable'. Katherine Hawley, for example, points to the importance of being a 'good promisor' who avoids commitments they cannot fulfil and can be relied upon to fulfil those they make and is cognisant of the limits of his or her competence (Hawley, 2019). O'Neill (2018) similarly emphasised the importance of competence, honesty and reliability. She also, however, points to the importance of communication in enabling trustors to make judgements of potential trustees' honesty in commitments, competence at relevant tasks and, finally, reliability in honesty and competence.

Just as trust is a domain-specific quality, the same restriction applies to trustworthiness (Jones, 2012). We place trust in individuals and institutions to perform the functions we need or desire from them, and outside of these relations, this 'trust' may no longer apply. Thus, for example, whether a scientist is a faithful friend or romantic partner is not related to that person's trustworthiness as a scientist, and similarly, whether a person is a faithful romantic partner is not related to whether that same person is a trustworthy scientist. In her discussion of trustworthiness, however, Jones (2012) moved beyond a straightforward question of what constitutes trustworthiness—in her argument, the responsiveness to dependency in a specific domain — to consider the responsibilities associated with trustworthiness, which she elaborates through the notion of 'rich trustworthiness'. Jones argues that given our limitations as trusting agents in deciding how to trust, we need more from others than simply that they are trustworthy. Instead, we want those who can be trusted to *identify* themselves so that we can place our trust wisely.

This presents a challenge to trustworthy institutions and individuals have a problem, which we term the TRP. On the one hand, institutions and people that are trustworthy may not be called upon as trustees—at times because of prejudice, but at times because they are unable to reveal that quality to those looking to trust. In Jones' terms, they are trustworthy, but not 'richly trustworthy'. On the other hand, institutions and people who are not trustworthy may signal that they are, exhorting people to 'Trust me!' through successful public relations campaigns and other techniques, leading to poorly-placed trust.

As with trust and trustworthiness, it follows that demonstrations of trustworthiness or indicators of competence and reliability are domain and trustor-specific. A scientist, for example, can demonstrate trustworthiness to other scientists through showing observance of conduct and norms of her field and so forth, but may need to demonstrate different qualities to the wider public (John, 2018). The qualities individuals and institutions rely on when making these demonstrations are contextual to the trust relationships with others.

From the perspective of health data governance, the TRP is a central challenge. Whereas there is growing evidence and knowledge about who is trusted and by whom, and what the qualities of trustworthy institutions might be, we have comparatively little knowledge of what this process of communication or signalling can or should look like.[2] As health data systems evolve, however, and in circumstances in which patients, doctors and researchers are all encountering unfamiliar actors throughout the system, this is a critical challenge.

3. The TRP requires an interdisciplinary solution

In theory, trustworthiness may be effectively signalled in a range of ways. As Jones (2012) points out, the first is by "walking the walk," by showing us that they are competent and can be trusted by doing something that anticipates what we would hope to trust them to do. However, behaviours prior to the commitment to a relationship are not an infallible guide to the quality of the relationship in practice. The second, she suggests, is by demonstrating homology, showing that they are relevantly similar to others

[2] See for example the Ipsos Global Trustworthiness Monitor (2023).

https://doi.org/10.1017/dap.2024.74 Published online by Cambridge University Press

who we know can be trusted in this area. At the same time, however, this presents concerns about mimicry and unreliable signalling.

In real-world scenarios where individuals are seeking to make use of a public good or service — such as systems for sharing health data, or enabling its use for the training of AI/ML systems — the TRP can prevent either a well-directed interaction with a trustworthy organisation, lead to misplaced trust in untrustworthy organisations, or indeed lead to a disavowal of all organisations involved in the domain (Figure 1). This highlights that the role of homology in recognising trustworthy actors cuts both ways, and has implications for those deemed similar to those seen as untrustworthy. For example, if the public does not believe a given NHS Trust keeps their health data private, they are more likely to mistrust other trusts, and potentially the NHS generally with data (the care.data scandal in 2014 is one recent example).

Furthermore, a focus on trustworthiness risks becoming simply the (untrustworthy) exhortation to 'trust me'—unless, that is, the mechanism by which trustworthiness ought to be demonstrated is clarified. Without this, principals—in this case members of the public—have no way to detect trustworthy agents, and thus, no matter the virtue of highlighting the distinction between trust and trustworthiness, cannot place 'well-directed' trust.

To expand and deepen this discussion of how to tackle the TRP, we first argue that it can be considered analogous to a range of social dilemmas in which individuals rely in their decision-making processes on the demonstration of unobservable qualities. This is a foundational problem in the evolutionary social sciences: individuals cannot effectively predict the behaviours of others if they cannot directly observe internal states.

Experiments in social behaviour have long played an important role in understanding trust relationships and their relation to the levels of trust reported in questionnaire surveys. Game theory may have particular value, however, in exploring how trustworthiness is signalled and recognised, particularly those that involve scenarios in which individuals must choose partners likely to yield them the maximum payoff. Examples include multi-player Prisoner's Dilemmas (Axelrod and Hamilton, 1981), Trust games (Han et al., 2020), Hawk-Dove games (Smith and Harper, 2003) and the Contract Game (Archetti et al., 2011), among other game theoretical frameworks with differential payoffs for players.

The last of these, the Contract Game, is particularly relevant for the relationships represented in, for example, health data sharing systems, in which some organisations—such as hospitals or GPs—control access to data that is potentially valuable or desirable to others such as academic researchers or insurance companies. In the Contract Game groups are divided into 'principals' and 'agents'. The former controls a finite resource which the latter need or desire. To obtain the resource, agents perform a service for the principals. The payoff structure of the game takes the form of the contract between the two players: the agent's payoff is maximised where it performs a low-cost service for the principal and yet receives a maximum reward; the principal's payoff is maximised where it exchanges the minimum amount of resources for a high-quality service from the agent (Archetti et al., 2011).

The Nash equilibrium—the point at which each player's strategy is optimal given the strategies chosen by all the other players—in the Contract Game is where principals selectively interact with high-quality agents, and create barriers that effectively block agents aiming to provide a low-cost, low-quality service (Rodríguez-Gironés et al., 2015). Failure to select high-quality agents leads to a loss of resources that, over time, leads to a breakdown in the system: principals cannot replace resources without receiving high-quality service from agents.

There are examples of real-world biological scenarios resembling the Contract Game, and the interface between principal/agent theory and biological markets is growing.[3] This body of work explores how rational agents choose between potential partners in economic relationships. Yet in scenarios where individuals can use language to disguise their lack of trustworthiness, the TRP becomes relevant. This is particularly the case in the move from simple to complex societies (Misztal, 2013). In smaller and less complex groups and societies, personal knowledge and group sanctions—such as social ostracism or

[3] For a review, see Hammerstein and Noë, 2016.

https://doi.org/10.1017/dap.2024.74 Published online by Cambridge University Press

physical attack—support direct trust relationships in which there is a continuing value to trustworthiness (Raihani, 2021). Yet as societies become larger and more complex, as with modern-day institutions, such personal knowledge is less common, punishment is not always an effective deterrent and principals must select among a group of potential agents (Dunbar, 2022; West-Eberhard, 1983).

Therefore, as societies grow, it becomes essential to discern high- and low-quality agents in the Contract Game framework. We must rely on signals to determine whom we trust—and insofar as agents may misrepresent the quality of the service they will provide, principals will be disadvantaged by an inability to predict, based on observable agent qualities, the quality of the service they will receive (Raihani, 2021). We suggest that the result of a Contract Game framework where fake signals are common is a low overall level of trust, and in the case of scientific institutions, social and science capital are likely to be damaged.

There is therefore a need for a reliable index of trustworthiness that solves the TRP, but that recognises the domain specificity of trust relationships. Thus, in the case of data, just as principals in the Contract Game seek high-quality agents for a specific service, so do members of the public seek out institutions that are likely to handle their private information with care and only in the ways to which the public agrees. However, while game theoretical scenarios have been central to trust research in general, and repeatedly report the importance of signals in competitive interactions, to date their role has not been considered in detail in the literature on trust in modern institutions, health data science and AI.

In what follows, we draw particularly on research in behavioural ecology, where the notion of signals has helped to solve paradoxes of recognition in the natural world. Evolution has led to Nash equilibria in competitive relationships across species, and an appeal to behavioural ecology reveals a multitude of scenarios where individuals form predictive models of how others will behave—which, for the purposes of this paper, we consider analogous to solving the TRP in modern human institutions. We characterise the contextual nature of such signalling, and show how insights from behavioural ecology can provide an index of trustworthiness, bridging trustworthiness to trust and offering a solution to the TRP.

4. Costly signalling, mimicry and social selection: lessons from behavioural ecology

Interdisciplinary work between biologists and economics has led to a body of literature around partner choice in human populations.[4] For example, Noe and Hammerstein (1995) developed the concept of the ‘biological market,’ which explores how individuals, including humans, choose partners in social dilemmas where it is possible to be exploited. In what follows, we highlight examples from the evolutionary sciences, which we believe illustrate how selection pressures solve principal/agent dilemmas in the natural world. Dynamics over evolutionary time solve TRPs by revealing equilibria in competitive social relationships—and, for the purposes of this paper, highlight how an interdisciplinary approach can be fruitful for offering solutions to seemingly intractable problems in modern institutions. Here, we give a short review of the large literature on signalling theory and explain its relevance to the TRP.

4.1 Signalling theory

Over the last several decades, researchers have shown that costly signals are more likely to reflect true underlying qualities in organisms—they are more likely to be ‘honest’ (Grafen, 1990; Zahavi, 1975). The terms relevant in this body of literature—‘signals,’ ‘costs’ and ‘honesty’—are, however, opaque, and consequently Zahavi’s (1975) insight, known as the handicap principle, was largely ignored until Grafen developed a mathematical model showing its utility in 1990. Grafen’s (1990) model shows that mimics cannot pay the costs of an honest signal; when a costly signal is received in the natural world, therefore, it is indicative of a true underlying quality.

Since this model was developed, there has been an explosion of research in the costly signalling space, comprising what is now known more broadly as ‘signalling theory.’ with important consequences for the

[4] F(for reviews, see Barclay, 2016; Baumard et al., 2013; see also Nesse, 2016).

zoological, biological and human sciences (Connelly et al., 2011; Smith and Harper, 2003). Previous work suggests that signals can determine outcomes in both cooperative and non-cooperative interactions among cells, bacteria and animals (Aktipis, 2020). For example, quorum sensing in bacterial populations can initiate cooperative interactions, particularly among related organisms (Diggle et al., 2007). Anthropologists have also suggested that success in hunting is a signal of high underlying genetic quality, and is consequently associated with reproductive success in hunter-gatherer populations (Stibbard-Hawkes, 2019).

A key principle in this body of literature is that signals evolve because of the effect they have on the receiver (West et al., 2007). This is for two reasons: first, biologists seek to distinguish signals from cues, which are sources of information that have no impact, on average, on evolutionary fitness (Smith and Harper, 2003). The sound a person's feet make when someone walks, for example, does not affect fitness in the vast majority of cases, and therefore did not evolve, but is rather a consequence of the evolution of the structure of human physiology. Second, the signal's importance is defined by how others—the receivers of the information—respond to it (Krebs and Dawkins, 1984; West et al., 2007). If female spiders stop using nuptial gift offerings to make decisions about mating, males will stop offering these gifts.

In general, honesty becomes a critical element of signals when it is possible to fake them—that is, when indices are not possible. This is a common problem across the biological and social worlds. For example, cancer cells effectively mimic normal cells to avoid immune responses (Aktipis et al., 2015). Similarly, some species of bird lay their eggs in the nests of others, who raise these invaders, known as brood parasites, without any genetic relationship to them (Jamie et al., 2020).

In each of these cases, cheaters ('free riders') exploit a signalling system for evolutionary benefit. Mimicry of a signal permits this: if a signal evolves because of the effect it has on a receiver, then falsely conveying that same signal yields benefits without the payment of any costs. The signalling system will break down only if too many free riders exploit it—and consequently free-riding frequency must remain low enough that receivers do not start ignoring signals (Smith and Harper, 2003).

The focus of signalling theory, and relatedly in the science of understanding cooperative systems in a Darwinian world, has been on how and why these systems do not break down with the presence of free riders. Zahavi's insight, followed by Grafen's analytic model, shows the central solution: free riders cannot replicate a costly signal. A bad hunter, to use Stibbard-Hawke's (2019) example from hunter-gatherers, cannot fake being a good one: what matters is what each hunter brings back to the camp.

As human social groups grew beyond the boundaries of small camps, however, determining underlying qualities, such as trustworthiness in economic interactions, became more important; evolutionary theorists have aimed to solve this issue using the social selection framework (Nesse, 2016; West-Eberhard, 1983). Social selection occurs when individuals non-randomly choose partners for cooperative tasks, assuming a market-like structure for relationship networks (Barclay, 2016; Noë and Hammerstein, 1995, 1994). There are many examples across the biological and anthropological literature, but the basic principle is that free riders will not, on average, be selected as cooperative partners, leading to a reduced amount of free riding over evolutionary time. Individuals will refuse to cooperate with free riders, and will instead favour other organisms that display their cooperative tendencies (Aktipis, 2004; Aktipis et al., 2015). This has been shown across the animal and plant kingdoms, with modelled examples ranging from cancers to humans (Bshary and Grutter, 2006; Cronk, 2007; Nowak and Sigmund, 2005).

The critical problem, however, remains assorting cooperators from non-cooperators, especially given that in complex organisms, free riding may be an opportunistic behaviour rather than a default action in social settings (Goodman, 2023). Moreover, while in smaller human societies, local norms, gossip and punishment mechanisms make free riding difficult, as societies grow larger, more opportunities to cheat will arise (Aktipis et al., 2015; Dunbar, 2009). People can effectively cheat anonymously in large, stratified societies in numerous ways—but the fundamental element of this cheating is anonymity. Hiding the proclivity for stealing, tax evasion, or any other free-riding behaviour creates a serious problem for social selection frameworks that rely on being able to predict an individual's likelihood of cooperating in the future.

https://doi.org/10.1017/dap.2024.74 Published online by Cambridge University Press

Yet, as with smaller groups, individuals attempting to make decisions about cooperation must rely on signals (Gambetta and Hamill, 2005). These may be signals of a shared social identity—helping others we believe are from a similar background to us, for example—or signals indicating a similar worldview, such as shared religious beliefs (Cohen and Haun, 2013; McElreath et al., 2003; Sosis and Alcorta, 2003). Anthropological data indicate both forms of social selection are common among hunter-gatherers, and moreover that some signals, especially those linked with religion, can become increasingly costly over time. Scarification, for example, indicates a commitment to religion that is difficult to fake, and two people with shared markings may be likely to help each other or to unite against common enemies (Bell and Paegle, 2021).

These signals are in effect representations of underlying qualities that drive behaviours from others. They indicate to acculturated receivers information about the signaller's identity that is likely to predict the person's future behaviour. This allows for an assortment of fellow group-members—understood in this context as individuals with shared cultural heritage, norms, or values—for cooperative purposes. Social selection therefore remains an effective evolutionary mechanism even in large societies because honest signalling acts as a barrier to free riding (Goodman and Ewald, 2021).

4.2 Paying costs: a solution to the TRP

The signalling theory framework applies not only to dyadic relationships between individuals but between groups, within groups and between individuals and the institutions of which their society is comprised. In each of these cases, trustworthiness is the underlying quality that individuals or institutions aim to signal. Our societies are now so large and stratified that trust is an essential element of our social selection networks: we must choose our partners, whether people, organisations or political parties, based on information we receive about them.

As trustworthiness is not readily observable, individuals must develop heuristics for deciding in whom to place trust, even (or perhaps especially) in the setting of modern technology (Kuipers, 2022) and institutions (Lounsbury et al., 2023). These will vary by context; the notion of costs, consequently, must also vary, as must what constitutes an honest signal.[5] This explains the use of rating systems across the commercial sphere, especially in the digital age: individual consumers rate their experiences to help guide the decisions of others (and of course, to punish businesses for bad experiences). Few people take seriously the claims a business makes about its own virtues without good reviews to justify them.

Furthermore, we have developed mechanisms for ranking relevant qualities of businesses both for signalling and sanctioning purposes. For example, restaurants in the United Kingdom must pass inspections and receive a Food Hygiene Rating from the Food Standards Agency (FSA); the rating is on an ordinal scale from 0 (urgent improvement required) to 5 (hygiene standards are very good) ("Food Hygiene Rating Scheme"). The costs of the signal—the rating, which is displayed online and often on shop windows—are investing in the hygiene practices that result in the score received. It is consequently difficult to fake the rating, though the costs are specific to both the industry and the regulatory framework in which the business operates (Worsfold and Worsfold, 2007).

Trust, in this case, is (or is not) placed in the institution—the FSA—that created and maintains the rating framework, and perhaps to a lesser degree in individual restaurants who have an interest in misrepresenting their standards. Trustworthiness is the underlying quality to which trust, at least we hope, attaches—but it is the context-specific costs that guide that connection. Our institutions aim to demonstrate their trustworthiness by paying these costs—solving the TRP in a domain-specific manner.

Determining whom to trust therefore relies on discerning signals indicative of trustworthiness from those designed to free ride on trust. Appealing to behavioural ecology and anthropology reveals how selection has aided in solving the TRP—or biological equivalents of the TRP—over time in a variety of

[5] Costs are therefore not necessarily economic, and may instead entail, for example, openness and transparency around institutional policies.

https://doi.org/10.1017/dap.2024.74 Published online by Cambridge University Press

populations. Yet to solve the TRP in the era of data-driven healthcare, we must appeal to the peculiarities of a given domain, such as the sharing of health data or the implementation of AI tools.

5. Modern institutions: health data and AI

In the context of health data sharing, members of the public have indicated a lack of trust in institutions regarding how their data will be used. Many individuals do not trust NHS institutions to safely store their data or, in some cases, not to sell data for profit (Lounsbury et al., 2021). It is argued that transparency around these issues, for example, by explaining the potential drawbacks of health data sharing as well as declaring any industry deals around health data, may improve public trust. However, as O'Neill (2018) pointed out, transparency that is passive, and only a matter of making material available is unlikely to result in trust. In the terms introduced here, organisations need to bear the costs of time (and money) involved in developing active forms of communication that take account of the specific capacities and concerns of actual audiences in a way that would allow those audiences to hold organisations to account. In doing so, they demonstrate an openness to bearing the reputational and potential financial costs of misbehaviour.

For example, Kundu (2023) noted that while hundreds of AI-powered tools exist in the healthcare space, healthcare is behind other industries in healthcare adoption. This is partially due to a lack of trust in AI tools: many physicians, as well as the public, do not trust the outputs from AI models that directly affect treatment-decision-making and diagnosis (Asan et al., 2020). Kundu (2023) suggested that AI should be regarded as trustworthy in the healthcare space only insofar as it provides the same explanations for its decisions as does a treating physician. This follows evidence that explainability directly affects levels of trust by physicians in model outputs: black box tools are much less likely to be trusted than tools that showed the steps taken towards a clinically relevant decision. Furthermore, Kundu (2023) argued, there must be transparency around training data and community vetting for any AI tool used. Asan et al. (2020) similarly emphasised the role of transparency about the working principles of AI algorithms and the algorithmic biases and biases due to underrepresentation in fostering trust. Given evidence that training data can exacerbate health inequalities by ignoring clinically relevant qualities of minority populations, no tool can be considered trustworthy unless the biases inherent in its training data are made explicit.

However, making this information available, and providing step-by-step explanations for decisions taken, can incur costs, associated with the work required to establish transparency, the consequences of scrutiny and the effort required to move from passive to active transparency and signalling that takes account of the specific capacities of audiences. Any designer of AI systems that pays these costs will, on our view, be more discernibly and richly trustworthy than those who leave their training data and decision-making pathways opaque or commits to minimum standards of transparency. This is, however, a domain-specific solution to the TRP, and different, though analogous solutions are necessary in other settings.

In both of these cases, costs are specific to the institutions, their goals, and the ways they interact with the public. Solving the TRP therefore relies on domain-specific evaluation of costs by considering public concerns and engagement at a local level to determine how trustworthiness can best be signalled, that is, actively communicated rather than passively observed. We suggest the cost-paying model, derived from behavioural ecology, may be the most effective way of solving the TRP.

6. Challenges and next steps

In this paper, we have explored the challenge associated with connecting trustworthiness and trust, suggesting that the TRP can valuably be approached as a problem of creating rich trustworthiness and effective signalling. We suggest, overall, that the domain specificity inherent to the signalling theory paradigm offers a potentially effective interdisciplinary solution to understanding and addressing the TRP, which could be foundational to whether and how rapidly improving technologies are integrated into the healthcare space. Further work to understand and qualifies the domain-specific costs of trustworthiness in these areas is, however, urgently needed.

https://doi.org/10.1017/dap.2024.74 Published online by Cambridge University Press

While an interdisciplinary approach that combines principal/agent theory through the Contract Game and signalling in behavioural ecology offers a theoretical solution to the TRP, we note that this argument is merely a theoretical step forward. Future work must explore whether and how taking a domain-specific approach to costly signalling can bridge trustworthiness to trust in modern institutions, particularly in the rapidly evolving area of digital health. We believe, however, that this appeal to other disciplines is an essential first step, and one that will, in the future, give organisations the power to signal their trustworthiness to the public effectively.

We note, however, two potential objections to the cost-paying framework we have introduced here. First, costs may create selective incentives to focus on some features rather than others. For example, in the food case, signalling hygiene does not tell anything about the quality of food, the sustainability of the supply chain, or the personal qualities of the chef. Similar arguments are made by O'Neill (2018) about how transparency skews activity: for example, demonstrating you are not doing anything wrong does not mean you are doing good. In the case of the FSA, the commitment to making all papers and documents publicly available may have led to changes in institutional behaviour, such as not writing things down or having more informal, hidden discussions (Worsfold and Worsfold, 2007). Costly signals can therefore give us information only about specific qualities of an individual or institution, but may not tell us how either would act in the absence of the need to display costs—there is an interest in maintaining the relevant form of capital.

Second, and relatedly, it may not always be clear to individuals themselves, and certainly not to institutions comprised of individuals, whether they are trustworthy. Self-deception and motivational pluralism—having several goals simultaneously—are possible and perhaps inevitable. For example, an engineer may design an AI model to make healthcare recommendations to help relieve burden on the NHS, but for reasons of career advancement may also wish to mask some of the model's shortcomings, such as poor-quality training data. A health data organisation may want to make data as widely available as possible for research but may have competing pressures to protect national economic or security interests by limiting access. Here, however, we suggest that individuals and institutions may overcome this internal confusion by appealing to explanations, as in Kundu's (2023) model of trustworthiness in AI. In malpractice cases, clinicians must explain the process by which they came to clinical decisions; we suggest, following Kundu (2023), that in these contexts, the governance systems for health data and development processes for algorithms must bear the potential time, resource and reputational costs to do the same. Active transparency is a cost that individuals and institutions must pay in specific, not just general, cases.

More broadly, we suggest that the costly signalling framework provides institutions with a toolkit for determining, in a domain-specific fashion, how best to demonstrate their trustworthiness to the public. Commitment to transparency may be a common mechanism for this demonstration, but we suggest active interactions with members of the public to determine their views on what qualifies a healthcare institution as trustworthy, and how institutions can best use their data in a trustworthy manner. Interviews and focus groups will help to achieve this, and in conjunction with effective educational strategies, will help members of the public to determine who is trustworthy effectively. Previous work with taxi drivers by Gambetta and Hamill (2005) showed a specific case where this has been achieved; we suggest an analogous approach in the healthcare and AI settings.

Data availability statement. None.

Author contribution. Conceptualisation and writing: J.G.; R.M. Both authors approved the final submitted draft.

Provenance. This article was accepted into the 2024 Data for Policy Conference and published in Data & Policy on the strength of the Conference review process.

Funding statement. RM's research was supported by grants from the Kavli Foundation to the University of Cambridge (G115418) and from Wellcome to Wellcome Connecting Science ((grant no. 220540/Z/20/A). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript." Or "None" when no specific funding exists.

Competing interest. None.

https://doi.org/10.1017/dap.2024.74 Published online by Cambridge University Press

References

Aktipis CA (2020) *The cheating cell: How Evolution Helps Us Understand and Treat Cancer*. Princeton, NJ: Princeton University Press.

Aktipis CA (2004) Know when to walk away: contingent movement and the evolution of cooperation. *Journal of Theoretical Biology 231*, 249–260. https://doi.org/10.1016/j.jtbi.2004.06.020

Aktipis CA, **Boddy AM**, **Jansen G**, **Hibner U**, **Hochberg ME**, **Maley CC**, **Wilkinson GS** and (2015) Cancer across the tree of life: cooperation and cheating in multicellularity. *Philosophical Transactions of the Royal Society B: Biological Sciences 370*, 20140219. https://doi.org/10.1098/rstb.2014.0219

Archetti M, **Scheuring I**, **Hoffman M**, **Frederickson ME**, **Pierce NE and Yu DW** (2011) Economic game theory for mutualism and cooperation. *Ecology Letters 14*, 1300–1312. https://doi.org/10.1111/j.1461-0248.2011.01697.x

Armstrong S (2023) Palantir gets £480m contract to run NHS data platform. *BMJ 383*, p2752. https://doi.org/10.1136/bmj.p2752

Asan O, **Bayrak AE and Choudhury A** (2020) Artificial intelligence and human trust in healthcare: focus on clinicians. *Journal of Medical Internet Research 22*, e15154. https://doi.org/10.2196/15154

Axelrod R and Hamilton WD (1981) The evolution of cooperation. *Science 211*, 1390–1396. https://doi.org/10.1126/science.7466396

Baier A (1986) Trust and antitrust. *Ethics 96*, 231–260.

Barclay P (2016) Biological markets and the effects of partner choice on cooperation and friendship. *Current Opinion in Psychology, Evolutionary Psychology 7*, 33–38. https://doi.org/10.1016/j.copsyc.2015.07.012

Baumard N, **André J-B and Sperber D** (2013) A mutualistic approach to morality: the evolution of fairness by partner choice. *Behavioral and Brain Sciences 36*, 59–78. https://doi.org/10.1017/S0140525X11002202

Bell AV and Paegle A (2021) Ethnic markers and how to find them. *Human Nature 32*, 470–481. https://doi.org/10.1007/s12110-021-09401-z

Braun M, **Bleher H and Hummel P**(2021) A leap of faith: is there a formula for "Trustworthy" AI?. *Hastings Center Report 51*, 17–22. https://doi.org/10.1002/hast.1207

Bshary R and Grutter AS (2006) Image scoring and cooperation in a cleaner fish mutualism. *Nature 441*, 975–978. https://doi.org/10.1038/nature04755

Carter P, **Laurie GT and Dixon-Woods M** (2015) The social licence for research: why care.data ran into trouble. *Journal of Medical Ethics 41*, 404–409. https://doi.org/10.1136/medethics-2014-102374

Cohen E and Haun D (2013) The development of tag-based cooperation via a socially acquired trait. *Evolution and Human Behavior 34*, 230–235. https://doi.org/10.1016/j.evolhumbehav.2013.02.001

Connelly BL, **Certo ST**, **Ireland RD and Reutzel CR** (2011) Signaling theory: a review and assessment. *Journal of Management 37*, 39–67. https://doi.org/10.1177/0149206310388419

Cronk L (2007) The influence of cultural framing on play in the trust game: a maasai example. *Evolution and Human Behavior 28*, 352–358.

Diggle SP, **Griffin AS**, **Campbell GS and West SA** (2007) Cooperation and conflict in quorum-sensing bacterial populations. *Nature 450*, 411–414. https://doi.org/10.1038/nature06279

Dunbar RIM (2022) Managing the stresses of group-living in the transition to village life. *Evolutionary Human Sciences 4*, e40. https://doi.org/10.1017/ehs.2022.39

Dunbar RIM (2009) The social brain hypothesis and its implications for social evolution. *Annals of Human Biology 36*, 562–572. https://doi.org/10.1080/03014460902960289

Food Hygiene Rating Scheme [WWW Document] (n.d.) Food Standards Agency. Available at https://www.food.gov.uk/safety-hygiene/food-hygiene-rating-scheme (accessed 11 9 23).

Gambetta D and Hamill H (2005) *Streetwise: How Taxi Drivers Establish Their Customers' Trustworthiness*. New York, NY, USA: Russell Sage Foundation.

Giddens A (1991) *The Consequences of Modernity*. Newark: Polity Press.

Gille F, **Jobin A and Ienca M** (2020) What we talk about when we talk about trust: theory of trust for AI in healthcare. *Intelligence-Based Medicine 1–2*, 100001. https://doi.org/10.1016/j.ibmed.2020.100001

Gille F, **Smith S and Mays N** (2015) Why public trust in health care systems matters and deserves greater research attention. *Journal of Health Services Research & Policy 20*, 62–64. https://doi.org/10.1177/1355819614543161

Goodman JR (2023) The problem of opportunity. *Biology & Philosophy 38*, 48. https://doi.org/10.1007/s10539-023-09936-8

Goodman JR and Ewald PW (2021) The evolution of barriers to exploitation: sometimes the Red Queen can take a break. *Evolutionary Applications 14*, 2179–2188. https://doi.org/10.1111/eva.13280

Grafen A (1990) Biological signals as handicaps. *Journal of Theoretical Biology 144*, 517–546. https://doi.org/10.1016/S0022-5193(05)80088-8

Graham M, **Milne R**, **Fitzsimmons P and Sheehan M** (2022) Trust and the goldacre review: why trusted research environments are not about trust. *Journal of Medical Ethics.* https://doi.org/10.1136/jme-2022-108435

Hammerstein P and Noë R, (2016) Biological trade and markets. *Philosophical Transactions of the Royal Society B: Biological Sciences 371*, 20150101. https://doi.org/10.1098/rstb.2015.0101

https://doi.org/10.1017/dap.2024.74 Published online by Cambridge University Press

Han TA, **Perret C and Powers ST** (2020) When to (or not to) trust intelligent machines: Insights from an evolutionary game theory analysis of trust in repeated games. *Cognitive Systems Research 68*, August 2021, 111–124.

Hardin R (2002) *Trust and Trustworthiness*. New York, NY, USA: Russell Sage Foundation.

Hawley K (2019) *How To Be Trustworthy*. Oxford, New York: Oxford University Press.

Ipsos Global Trustworthiness Monitor (2023) Available at https://www.ipsos.com/sites/default/files/ct/publication/documents/2023-01/ipsos-global-trustworthiness-monitor-stability-in-an-unstable-world.pdf

Jamie GA, **Belleghem SM**, **Hogan BG**, **Hamama S**, **Moya C**, **Troscianko J**, **Stoddard MC**, **Kilner RM and Spottiswoode CN** (2020) Multimodal mimicry of hosts in a radiation of parasitic finches. *Evolution; International Journal of Organic Evolution 74*, 2526–2538. https://doi.org/10.1111/evo.14057

John S (2018) Epistemic trust and the ethics of science communication: against transparency, openness, sincerity and honesty. *Social Epistemology 32*, 75–87. https://doi.org/10.1080/02691728.2017.1410864

Jones K (2012) Trustworthiness. *Ethics 123*, 61–85. https://doi.org/10.1086/667838

Jones K (1996) Trust as an affective attitude. *Ethics 107*, 4–25. https://doi.org/10.1086/233694

Kalkman S, **van Delden J.**, **Banerjee A**, **Tyl B**, **Mostert M and Thiel G. van** (2022) Patients' and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence. *Journal of Medical Ethics 48*, 3–13. https://doi.org/10.1136/medethics-2019-105651

Krebs J and Dawkins R (1984) Animal signals: mind-reading and manipulation. In Davies NB, Krebs JR (eds.), *Behavioural Ecology*: *An* Evolutionary Approach. Hoboken, NJ, USA: Blackwell.

Kuipers B (2022) Trust and cooperation. *Front. Robot. AI. Sec. Ethics in Robotics and Artificial Intelligence 9*. https://doi.org/10.3389/frobt.2022.676767

Kundu S (2023) Measuring trustworthiness is crucial for medical AI tools. *Nature Human Behaviour 7*, 1812–1813. https://doi.org/10.1038/s41562-023-01711-9

Lounsbury M (2023) the problem of institutional trust. *Organization Studies 44*, 308–310. https://doi.org/10.1177/01708406221131415

Lounsbury O, **Roberts L**, **Goodman JR**, **Batey P**, **Naar L**, **Flott KM**, **Lawrence-Jones A**, **Ghafur S**, **Darzi A and Neves AL** (2021) Opening a "Can of Worms" to explore the public's hopes and fears about health care data sharing: qualitative study. *Journal of Medical Internet Research 23*, e22744. https://doi.org/10.2196/22744

McElreath R, **Boyd R**, **Richerson PJ** (2003) Shared norms and the evolution of ethnic markers. *Current Anthropology 44*, 122–130. https://doi.org/10.1086/345689

Milne R, **Morley KI**, **Howard H**, **Niemiec E**, **Nicol D**, **Critchley C**, **Prainsack B**, **Vears D**, **Smith J**, **Steed C**, **Bevan P**, **Atutornu J**, **Farley L**, **Goodhand P**, **Thorogood A**, **Kleiderman E**, **Middleton A** and **on behalf of the Participant Values Work Stream of the Global Alliance for Genomics and Health**, (2019) Trust in genomic data sharing among members of the general public in the UK, USA, Canada and Australia. *Human Genetics 138*, 1237–1246. https://doi.org/10.1007/s00439-019-02062-0

Misztal B (2013) *Trust in Modern Societies: The Search for the Bases of Social Order*. Cambridge, UK John Wiley & Sons.

Möllering G (2001) The nature of trust: from georg simmel to a theory of expectation, interpretation and suspension. *Sociology 35*, 403–420. https://doi.org/10.1177/S0038038501000190

Nesse RM (2016) Social selection is a powerful explanation for prosociality. *Behavioral and Brain Sciences 39*, e47. https://doi.org/10.1017/S0140525X15000308

Noë R and Hammerstein P (1995) Biological markets. *Trends in Ecology & Evolution 10*, 336–339. https://doi.org/10.1016/S0169-5347(00)89123-5

Noë R and Hammerstein P (1994) Biological markets: supply and demand determine the effect of partner choice in cooperation, mutualism and mating. *Behavioral Ecology and Sociobiology 35*, 1–11. https://doi.org/10.1007/BF00167053

Nowak MA and Sigmund K (2005) Evolution of indirect reciprocity. *Nature 437*, 1291–1298. https://doi.org/10.1038/nature04131

O'Neill O (2018) Linking trust to trustworthiness. *International Journal of Philosophical Studies 26*, 293–300. https://doi.org/10.1080/09672559.2018.1454637

Putnam R, **Leonardi R and Nanetti R**(1993) *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton: Princeton University Press.

Raihani N (2021) *The Social Instinct: How Cooperation Shaped the World*. London, UK: Jonathan Cape.

Rodríguez-Gironés MA, **Sun S and Santamaría L** (2015) Passive partner choice through exploitation barriers. *Evolutionary Ecology 29*, 323–340. https://doi.org/10.1007/s10682-014-9738-3

Smith, JM and Harper D (2003) *Animal Signals, Oxford Series in Ecology and Evolution*. Oxford, New York: Oxford University Press.

Sosis R and Alcorta C (2003) Signaling, solidarity, and the sacred: the evolution of religious behavior. *Evolutionary Anthropology 12*, 264–274. https://doi.org/10.1002/evan.10120

Stibbard-Hawkes DNE (2019) Costly signaling and the handicap principle in hunter-gatherer research: a critical review. *Evolutionary Anthropology*. https://doi.org/10.1002/evan.21767

Wellcome Global Monitor (2018) *Wellcome Global Monitor: How Does the World Feel about Science and Health*. London, UK: Wellcome Global Monitor.

https://doi.org/10.1017/dap.2024.74 Published online by Cambridge University Press

West SA, **Griffin AS and Gardner A** (2007) Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology 20*, 415–432. https://doi.org/10.1111/j.1420-9101.2006.01258.x

West-Eberhard MJ (1983) sexual selection, social competition, and speciation. *The Quarterly Review of Biology 58*, 155–183. https://doi.org/10.1086/413215

Worsfold D and Worsfold PM (2007) Evaluating food hygiene inspection schemes: 'Scores on Doors' in the UK. *International Journal of Consumer Studies*, *31*: 582–588. https://doi.org/10.1111/j.1470-6431.2007.00612.x

Zahavi A (1975) Mate selection—A selection for a handicap. *Journal of Theoretical Biology 53*, 205–214. https://doi.org/10.1016/0022-5193(75)90111-3

Cite this article: Goodman JR and Milne R (2024). Signalling and rich trustworthiness in data-driven healthcare: an interdisciplinary approach. *Data & Policy*, 6: e62. doi:10.1017/dap.2024.74

https://doi.org/10.1017/dap.2024.74 Published online by Cambridge University Press