Trust and Trustability (v0.5)
An idealized operational theory of economic attentiveness

Mark Burgess

May 26, 2023

Abstract

This work presents a summary of an operational model for trust and trustability (trustworthiness), building on Promise Theory, as a memory mechanism for time managing cooperation against future returns. Unlike so called rational actor models of Game Theory, we don't look a trust as an optimization game, rather we look at the operational roles of semantics and dynamics played by mutual assessment. The model here, proposed and partially developed more extensively in previously available notes, views trust as a resource issue— effectively a policy for inattentiveness in scenarios that balance opportunities and risks. The economics of trust are not altruistic, nor are they selfish maximizations of benefit. Trust is rather a cost saving policy, whose complement 'antitrust' or 'risk taking' allows 'selfish agents' to allocate resources for self interest against countervailing costs. In this work, we consider the basic two-agent interaction to be generalized to more realistic scenarios elsewhere.

Contents

1 Introduction **1**
- 1.1 Physics of agents . . . 2
- 1.2 Promise Theory as an agent model . . . 3
- 1.3 Assumptions about promises and trust . . . 3

2 Agent work and output **5**
- 2.1 Productivity of supply, affinity for demand . . . 6
- 2.2 Dimensions of key quantities . . . 6
- 2.3 Definition of trust . . . 7
- 2.4 Mistrust and feedback in promise keeping . . . 8
- 2.5 Promise keeping rate or process velocity . . . 9
- 2.6 Trusting and attraction to trustworthiness . . . 10
- 2.7 From energy to velocity, from trust to sampling rate . . . 12

3 Scaling of assessment and trust in simplicity **13**
- 3.1 Examples . . . 13

4 Other cooperative potentials **16**
- 4.1 Vulnerability and threats . . . 16
- 4.2 Agent confidence and doubt . . . 16
- 4.3 Hope and blame . . . 17
- 4.4 Risk and risk appetite and recovery races . . . 17

5 Automating operational trust policy **19**

6 Summary **20**

1 Introduction

Trust is one of the most commonplace ideas in our lives, at the root of almost everything we do, and yet it remains an elusive idea that we handwave about and even dismiss as part of the moral mystique

of the human condition. In the humanities, trust is held in a category of social capital, with other characterizations like confidence, vulnerability, and hopefulness, curiosity, and interest; researchers focus mainly on human expressions to explain its role in society. There is some neuroscientific correlation with motivational and intentional processes [1]. In computer science and technology, by contrast, trust is trivialized as an identity issue, and is sidestepped using passports and signatures as an alibi for verification.

Over the past two decades, the stakes for understanding trust have been raised as we attempt to integrate cybernetic processes into our lives to form a more integrated human-machine society. This work offers a reasonably concise summary of a rather different model of trust, based on information and intent. It stems partly from the need to automate responses for interactions where machinery operates as proxy for human intent, and is based on an ongoing project to define the role of intent and promises in more formal terms, but it is also meant as a serious contribution to divine the meaning of trust from a sociological and indeed neurological perspective.

This work builds on Promise Theory [2], which incorporates a set of useful principles to build on. Promise Theory holds that agents are initially autonomous, and influence one another by expressing their intentions to cooperate by assessing one another's semantics and dynamics, all of which build on a presumption of trust [3]. It has previously been used to describe money and economics as proxies for trust dynamics [4]. Promises are closely associated with trust, but on a deeper level they imply a *process oriented* view of trust and intent, operationalized through promise keeping.

A simple operational (process-oriented) view of trust can this be offered here to explain the properties and features described in the multifarious literatures in as concise a form as possible. A summary of literature has been given elsewhere for the sake of brevity [5]. The model leads to some simple predictions that can be tested by willing empiricists. Processes are phenomena that couple change in space and time together in a prescribed manner. They have specificity and direction (which maps to intentionality), magnitude, and rate. Trust relates not to the acts of keeping arbitrary promises, but rather to the meta process of oversight (as an overhead) for administering and verifying the veracity of their outcomes.

Apart from striving for clarity in the semantics of trust, the larger goal of this effort is to be able to use trust as a predictive and operational tool, both for technology and social science. Thus we look for clear definitions and relationships build on. It is roughly analogous to the way we use the concept of energy is physics, combining the dimensions of measurement with the driving forces for change.

1.1 Physics of agents

How might one trust such an idealized operational model of trust (or indeed any other model in the physical sciences) faced with something as complex as the human condition? Should one trust its originator, its promises, or its outcomes? These are key distinctions. We tend to associate trust with persons, but in fact that is by inference of their observed behaviours.

By associating too quickly with humans, we tend to invoke moral notions that confound simplicity and compound the ambiguities of defining the phenomena. Our goal here then is to eschew complexity as a starting point, to see if this allows us to understand how and why such complexity enters later—to avoid confronting all matters at once; indeed, the broad strokes of behaviour may not be as complex as we imagine. Physics has a long history of this kind of approach, so we can try to learn from its successes. It deals with how phenomena appear across different scales, from one size or measure to another as this usually provides the most orderly separation of concerns. Phenomena on a large scale are sewn together from contributions on a smaller scale, with additional boundary information at each separable level. For agent models, Promise Theory provides a scaffolding consistent with this approach [2]. That is not to imply that complex phenomena can be expected to obey simple and deterministic algebraic laws as elementary systems do. Trust will place us in the realm of *assessment* (the so-called 'measurement problem' in physics), which is an information problem to be made consistent with Information Theory [6]. These are some of the challenges.

Agent modelling involves the encoding of *intent*, or the interior 'algorithms' of 'policy driven behaviours'. In contrast to the elementary phenomena usually studied by physics, complex agents will tend to behave non-deterministically. Agents will bond together based on their promiseable behaviours, in order to share finite resources and specializations and achieve outcomes on a larger collaborative scale. This is true of atoms and it is true of humans. Trust will enable agents with sufficient memory to borrow against future resolutions, as a form of gambling or long term accounting. Reliability and thus trustworthiness are agents' assessments of performance. While this idea is simple enough to express in words, we need a more formal and quantitative explanation relating to investment of effort by agents with their finite resources. Only experience in applying a model to many cases will build up the necessary learning

to assess the model itself. We should expect setbacks as we confront the realities of approximation and stochastic development.

1.2 Promise Theory as an agent model

Promises are statements of intent, that agents attempt to keep by best effort. They may be used as targets to achieve or as lines in the sand, ('positions' or 'stances') that an agent adopts to define a problem in terms of 'directions' represented in the subject or 'body' b of a promise. Promises are not commands or deterministic rules whose outcomes are assumed, they are statements of intent to be fulfilled by best effort. In earlier work, it was shown how trust is related to the notion of promise keeping and that Promise Theory offers a convenient framework in which to formulate an agent based model of its semantics, its dynamics, and its scaling [3, 5, 7]. Trust also has a number of related and derived interpretations to consider, with variations on semantics, which are adapted to different purposes. We speak of risk, confidence, hope, etc.

1.3 Assumptions about promises and trust

Consider to begin with the simplest scenario in which trust makes sense, we focus first on a single relationship and its observation by a neutral third party. This is the 'three body problem' for intentionality (see figure 1).

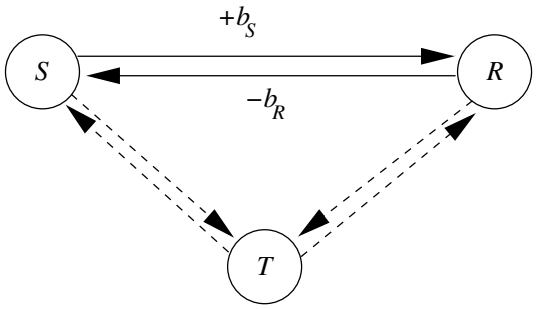

Figure 1: The prototypical three body building block for trust..

As a helpful mnemonic, we choose a scenario where agents assume the roles of 'sender' S, 'receiver' R, and 'third party' T. S will be the agent to offer a typical promise, R will be the 'receiver' or one to choose to accept the promise, while T will be a neutral third party. Cooperation is rarely as simple as two agents interacting, however we start by looking at this atomic scenario (a half duplex channel of service from one agent to another) as the basic building block of cooperation and trust dynamics. We write the promise by S of b_S, and its independent acceptance by R in the amount b_R as two promises:

$$\pi_S^{(+)} \quad : \quad S \xrightarrow{+b_S} R \tag{1}$$

$$\pi_R^{(-)} \quad : \quad R \xrightarrow{-b_R} S, \tag{2}$$

where the b's represent the promise bodies, and the signs (+) and (-) represent offer and acceptance directionality. The bosies are assumed to include both a magnitude and a schedule for what is offered or accepted respectively. The result of these two promises is to intend a shared realization in the promised amount $b_R \cap b_S$ between them.

Promise signalling and measurement must satisfy the laws of communication as described by Shannon in what is now known as Information Theory [6]. Our basic premise for trust dynamics is that trust dynamics are composed of two parts (called trust and trustworthiness or kinetic and potential trust), whose separation is a consequence of the autonomy of agents, whence the underlying (+) and (-) polarity of promises:

- The assessment, formed by one agent about another agent's reliability with respect to keeping its promises, is what leads to a concept of *trustworthiness* (or potential trust).

- Once assessed, trustworthiness may be used to inform a policy (strategy), either partially of completely at the behest of the agent concerned. This amounts to an allocation of work for verifying its dealings with the other. An agent thus invests a level of *trust* (actually the complementary level of mistrust is the relevant allocation) as a strategy for checking and verifying the other's promise keeping reliabilty. The allocation of mistrust also acts as a signal that feeds back to the origin agent and influences the efficiency of the relationship over a timescale of multiple interactions.

Although independent, it would be perverse if the levels of trust and trustworthiness were not correlated positively by some monotonic function, but this does not imply a simple proportionality between them as there are many factors involved in allocating resources, particularly under strain or competition [8]. Note that assessments play a central role in both cases. The ability for an agent to form an assessment depends entirely in its capabilities. In the case of human agents, one expects Kahneman's system 1 and system 2 modes of assessment to play a role in cost saving such assessments [9]. The former is the coarse and simple low cost estimator and will probably therefore be the dominant one in forming trust related assessments. For mechanistic agents the assessments may be as simple as true versus false, like finite state machines.

Let's note some key assumptions of Promise Theory:

1. Autonomy (causal independence): agents can only make promises about their own intentions and behaviours. Attempting to promise on behalf of another is generally without merit.

2. Economy: agents have finite a budget of internal resources.

3. Observability: an agent's knowledge of other agents (its exterior world) is based entirely on its own promised intent to receive information and form assessments, and is characterized by a sampling process which tests the information at some interval $\Delta\tau$ to be decided. This sampling process is subject to the Shannon Nyquist sampling law, so an agent can only accurately detect changes that occur slower than half its sampling rate.

4. Parsimony: all agent characteristics are modelled as promises (in the promise theoretic definition of [2]), autonomously made, that may or may not be known to others. In general only that which is promised is observable by a matching receptor promise.

5. Relativity: only a third party T can observe and calibrate an interaction between S and R, against a single shared scale of assessment, if both S and R promise to report their part in the interaction T:

$$\pi_{ST} \quad : \quad S \xrightarrow{+b_{ST}} T \tag{3}$$

$$\pi_{ST} \quad : \quad T \xrightarrow{-b_{TS}} S \tag{4}$$

$$\pi_{RT} \quad : \quad R \xrightarrow{+b_{RT}} T \tag{5}$$

$$\pi_{RT} \quad : \quad T \xrightarrow{-b_{TR}} R. \tag{6}$$

   The third party T can discern only what it is capable of by virtue of its own promised capabilities, subject to the Nyquist-Shannon sampling law, and assuming that neither S nor R are lying in their promises to T about to one another.

Some corollaries of these points may be underlined:

- By assumption 1, in order for an agent to benefit from the offer of a promise by another agent it must autonomously promise to accept the offer. This cannot be induced from outside. Thus the offer in (1) must be accepted by a directed intention to accept and use the offer in (2), in order for an effective promise with body $b_R \cap b_S$ to be realized.

- Agents assess phenomena by sampling the information they can observe, using whatever receptors and internal capabilities they may have. Each agent then 'measures' or calibrates its assessment of something by projecting the outcome of its own sampling process into a set of internal states. Elementary agents make elementary assessments, since their internal resources are limited. Sophisticated agents, with greater resources, can form more sophisticated assessments. With finer grained states, assessments are less easily distinguishable than elementary observations and in the limit of infinite size become a continuum.

- Assessments are written like functions $\alpha()$. We distinguish two main kinds of assessment (written with and without the caret symbol $\hat{\cdot}$) in an attempt for clarity:
  - A careted assessment maps a process of observation to some dimensionless scale e.g. $[0, 1]$. e.g. the assessment of a promise $\hat{\alpha}(\pi)$ might simply be 'kept' or 'not kept', or it could be a number without dimensions.
  - An uncareted assessment of a variable quantity $\alpha(b)$ maps the argument to a value in the same dimensions as b, as a representative value perceived by the agent.

  The difference between b and $\alpha(b)$ is that b is a policy or intention, while $\alpha(b)$ is a sampled or observed event. An agent A might know both these, but its observations are filtered through its capabilities, represented as the function α_A.

An agent that makes a sequence of assessments of some variable can collect these into an ordered set, or directional set, which forms an episode (observed from a point) or trajectory (sourced over a spacetime path).

2 Agent work and output

Framed by the alignment of promises both in configuration space and intention space (see figure 2), we can characterize the work done by agents entirely in terms of the efforts to keep promises. The activity promised between agents involves multiple channels between sender and receiver, which these fall into two main categories.

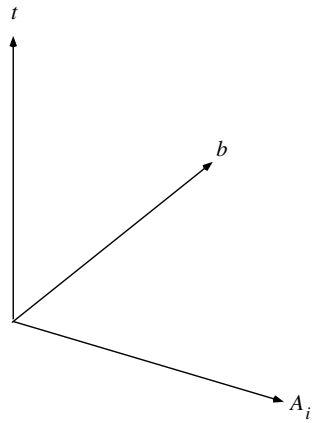

Figure 2: The three kinds of degrees of freedom for an agent model are shared time t, agent identity (location or exterior space A_i) and promise state b (interior space). Together these form a model embedding space known as Semantic Spacetime.

- Channels for transporting promised results (goods or services, etc), characterized by the body of the promise b.

  The generalized economic concepts of supply (promised productivity) and demand (promised acceptance), are associated with this first kind.

- Channels for monitoring and overseeing the flow of results, to detect surpluses and shortfalls with respect to the promised amounts.

  The concepts of trust and trustworthiness belong to this second channel, which may be considered a special subset of the former. They provide for the attentiveness, observability, and 'transparency of process' granted to agents to monitor the flow of the former through repeated sampling. The absence of this such observability may turn deliveries into opaque 'impositions', which tend to reduce assessments of trust.

Based on their assessments of these exchanges of information, agents may alter their promises, i.e. the type or level of their delivery through these channels in order to adapt their intent (see table 1).

For future reference, let's briefly define some terminology in terms of promise theoretic variables.

Supply Monitoring (+)	Demand monitoring (-)
Attentiveness to monitoring receipt	Attentiveness to monitoring delivery
Sampling rate for $\hat{\alpha}_{\text{kept}}(\pi_R)$	Sampling rate for $\hat{\alpha}_{\text{kept}}(\pi_S)$
Mistrust of receipt	Mistrust of delivery

Table 1: Correspondences between trust, trustworthiness as a currency for attentiveness to promise-keeping activity.

- We define an agent's *capability* as the assertion of a (+) offer promise $\pi_S^{(+)}$ or a (-) receptor promise $\pi_R^{(-)}$. Both offer (+) and acceptance (-) promises are capabilities, by symmetry, since the complementarity law of promises implies that is always a valid reinterpretation of a promise system in which (+) and (-) are reversed [2].
- We define a *need* as the exposing of a promise to receive something from another agent by a receptor promise $\pi^{(-)}$ forms the basis for signalling an agent's need. Thus a need implies a capability to receive.

2.1 Productivity of supply, affinity for demand

Equal principles apply to both the delivery and monitoring channels mentioned above, but the function of monitoring has a special temporal significance. It is used to learn and predict possible future behaviour. An agent that can learn, expect, and prepare for future behaviour can allocate its resources efficiently.

The *productivity* of a supply channel corresponds to *potential trustworthiness*, and effort invested in checking, collecting and processing the product is related to *potential mistrust* or 'anti-trust'. In other words, a decision to trust corresponds to a decision to forego work in monitoring a promise delivery channel. We frame these assumptions here (table 1):

Assumption 1 (Trustworthiness) *Trustworthiness is the assessment that an agent reliabily intends to keep its promise, independent of the final level of success, which may be beyond an agent's control. This forms the basis for a common and exchangeable currency of reliabilty, which may be propagated as a reputation.*

Assumption 2 (Trust and mistrust) *The assessment of trustworthiness is a possible indicator that less attention could be afforded to monitoring compliance with a receiver's expectations, since it has gone well in the past. Trust is an allowance to forego monitoring, while mistrust is an affordance prescribed by each agent as an overhead to watch the delivery channel for promise compliance.*

Watching and monitoring of interactions (i.e. mistrust) costs an agent effort. The default position of doing nothing corresponds to complete trust as well as complete indolence. There is thus a similarity between trust and arguments that surround energy in physics that we can now try to exploit to create a dynamical guiderail for agent reponses.

2.2 Dimensions of key quantities

We can adopt a simple accounting principle for trust by looking to mechanical energy from classical mechanics. A force is a (tautological) influence that performs work to accelerate the rate of an agent, over some displacement Δx in a space of change. If we assume piecewise continuity, a force can be represented by a potential function $V(x)$ with the dimensions of energy. This classical representation was divined for ballistic systems, with trajectories over physical Euclidean space. The dimensional relationship between these quantities is summarized as follows:

$$\begin{aligned}
\vec{F} \cdot \vec{dx} &= \vec{F} \cdot \vec{v}\, dt && (7)\\
&= \frac{\vec{dp}}{dt} \cdot \vec{v}\, dt && (8)\\
&= \vec{v} \cdot \vec{dp} && (9)\\
&= m\vec{v} \cdot \vec{dv} && (10)\\
&= \frac{1}{2} m d(\vec{v} \cdot \vec{v}) && (11)\\
&= d\left(\frac{1}{2} m v^2\right) && (12)\\
&= dT. && (13)
\end{aligned}$$

Although we associate these terms with ballistic physics, this is only a statement of dimensional relationships changing at a deterministic rate.

In the case of a discrete agent theory of promises, with both exterior spacetime positions and interior promise states, we can map the displacement x to a change of state in two different ways (figure 2):

- As a difference in agent $A_i = \{S, R\}$ as a discrete version of the usual exterior $\Delta x \mapsto 1$.
- As a difference in the assessed output state $\Delta\alpha(b)$ of the agent pertaining to its promises Δb.

For discrete agents continuum functions map to assessments made by agents about one another. In configuration space, this requires special definitions (see [7]). We shall mainly focus on the interior states, and represent changes schematically as if they were continuum variables for familiarity.

$$\vec{F}_{(ij)} \equiv -\overrightarrow{\nabla}_j\, V \quad \mapsto \quad -\overrightarrow{\nabla}_j\, \alpha_i(\pi). \tag{14}$$

A potential function V is just an alternative representation of the exterior influence that we call the force, with the dimensions of energy. It's useful precisely because it forms a part of the energy accounting picture. Now V or $\alpha(\pi)$ becomes the fitness landscape or potential surface that we're familiar with in smooth classical systems. Such landscapes have long been used in many fields from economics to machine learning as ways of encoding behaviour. using methods like hill climbing or steepest descent. At every stage in this formulation, agents are behaving independently, but are signalling one another at a certain rate with updated information through channels formed from promises.

The analogous expression in Promise Theory was given in [10]. We can replace x with either agent position A_i, or agent state $\alpha_R(\pi_S) = \alpha_R(b_S)$, the assessment of one agent R concerning the state of a promise π_S with body b_S. In other words each agent views this displacement as as 'my estimate of your level of promise keeping' in units independent of b:

$$x \mapsto \alpha_{\text{me}}(\pi_{\text{you}}) \mapsto \alpha_{\text{me}}(b_{\text{you}}), \tag{15}$$

where π may be any promise. The effective velocity is the rate of change of assessment, or the sampling rate:

$$v \mapsto \partial_t \alpha_{\text{me}}(b_{\text{you}}). \tag{16}$$

The work done in this formulation is a still a total derivative, by design, and thus its sum depends only on the initial and final states. It is not path dependent and is therefore conserved by assumption of continuity. The only process memory lies in the externalization of the field We can integrate it along any path and it will appear to be conserved, by virtue of exterior continuity.

$$\vec{F} \cdot d\hat{\alpha}(\vec{b}_R) = d\left(\frac{1}{2} m_R J_R^2\right). \tag{17}$$

2.3 Definition of trust

Using these dimensional correspondences for rates of work, we can now define the two components of trust in terms of promise-keeping and monitoring work. Trustworthiness corresponds to a potential V (i.e. potential for trust) and is an onlooker's assessment of an agent's reliability in keeping a particular

promise. The reliability can (in principle) be assessed objectively, up to the Nyquist sampling limit of the assessing agent [6, 11], but the translation of that into an estimate of the average required sampling rate turns it into a subjective risk assessment. Seeing that another agent is more trustworthy than itself, and agent may seek to obtain the outcome from the other instead of its own resources. The assessment of trustworthiness thus acts as a causal motivation or 'force' supplying intent for external processes. Thus we now define the potential and kinetic forms of trustworthiness and trust:

Definition 1 (Trustworthiness (potential trust) V) *For the promise in equation (1), R will assess the trustworthiness of $\pi_S^{(+)}$ to be some value V. This is an estimate by R of the rate of promise keeping J_R in relation to that promised by S. The method of assessment is unspecified.*

There is a distinction between the cost associated with keeping a promise, and the cost of assessing whether that the outcome kept the promise. The former may involve assessing whatever was received from the sender. The latter is an additional overhead, which involves a separate sampling process of measuring compliance.

Mistrust is semantically distinct from either of these assessment, since it is a policy that involves an active decision to invest an agent's finite resources in monitoring the collaboration with the assessed party. We define trust T and its mistrust $\overline{T}$ as complements.

Definition 2 (Mistrust $\overline{T}$ (kinetic antitrust)) *An strategy or policy, by an agent R to invest an amount of work $\overline{T}_R(\pi_S)$ in the continuous sampling and assessment of a promise π_S. This work corresponds to a sampling rate $J_R = 1/\Delta\tau_R$ measured in time intervals $\Delta\tau_R$ according to R's internal clock, from the association $T = \frac{1}{2}mJ_R^2$.*

Mistrust corresponds to work done, while trust implies a saving of work. We speak of trust rather than mistrust, perhaps due to our moral biases, but here the contention is that mistrust is the simpler complementary variable to understand the role of trust[1].

Note that these definitions have the following logical properties. An assessment that an agent is not trustworthy to keep π is equivalent to the assessment that the same agent is trustworthy not to keep π. However, not trusting the agent to keep π is not necessarily equivalent to trusting the same agent not to keep π, since the decision not to trust may not involve establishing an attentive relationship of any kind, nor is there any causal dependence between an allocation of trust and rational factors.

It seems natural that low trustworthiness tends to be matched by a low allocation of trust, if the promise referred to is of sufficient value to the agent, since the the probable return on this investment of trust is, by definition, low. Nevertheless, this decision is entirely up to the agent R to decide according to its capabilities and resources. One should not imagine a universal law of rational agent utility at play.

If trustworthiness is built up ratchet-wise as 'credit' earned in increments of promise keeping work, like extending a spring, then once released the effect on sampling rate (velocity) would be only the square root of the trustworthiness.

We have only considered two agents here. Processes are rarely as simple as a single promise single link in a chain. Usually, there is conditional reasoning built into cooperative processes. Trust and trustworthiness still act as a guiderail for such processes.

2.4 Mistrust and feedback in promise keeping

In elementary systems, like physical primitives, one often assumes that promises are kept simply by virtue of being stated. Such a high degree of reliability is what is often referred to as determinism or the invariance of 'physical law', implying an inevitability on over some scale of behaviour. In more complex systems, found on a more macroscopic level, there are many uncertainties that might prevent such determinism from being realized in a simple way. Thus, we seek to verify to assess trustworthiness on a more statistical basis.

It's beneficial in terms of cost saving for an agent to cache or remember its past assessments. Not all agents have a long term memory however. At the elementary level, Markov memoryless processes dominate. Trust is a mechanism for memory processes, so we expect expensive trust variables to be slowly varying relative to primary needs. The specificity of trust, by promise, is a multipler of the memory requirement for recalling agents reliability, thus there is also an economic imperative to simplify

[1] It has been pointed out to me by Daniel Mezick that a view of trust as attention was alluded to in reference [12], which discusses cyclic processes in a machine learning context. The language affects a mysticism which conceals the insight, but it seems to be independently identified before this work.

assessments to the lowest level of detail, by grouping or inferring assessments for different promises based on assessments of partially related promises (see section 3).

A second point makes trust relevant to promised interactions: type and timing influence the coupling of processes. Keeping promises relies on cyclic processes (repeated sampling, delivery in batches, etc). Promise keeping is thus something like a wave process, or may be composed from waves in the Fourier sense. If the phase between sending and receiving is misaligned, then the receiver will not be able to register the event. This implies that unplanned *events* (called impositions in Promise Theory) have a higher chance of being ineffective, i.e. ignored or go missing. This can only tend to lower the assessment of reliability of the sender by the receiver. Thus we have the prediction:

Hypothesis 1 (Impositions reduce trust) *Attempts to implement intentions by imposition will tend to reduce the assessment of trustworthiness in the sender and by implication the trust allocate by the receiver too.*

Conversely, promises that are planned and aligned by a binding process will tend to increase the chance of successful delivery, whether they are short lived or long lived.

Hypothesis 2 (Promise alignment increases trust) *Attempts to implement intentions are maximized by alignment of bodies $b_R \cap b_S$ and the phase of the send-receive cycles and thus trust grows at the highest rate when an equilibrium of regular transfer between sender and receiver is maintained.*

Steady state promises will confer the largest accumulation of trust over time.

If b_R and b_S refer to completely different types of promise, one might choose to say that trust is undefined. However, since we do not expect or demand that agents be rational in anyone else's judgement, an agent might choose to combine this with its coarse assessment of the agent. Miscommunication or failure to comprehend on similar terms would lead to random alignment of intent, and a spurious assessment of trustworthiness.

Indeterminism in the feedback, coupled with the dependencies between the agents' assessments, is a reason to expect chaotic changes in levels of interaction for service relationships [13, 14] if agents reassess their estimates and policies too frequently. Hysteresis or coarse grained responsiveness, over a longer assessment timescale, is a strategy for stabilizing and reducing the cost of assessment. Indeed, we claim that this is the point of trust.

A decisions to mistrust another means an agent might not even be willing to invest the effort to assess their trustworthiness up front. Thus, trust doesn't imply favour. It implies consistency of behaviour (good or bad, positive or negative). The meta policy behind these interactions is a desire to save on the penalty of work cost.

2.5 Promise keeping rate or process velocity

The promise-keeping current may be defined by the Shannon-Nyquist sampling process as a series of assessments $\alpha(\pi)$ or a promise π, and counted over time. Note that no agent can measure something without going through the process of assessment, so the only measurables are functions of assessed events.

$$\vec{J}_{SR}(\pi_S) = \sum_{\tau=1}^{\tau_{\max}} \frac{\alpha_{\text{kept}}(\pi_S)}{\tau_{\max}}, \tag{18}$$

where $\alpha_{\text{kept}}()$ is the assessment of the promise that the outcome fell within the bounds of $b_R \cap b_S$. Similarly, for the promise in equation (2), S will assess the trustworthiness of $\pi_R^{(+)}$ to be:

$$\vec{J}_{RS}(\pi_R) = \sum_{\tau'=1}^{\tau'_{\max}} \frac{\alpha_{\text{kept}}(\pi_R)}{\tau'_{\max}}, \tag{19}$$

where $\alpha_{\text{kept}}()$ is once again the assessment of the promise that the outcome fell within the bounds of $b_R \cap b_S$. The assessment current has minimum and maximum cut offs by virtue of Nyquist sampling (max rate). The process of keeping a promise of contemporaneous with the process of assessing the outcome, so these compete for the resources of the agent making the assessment. Over long timescales, with statistical continuity, we can presume to write this as a derivative with respect to the timescale of the agent making the downstream assessment (denoted τ_R). The short term 'instantaneous' current is:

$$J_R(\pi_S) = J_{\text{monitor}} = \partial_{\tau_R} \hat{\alpha}_R(b_S) \quad \text{(receiver assessment)} \tag{20}$$

$$J_S(\pi_S) = J_{\text{assert}} = \partial_{\tau_S} \hat{\alpha}_S(b_S) \quad \text{(sender self-assessment)}, \tag{21}$$

where the derivatives are with respect to the proper time clocks τ_S and τ_R of the agents making the assessments. Over sampling of outcomes is costly[2]. We expect the cost of monitoring a promise to be optimized when the send and receive processes are aligned in phase and the rates are identical:

$$\text{cost} \sim \begin{cases} \text{impact of missing an event} & (J_S > J_R) \\ \text{cost of sampling} & (J_R < J_S) \end{cases} \tag{22}$$

So we can postulate that the simplest function satisfying these conditions has is quadratic in the difference: If μ is the impact of missing an promised event.

$$\text{Potential Cost} \sim (J_S - J_R)^2 \tag{23}$$

So, we expect the work cost to vary like the square of the relative velocity in promise space.

Trust need only deal with the counting of attempts to keep a promise rather than in the level of compliance. This is a matter for each agent to decide. In other words, it deals with a flow of assessments rather than the flow of bulk goods, services or other subject matter. The maximum rate of checking the status of the slow depends only on the receiver, and this could be updated more than the delivery rate leading to inefficiency, or less than the delivery rate leading to missed events. The amount of work invested in monitoring a service is thus not limited by the process and is an independent overhead. This monitoring relates to the trust, because it's about the assessment of promise keeping. The finite resources of an agent imply that there is a maximum limit on J for any agent, which will lead to 'clipping' of the sampling signal and possibly chaotic transients from discrete finiteness effects.

2.6 Trusting and attraction to trustworthiness

As defined, trust relates to a desire to save effort in monitoring and verifying work, where agents with finite resources. What distinguishes (mis)trust from ordinary productivity assessments for supply and demand levels is that it is an overhead, and effort is saved when it is unspecific. We return to this point in section (3). This aligns its properties well with the requirements for a common currency, involving all kinds of promised activity. So trust obtained in relation to one promise may be transferred in a limited sense to other promises. This desire for a common currency illustrates the common roles attributed to trust, money, and energy.

Mistrust or anti-trust corresponds to an investment of attention (employing available resources) for monitoring the intention to keep promises between agents with a promise binding as in (1) and (2). Our operational hypothesis can then be states simply as two biases:

Hypothesis 3 (Parsimonious trust biases) *Agents may have a bias towards overestimating kinetic trust, because this minimizes resource consumption. Agents may have a bias towards underestimating trustworthiness, since there are more ways for reliability of a cyclic process to fail than to succeed, and inattention will lead to reduced assessment reliability.*

We may ask what purpose trust serves: what value is there to finding out a promise is not being kept? The only possible answer concerns the agent's needs and the long term tally a kind of debt that one expects to recover in the future. Trust is thus about smoothing or averaging of statistical processing—*time management* of flows and deficits, potentials and exchanges. The implication here is that receivers may be attracted to bind to new promise providers or to change promise terms by their relative assessed trustworthiness V. Colloquially:

- In configuration space: (figure 3) if I see that you are more trustworthy in delivering a result than I am and our neighbours are, I will gravitate towards you to obtain the service. There is thus a motivational force

  $$\vec{\nabla}\alpha_R(\pi(A_i)), \tag{24}$$

  which is effectively the gradient of the local promise assessment field $\alpha_R(\pi)$. Conversely, they might be repelled by low trustworthiness (high risk) where there are alternatives.

- In promise body space: (figure 4) if I see that a different choice of my promise body b can be delivered with more reliability I am motivated to change it and perhaps adjust my own acceptance promise to maximize the reliability, hence reducing the need for monitoring.

[2]Calling out "are we there yet?" from the back seat doesn't help the family road trip to arrive earlier. This is sometimes called busy waiting in computer science.

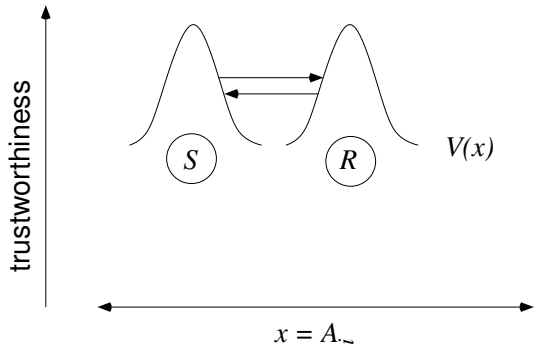

Figure 3: Potential is agent configuration space. A stable promise relationship has trustworthy agents in configuration space and a 'contained' promise relationship between them.

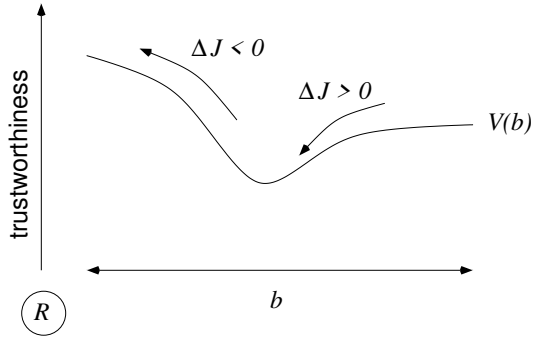

Figure 4: Trustworthiness as a potential function in promise b space. Variations in b might increase of reduce the trustworthiness or supply confidence. If trustworthiness (supply confidence) goes down, agents are attracted by the risk but increase their (antitrust) sampling velocity to compensate. If trustworthiness (supply confidence) goes up, agents decrease their sampling velocity, i.e. offer more trust.

- The association of kinetic (mis)trust with attentiveness is equivalent to a determination of the sampling rate of the receiver. The work involved in monitoring at a rate J is based on the dimensional arguments for work in equation (13), which in turn is based on the motivational alignment of work with the directionality of source and receiver. This is what we call intent.

- Antitrust $\overline{T}$ is the complement of trust allocation T and is associated with an affordance of work for due diligence in monitoring an exchange. For any promise we can define the implementation work as antitrust or overhead work of verifying the promise.

$$W_R = \frac{1}{2} m_R J_R^2(\pi_S) \mapsto \overline{T}_R, \tag{25}$$

- The suspiciousness, riskiness, or potential for loss is the complement $\overline{V}$ of trustworthiness or productive power V. By convention potentials are more negative when attractive, so this power is analogous to a potential well. If an agent moves from a promise position (described by the terms of the promise body b) of trustworthiness V to one of less trustworthiness $V - \Delta V$, then this is accompanied by a potential difference ΔV over the displacement Δb, as above, and this tends to increase the antitrust $\overline{T}$ to increase the rate of work in sampling to verify the output. For time continuity:

$$\overrightarrow{\nabla} V_b \cdot db = d(\frac{1}{2} m_R J_R^2). \tag{26}$$

  The square relationship is a consequence of the involvement of the intent with the direction and magnitude of the rate itself, and the effective mass or involvement of the agent in local relationships that draw on its finite resources and tend to slow it down.

2.7 From energy to velocity, from trust to sampling rate

In physics, energy is the generator of time dependent behaviour. It's conservation over some scale implies continuity of behaviour in time. As in the case of energy, the absence of kinetic trust is related to the work done. Trusting another agent is thus the minimal default position, while mistrust is an cyclical attentive process in assessing compliance. The work is done at a rate $\dot{x}$ or J, which is the flow rate of information along the channel. This in turn is an agreed level determined by the production rate and sampling rate of S and R respectively.

From trust (which has dimensions of energy in order to generate time steps for assessment), we look for a monotonic function of sampling interval or frequency. The sampling cost of increasing the sampling rate is the anti-trust:

$$\Delta\text{Work of overhead}_R = \Delta\overline{T}_R(S), \tag{27}$$

$$\Delta\text{Work of overhead}_S = \Delta\overline{T}_S(R), \tag{28}$$

i.e. the change in work done by R goes to reducing or changing the kinetic trust R assigns to receive (-) from S. Similarly, any change in the rate for S which applies to sampling or keeping the complementary (+) promise.

Although we tend to think of energy (as we think of money) as an independent and shared standard of flux, the actual accounting is entirely local, more like international currency exchange. This is emphasized in an agent model, where all activity happens within the agent's container. Trust values are not universal; they are private interior assessments and the agents remain autonomous. Thus, there is no sense in which one agent can force another to change. Each agent maintains its account only of its own investments and observed outcome receipts.

However, indirectly, over a longer timescale, extra effort (cost) expended by one side may be counted as a benefit on the other, but only if it adjusts its own rate of work to absorb what is sent. Without cooperation, the extra effort will simply be lost. Thus, while there is no determinism, there can be a stochastic conservation on average by continuity. Impositions, as singular events, will typically fail this test, and thus trust or energy will not be conserved when imposing on a receptor that is not adjusted to accept it. This is the situation for all foreign currencies.

Close to balance, a change in potential will cancel a change in activity, so a variational principle of least action applies:

$$\delta I = \delta\overline{T} - \delta V \simeq 0 \tag{29}$$

This immediately leads in turn to the Newtonian differential equations. So the accounting is a self consistent hypothesis.

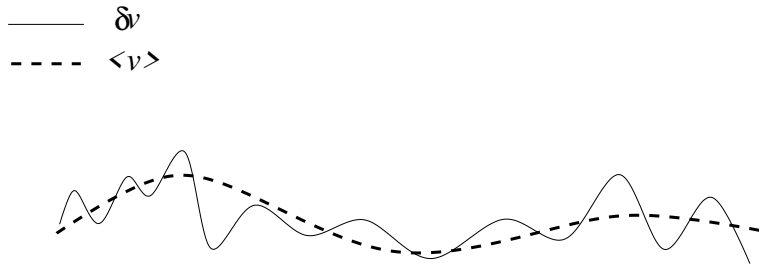

Figure 5: Separation of timescales into fast and slow variables.

In dynamical descriptions, it's expedient to define fast and slow variables as these couple weakly and allow us to use a perturbation expansion based on the relative stability of a slowly varying context against the more unstable ephemeral fluctuations that we can't rely on for prediction. Thus variables are often written as a decomposition

$$v = \langle v\rangle + \delta v, \tag{30}$$

in which $\langle v\rangle$ is a coarse grained moving average over some assessment time interval, and δv is an instantaneous sample relative to the moving average.

Stability is greater when trustworthiness is a slow variable, involving statistical learning of the behaviour of an agent. Trust may be a faster variable that responds to impositions and other transients. We

must therefore define the timescales agents use to distinguish these assessments. Ironically, impatience on the part of an agent can result in trust instability, since assessments may be dominated by fluctuations. This is why transient impositions typically lead to a reduction of trustworthiness while the continuity of promises tend to build it.

3 Scaling of assessment and trust in simplicity

The foregoing sections have been leading up to an important consequence of the model of trust proposed here. Hypothesis 3 almost implies the result as does the suggestion of a least action principle. Let's state it like this: if a promise is too complex to accept, agents may assess it as untrustworthy, even though it might be reliably kept—because the receiving agent may not be willing or able to verify it. Thus there is a probabilistic scaling of the economics of trust based on relative complexity. Let's state one more hypothesis in a suggestive form:

Hypothesis 4 (Preference for rough assessments) *In allocating trust for cost saving, agents will tend to trust coarse and low detail promises more than highly detailed promises as trusting is saving.*

Agents may therefore default to the coarsest representation of a promise, e.g. associating any detailed terms with the agent's name, image, or with the name of a group or pattern it belongs to.

In other words, the less specific an issue, the easier it is for an agent to trust. Although this is a simple cost issue its implications are profound. As a corollary, one could add that agents with memory will tend to replace detailed assessments with cached associations to identity.

One might look for an objection. The following example miight seem to contradict this principle: if something is too cheap, we are not happy, we mistrust it as a probable deception. In this case the cheapness of the product is the simple coarse assessment that we focus on. The cheapness of the assessment lies in choosing the cheapness of the product as an assessment. So we trust our judgement that cheap things are probably not all they claim to be. It's interesting that we are sometime less able to apply this principle to trust itself. See, however the matter of credit checks under risk below (section 4.4).

The effects of scaling go beyond this simple characterization, because processes and things are, in general, formed by the composition of smaller processes and things. This in turn means that their assessment can be watched over on multiple scales, depending on one's level of involvement with the whole. We need to ask: how long with trust or confidence last, given the cyclic nature of interaction at steady state, or the ephemeral nature of transients (impositions).

An example helps to illustrate this quickly. Consider the design of a building or the playing of music. An entrepreneur or musician cannot be relied upon to produce something that aligns with expectations without guidance. One could micromanage the promise keeping either in realtime or by writing plans, recipes, and scores for the desired outcomes. This effort is, on one level an expresssion of antitrust. However, on another level, the promise has now been altered to keeping to a process guide (the recipe or musical score). Now the mointoring only needs a manager or conductor to assess and potentially correct significant deviations. The level of checking can now be dialled back at this level, because the effort has been put into guidance from the start. Thus the monitoring of promise keeping can be viewed on different levels, and with different strategies in space and time. We can defer elaboration on this matter for later work.

3.1 Examples

Example 1 (Late and unreliable, kick the TV) *Tardiness in a service or an agent goes to the assessment of reliability (tristworthiness) and mistrust kicks in, provoking a higher rate of attention. We respond with the only thing we control (our own due diligence—for condfidence we may introduce detailed testing both dynamic and semantic tests): if confidence in another party fails, what can we do to mitigate the risk? If we are dependent on the other, we still want to try to affect the system causally so often we impose blame (a transient response to provoke a change, like kicking the TV or pushing a broken down car).*

Example 2 (Kin, tribe, language, country) *Identifying promises with agents who have prior familiarity makes trust is cheap. If we trust simplicitly, then the simplest assessments are the most familar since we've cached these items in memory which is quick and cheap to look up.*

Example 3 (Man of the people) *When leaders make promises, they may benefit from making vague promises, since these are easier to trust than complicated promises. People who talk in simple terms are 'folksy' unless they can talk the same jargon we can. The language of promises is a coarse tribal identifier.*

Example 4 (Brands and logos and icons) *Marketeers can hack the tendency to prefer simplicity by using visual symbols, catch phrases, and other low information patterns for promise identity. This increases the likelihood that agents will trust their promises.*

Example 5 (Terms and conditions) *When there are fewer terms and conditions to satisfy it's easier to trust.*

Some might argue that working out those terms and condition together brings greater trust in a partnership. This would be a misunderstanding. The process of agreement might bring trust, and this might be summarized as terms and conditions that are effectively already promised. However, imposing terms and conditions without that relationship growth would be entirely detrimental to trust. The same is true of laws that seem to make no sense.

This means generic promises that are easy to understand are more likely to be trusted, even if they are unlikely to be kept, than very complex promises of high fidelity. Long contracts feel less trustworthy than short ones. Complicated arguments are harder to accept than simple arguments, etc.

As an agent learns, it will achieve greater capacity or experience to draw on which makes the cost of assessing complex scenarios relatively cheaper. This seems to contradict the idea that trust is a cost saving issue. There is a difference between the policy for allocating trust and the assessment of trustworthiness. If no assessment is made (because it's expensive) then it might simply default to a protective no.

Example 6 (Patterns of behaviour) *Agents assessing the trustworthiness of collectives and institutions may look to the sum of their promises and behaviours and infer a pattern of behaviour which they assess together. Thus, clients of a company, or visitors to a region or country might look past the detailed or specific promises to imagine a coarse grained 'cartoon' entity judged on simple terms as trustworthy or untrustworthy. Such simple characterizations can be easily kept in memory, cached for avoiding future work.*

Avoiding generic interactions is one way to avoid generic assessments. Detailed and more intimate contact that aligns with the recipients needs (receptor promises) focuses attention on relevant aspects of a relationship rather than on generic patterns already prejudged.

As agents capacity increases and they become more complex, they can afford to make more complex assessments, and even cache previous results in memory so complex recalculations can be optimized.

Example 7 (Trusting one's own judgement) *People seem to trust their own opinions about things they know little about (political ideas etc) especially if these align with their existing receptors (biases). They are more careful when they have some knowledge because a blunt receptor can choose any level at which to accept an assertion. Thus people are easily brainwashed by simple slogans and symbols, but less easily brainwashed by complicated facts.*

We should beware the mixing up of trust with a rational assessment of confidence or belief in one's ability. The decision to trust one's own judgement is to forego rational evaluation, and proceed without thinking. Many misunderstandings about trust revolve around this point.

Example 8 (Submission and peer review) *An author writes a paper and sends (imposes) it to a journal or to a preprint archive, which has a steady state promise to receive and evaluate submissions, making the conditional promise to publish or reject based on the assessments of reviewers.*

$$\textsc{Author} \overset{+submit}{\longrightarrow\!\blacksquare} \textsc{Editor/Gatekeeper} \tag{31}$$

The editor or gatekeeper makes an assessment. Depending on the agent this might be detailed or rough. On the principle of cost saving, it will probably be assessed coarsely. The receiver might look at the names and addresses of the authors and judge them based on reputation or prejudice about their institution. An assessment of low trustworthiness for these identifying characteristics in relation to submissions. This is an assessment of the intent to publish. The receiver might also assess its confidence in the submitter (e.g. by testing): how competent is the author likely to be? These assessments are quick, before any detailed work. On this basis alone, the gatekeeper or editor might reject the paper as 'not suitable

for this venue', much like entry to a VIP club. This the effort of a detailed assessment is saved by mistrusting. The gatekeeper might see some future advantage to publishing the paper, e.g. publication charges or reputational association, in which case it accepts the risk of arranging a detailed assessment with monitoring of that new process.

Example 9 (Interview person of interest) *An interviewer imposes questions that suggests that they have a low estimate of the interviewee's trustworthiness, and suspect (have low trust for) some deception. An interviewee with low self-confidence may not try to apply effort to answer the questions correctly or accurately. Instead of assessing the interviewer in detail, they cheaply assume hostility or low trust. The agent is looking for a promise of personal safety, and finds hostility or risk. Lying could be a cheaper option.*

Example 10 (Quantum mechanics and measurement) *The similarities between virtual processes and quantum mechanics are hard to avoid. The matter of trust does not enter into quantum mechanics in a direct way. However, measurement or observation does. One might be tempted to interpret Quantum Theory in terms of energy and its relation to sampling. Energy appears in two forms in the Schrödinger equation: as a time derivative and as a boundary condition. In a sense this is hard coded by the Hermitian symmetry of the formalism. There is a kind of reliability of trustworthiness to the probability expression $\psi^\dagger\psi$, but there is no agent on the receiving end of the standard theory taking on a role to invest trust in the wavefunction. One has no way of calibrating the results, so one has the curious position of having no alternative but to trust the predictions of the theory completely, but to mistrust the outcomes by measuring as much as possible. That is the only status for a theoretical prediction.*

However, one can more easily apply the trust theory to the operation of a detector, where the efficiency of detection may only be a fractional percentage. One might trust the detector to detect what it should, but with limited confidence in its efficiency. In order to overcome this, the observer calibrates the detector, then trusts it within the experiment to have that level of efficiency. So finite trust still plays a role in cases where measurement can be independently calibrated.

Example 11 (Compliance with standards) *Standards are proposed promises that others are expected to adopt and promise themselves. The body of promises is a standard is given a name, e.g. ISO123456 or NIST ABCDE, etc. Agents that keep their promises are trustworthy, but this might be expensive to evaluate so few will try to check. Agents who promise compliance by name dropping the standard name (whether true or false) may easily deceive receivers that they are trustworthy, since the coarse brand of the standards associated with trust makes them easier to trust according to the hypothesis.*

Example 12 (The CAP conjecture for databases) *Consistency is a simple promise, but a hard technical thing to achieve.*

The simple narrative about databases in terms of three alternatives is easy to trust, but more importantly it was quickly branded in the industry as an icon. Availability is simply related to reliability of communication, which is thus simply related to trustworthiness.

Consistency is another promise for which assessment is required but expensive to establish.

Example 13 (Security and state monitoring) *We monitor systems because we expect occasional shortfalls in the keeping of their health or promise keeping. Many monitoring systems are not well aligned with the promises the system is trying to keep, so there is a large element of trust in those who do the monitoring. The cost of monitoring is often quite high, because one places more trust in the watchers than in the system itself....*

In each case, the scaling law for coarse graining plays a prominent role in the way we use trust.

Axelrod made the notion of tit-for-tat game play famous in his studies of Prisoner's Dilemma models [15]. These have become the archetypal model for cooperative dealings in economics. Promise Theory predicts the Prisoner's Dilemma model in a certain limit, but can be richer and more dynamic in its formulation.

In the example of a paper being rejected above, it's reasonable to think that a human agent suffering the rejection would harbour ill will towards the referees and might react by making accusations (rightly or wrongly). In Promise Theory, accusations are typically impositions that are aligned counter to promises of the recipient. If the agent receiving an accusation assesses this to be potentially harmful to it (e.g. by reducing its reputation or others' assessments of its trustworthiness) this may lead to a downward spiral of events turning from cooperation to anti-cooperation. This is a possible side-effect of the coarse graining hypothesis. The accuracy of assessments and promises or accusational impositions based on them is a fragility in cooperation. This is one reason for due diligence to be so important in sensitive matters where risk is in play.

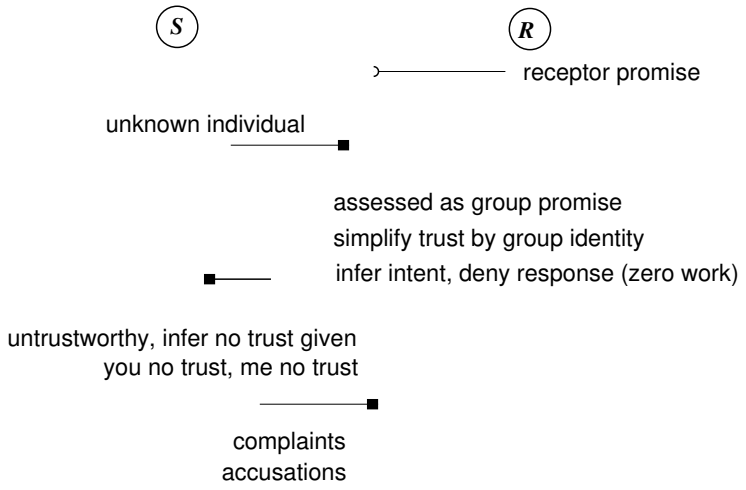

Figure 6: The perils of transient impositions. Cost saving can quickly lead to a downward spiral of trust and a storm of wasted effort.

4 Other cooperative potentials

Let's distinguish trust from a small number of superficially related characteristics in common usage: namely risk, confidence and hope, in as precise a way as we can. Any theory should match commonplace meanings and intuitions as far as possible. All these potentials play the role of predictive Bayesian potentials, derived from prior learning. It may be informed from experience, but it has a speculative prior assessment as its input.

In the literature of trust, authors sometimes remark that the willingness to trust has something to do with state of vulnerability of an agent. We can try to explain this remark to the extent that it makes sense in the framework of Promise Theory. In the present model, this is a misinterpretation of a potential we might call hope.

4.1 Vulnerability and threats

Vulnerability is a potential for extracting a high cost from an agent, e.g. a promise that costs too much to fulfil, or which exposes the agent to an unwanted delivery. Vulnerabilities are most intuitive when they lie in the promises a receiving agent, as exposures to threats. Threats are promises or impositions that are intended to cause harm, in some interpretation of the receiver. In an autonomous agent, harm can be caused by exhausting its resources or by otherwise altering its processes somehow.

An agent may thus find itself lured by a potential V, by apparent trustworthiness, into keeping a promise that depletes its budget. Caught servicing too many promise relationships or possessing insufficient interior resources to keep its own promises would then motivate it to reduce its expenditure on overheads, which includes being coerced into increasing its trust level in promise relationships. This is consistent with the idea that a state of vulnerability can increase a tendency to trust, but this is anomalous behaviour. In other ways, potential vulnerability would likely make agents more suspicious to avoid trust relationships if they were exposed to harm.

Vulnerabilities often lie in a promise relationship being too general and non-specific. This is one reason why trust itself can be a vulnerability, and is abhorred in computer security: if we accept hypothesis 4, then trust prefers the unspecific, but detailed specificity is the way to narrow risk.

4.2 Agent confidence and doubt

Confidence is a supericially similar assessment to trust, but with different semantics. If *trust* is a policy, which means 'no need to check for activity', then *confidence* in a promise outcome means 'no need to doubt' that the agent is capable of doing so, or that its promised outcome will be all it's claimed to be (fit for purpose). Confidence is an assessment about a forthcoming promise outcome, or by coarse transference in the promising agent itself. Confidence is often assessed by detailed testing, such as fire drills and software testing.

The semantic of confidence are more specific than those of trust. They imply that an agent is both willing and able, i.e. both *capable* and *trustworthy*, to keep its promise to the full extent defined by the promise body. Like trust, confidence may be assessed over different timescales, and we resist the urge to write a meaningless formula for confidence, since it could be assessed in any manner the agent pleases or is capable of. One may, for instance, assess our confidence in an agent to assess confidence in itself and others.

The complement of confidence is doubt. One can doubt the veracity of a claim, while trusting that the claim is well intentioned and accurately formulated.

Example 14 (Can I get there by bus?) *In waiting for a bus, one might not trust the buses, but one might have confidence that the bus can get you there in time. Confidence may look beyond qualifying conditions, such as 'if the bus even comes'.*

Example 15 (Passed the test?) *Confidence in the promise that someone has passed their driving license, implies a competence to drive, whereas as the trustworthiness of the statement that someone has passed their driving license only involves only the assessment that they are reliable in making such statements or not lying.*

Example 16 (Confidence in statistics) *A confidence interval in data is a range in which a specific hypothesis promises to be true, whereas trustworthiness of data only implies that the data are real and present.*

4.3 Hope and blame

When an agent has low confidence in a promised outcome being kept and no other alternatives to quench its needs (alternative redundant suppliers, see the Downstream Principle in Promise Theory [2]), then it enters a state of 'hope' or hopefulness, which is a state of hightened anticipation of an outcome. It's a form of antitrust, in the sense that expressing hope or hopefulness is a passive reaction of individual urgency that tends to stimulate 'busy waiting' or oversampling to capture the earliest possible response ('are we there yet'?). The degree of oversampling or *hopefulness* is amplified by the economic neediness of the agent which is downstream of the promise, by some individual policy.

For a sender, hopefulness refers to the acceptance of its offer. For the receiver, hopefulness refers to the delivery of what is offered. Hopefulness is thus dynamically or operationally equivalent to an excited state of mistrust.

If hope is a passive response to imminent failure, an active response is to impose blame. Agents sometimes express their lack of planned redunancy alternatives by imposing blame (as a form of accusation), which is often a futile gesture since the loss may not be intentional and the accusation may only reduce future trust further (see Bergstra's work on accuations in Promise Theory [16]). The purpose of blame, like kicking a machine that doesn't work, is probably the hope that it will induce cooperation. Impositions are attempts to induce cooperation [2].

4.4 Risk and risk appetite and recovery races

The action potentials and assessments we've looked at so far all concern the secondary promise monitoring channel referred in section 2, and a decision to choose the time between samples $\Delta\tau_R$ for processes that are dominated by steady state behaviour. Risk is similar in spirit to these other potentials, but its semantics also include assessments about the amount of loss $\Delta(b_S \cap b_R)$ in the transport channel, when a promise is not kept. It thus combines the uncertainties in both channels.

Although the term can be used in an ironic way to refer to any outcome, e.g. to risk making someone smile, risk is most commonly used (especially in economics, safety, and security professions) with the assumption that it relates to harmful future changes—with potentially major impacts. Moreover, risk accounts for the fact that the impact may grow with time, and thus leads to a race to mitigate the time dependent impact, like plugging a leak or a wound. This where the trust sampling rate enters: noticing a problem quickly may result in cost savings, so risk pits transport cost against monitoring costs in a way that trust does not. A high risk event motivates the cost of mistrust in high frequency monitoring.

Risk is an ad hoc assessment, like the other potentials, that agents use all possible methods at their disposal to define. Some authors have suggested to define risk as the probability of an event multiplied by its impact, i.e. a kind if expectation value of impact. However, neither probability of occurrence nor impact are calculable in practice for events of interest. Risk is commonly associated with natural

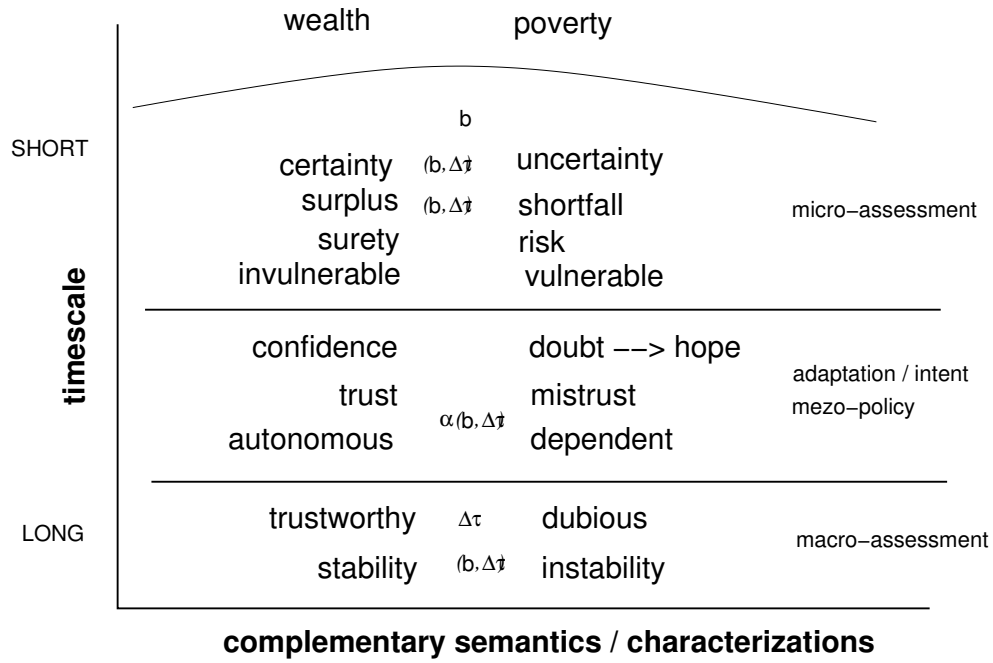

Figure 7: Polarization of assessed characters. Without moral interpretation, the terms we use to characterize the state of an agent tend to align with its condition of resource independence. This is a natural alignment due to the ground state of autonomy (causal independence) and the ability for agents to obtain semantic and dynamic capacity from interacting with others. This explains why there is a tendency to respect wealth and despise weakness in society.

disasters, acts of god, theft and willful harm imposed by others, which are not regular or predictable events. In that case, the emphasis shifts to the race against time following an event to recover, which goes beyond the scope of the present discussion.

More important than calculabilty is that risk fits into the scheme of semantic potentials in figure 7 and thus it has two parts: potential and kinetic assessments, though these might be less useful in a predictive sense than they are for more steady state processes. Let's simply define.

Definition 3 (Risk) *A tuple of measures* $(R, \Delta\tau_{repair})$ *characterizing a supposed rare transient event. Once such a transient event has occurred, the agent expects to lose its promised outcome at a bleed rate of* $R = \partial_\tau \Delta(b_S \cap b_R)$*, as time* τ_{repair} *increases, until the promise can be repaired, where* τ *is the proper time of the agent. The agent's estimate of the expected impact cost* $C = R\Delta\tau_{repair}$ *to the promise outcome, measured in the units of the promise.*

Note that there is no 'Mean Time To Repair' (MTTR) for transient risk events, since we cannot collect data for averages. If an agent has no repair strategy in place, then the risk is in principle unbounded. If one applies the same characterization to a steady state, then risk is simply the antitrust for the cost of the system relaxing due to perturbations. The kinetic response complementary to risk may be called a policy for *risk appetite* or willingness to take on risk.

Definition 4 (Risk appetite) *A policy for foregoing the cost of risk preparations.*

Note that one cannot forego the cost of repair as long as the risk is not zero.

As usual, each agent may define these assessments in its own way. Operationally, risk is simply another assessment of potential cost, similar to trustworthiness but with different semantics. Agents are attracted to configurations with less risk (greater trustworthiness), but there is a complication in the cost accounting for risk. A policy for minimizing the bleed time of a transient is in opposition to the policy for trust (saving effort of monitoring). Where the potential loss Δb is large, it will outweigh cost benefits of trusting (see the credit check example below).

The potential and kinetic work interpretation of risk is less useful than for trust, because risk involves mainly transient change events where history has little relevance. Planning risk mitigation then has to rely on having reserves to draw on (a reservoir of savings to draw on in order to recover), which in turn may involve a long term strategy for accumulating such a reserve.

Working with risk rather than trust, one supposes that some effort has been made to confront the potential impact or penalty of the unwelcome event in advance. We expect to handle the detection of risk events in the same way as detecting infractions of trust, by frequent sampling. However, whereas trust does not take into account the impact of promise infractions, risk tries to account for this. An receiver agent's willingness to take on risk is to promise about a certain level of confidence in its own capabilities to contain the magnitude of and/or recover from the impact of, changes imposed by the sender for self preservation. Both imply foregoing the allocation of resources to avert loss. The cost of work (being over attentive) is pitted against the risk of being inattentive. We remind readers that agents can only promise their own behaviour, so the fact that risk is a distributed phenomenon makes it impossible for any agent to mitigate risk on its own.

Example 17 (Intentional misinterpretation of intent) *In the 21st century, language and behaviour have often been weaponized by for political purpose, by imposing unintended interpretations on imprecise statements in language. In order to take offence and win moral highground. There is thus a risk in making statements on sensitive topics that the formulation of a promised statement can be redefined by the recipient to impose an accusation [16].*

Example 18 (Credit checks) *One example where mistrust is common is in financial matters. The shift from trust based banking to performing fervent credit checks (a form of interbank reputation) before every significant business relationship is a notable shift away from trust in online society, which surely came about in response to an increase in fraud.*

5 Automating operational trust policy

In the case where automated systems wish to emulate a trust evaluation of a client server relationship, we can summarize a protocol to evaluate the key aspects of the model and examine how feedback may stabilize or destablize relationships in meaningful ways, as a service level objective. This is a version of that discussed in [13, 14], and will be studied elsewhere.

1. Do we have a prior history promise keeping π_S for which we can assess $\alpha_R(\pi_S) \mapsto V_S$, at the starting time τ_0 for R

2. Do we have either a globally agreed reputation assessment $\rho(S)$ or a number of individual assessments from other agents, from a local network in which S and R are embedded? Otherwise choose a default reputation $\rho_0(S)$.

3. Decide and compute the initial trustworthiness $V_S(\pi)$, using the two information sources in 1 and 2.

$$V_S(\tau_0) = V_S(\alpha_R(\pi_S), \rho(S), \tau_0, \ldots). \tag{32}$$

   The ellipsis refers to the fact that R is free to make any assessment of π_S, based on any variables it chooses. This is entirely R's choice and only R has to make use of it.

4. Select a policy for allocating kinetic antitrust $\overline{T}_S$, and map this to a sampling rate for observing and verifying the promise outcome as some monotonic function of the antitrust:

$$V_S(\tau_0) \mapsto \overline{T}_S(\tau_0) \mapsto 1/\Delta\tau_R \tag{33}$$

5. As time elapses by transactional ticks: promise kept, promise assessed, etc trust will increase or decrease as the system reaches a steady state. This corresponds to a change in the sampling rate of the order of the square root.

$$\overline{T}_S(\tau) \mapsto (\Delta\tau_R(\tau))^{-2}, \tag{34}$$

   corresponsding to a change in activity.

   Periodic reevaluations of the trustworthiness based on new data give rise to a Bayesian learning, with fast and slow variables. The memory function V is more slowly varying than the immediate response $\overline{T}$.

$$\begin{aligned} \text{Total trust} &= \langle v \rangle + \Delta v & (35)\\ &= V_S + \overline{T}_S. & (36) \end{aligned}$$

6. When R in future comes to promise to S, it will make the decision based on its assessment of total trust, and the refresh rate for doing work to keep its promise will be a function of the kinetic antitrust $\overline{T}_S$.

As an agent uses more resources to ensure its own outcomes, it has less of them to spend on attending to checking others. So it would have to become more trusting of them, by a resource reckoning. There is an apparently virtuous reinforcement cycle in this reasoning, which sounds appealing, but it has nothing to do with being morally good. The trigger to mistrust may lie in general motivational stimuli [1] that we call interest or curiosity: anomalous events compared to steady state behaviour. Anomaly detection is a common approach to monitoring.

6 Summary

An application of Promise Theory reveals a simple model of trust and trustworthiness between a pair of agents, as action potentials used in paying attention to promises. Some evidence for this comes from neuroscience [1] as well as a range of other disciplines; but ultimately this is a speculative interpretation that remains to be verified. In this hypothesis, potential trustworthiness is an individual assessment of an agent's reliability in intending to keep promises. Kinetic trust is an individual willingness to forego the overhead of attending to (monitoring) a promised outcome. As in physics, where potential and kinetic energy guide the average motion of projectiles, the components of trust acts as an accounting system for summarizing past behaviour and for guiding reponses. The likeness to energy is not a metaphor; it is an analogous process scheme on a macroscopic scale.

In this summary, we focus on just two agents, but few social interactions involve such simple dynamics. Usually multiple co-dependencies between agents have to settle into a steady state of network behaviour to understand how agents will perform in clusters. There are many details missing from this idealized two-agent interaction: conditional promises and their role in dependencies will become important on a network level, for instance [2].

Trust likely is a helpful emergent regulator of cooperative workflow, which makes it plausible to summarize independent behaviours as part of a common single-valued 'landscape' or potential function. This is a technique that has been highly successful in physics and which provides an explanation as to why sociophysical studies based on physics analogies can work [17,18]. By formulating trust in this way, we take a first step to deriving patterns and 'laws' for social behaviour, in which the characteristics of individual agents are either upheld or purposely scaled away, in a quantitative manner that withstands scrutiny. This is the ultimate role of science.

We do not need to rely on moral notions to understand trust and trustworthiness; they are individual assessments that play a role in managing intentional behaviour under limited resources. The allocation of trust by a promise recipient is a policy for foregoing the overhead of 'due diligence' or the busy work of monitoring. It could be repaid through savings on overheads by similar investments in the future, resulting in an emergent ledger of cooperative debt. Agents average return on an investment of trust is a matter for of statistical balance in the system as a whole. Trust also lubricates a promise relationship, because foregoing the cost of promise validation tyically speeds up processing by eliminating the inertial bottleneck of ambient monitoring.

Trust relates to intent as in the manner of a correlation. Several derivative notions, like confidence (expected success) and risk (possible impact to vulnerability) can be defined in the same way by considering the channels of cooperation. These operate in the same basic way as trust, but have different operational semantics, and are frequently muddled together. For example, in recent times we see a shift in hiring practices away from managing trust (optimistically) to managing risk (a more fearful stance) of employing someone. Instead to seeing qualities of intent that can grow within a group, one looks at how much the person will cost as a risk. It seems doubtful that such a strategy would be optimal in the long run, since individual potential will only be realized on the scale of the group through sustained learning. Such matters remain to be formalized in future work, though some network effects were discussed in [3,7].

Acknowledgment: This work is supported by NLnet project Trust Semantic Learning and Monitoring. I'm grateful to Edmund Humenberger for helpful discussions.

References

[1] S.I. Di Domenico and R.M. Ryan. The emerging neuroscience of intrinsic motivation: A new frontier in self-determination research. *Frontiers in Human Neuroscience*, 11, 2017.

[2] J.A. Bergstra and M. Burgess. *Promise Theory: Principles and Applications (second edition). $\chi t Axis$* Press, 2014,2019.

[3] J.A. Bergstra and M. Burgess. Local and global trust based on the concept of promises. Technical report, arXiv.org/abs/0912.4637 [cs.MA], 2006.

[4] J. Bergstra and M. Burgess. *Money, Ownership, and Agency.* χt-axis Press, 2019.

[5] M. Burgess. Notes on trust literature, bridging the perspectives of social philosophy and technology. *Personal notes available on Researchgate*, February 2023.

[6] C.E. Shannon and W. Weaver. *The Mathematical Theory of Communication.* University of Illinois Press, Urbana, 1949.

[7] M. Burgess. Notes on trust as a causal basis for social science, v0.2. *SSRN Archive, available at http://dx.doi.org/10.2139/ssrn.4252501*, August 2022.

[8] M. Aubé and A. Senteni. A foundation for commitments as resource management in multi-agents systems. In *European Workshop on Modelling Autonomous Agents in a Multi-Agent World*, pages 13–25, 1995.

[9] D. Kahneman. *Thinking, Fast and Slow.* Penguin, London, 2011.

[10] M. Burgess. Motion of the third kind (ii) notes on kinematics, dynamics, and relativity. (notes available on Researchgate), 2022.

[11] T.M. Cover and J.A. Thomas. *Elements of Information Theory.* (J.Wiley & Sons., New York), 1991.

[12] Cognicist Collective and Prophet Mind. The cognicist manifesto. Google document, 19 October 2017.

[13] K. Begnum, M. Burgess, T.M. Jonassen, and S. Fagernes. Summary of the stability of adaptive service level agreements. In *Proceedings of the 6th IEEE Workshop on Policies for Distributed Systems and Networks*, pages 111–114. IEEE Press, 2005.

[14] K. Begnum, M. Burgess, T.M. Jonassen, and S. Fagernes. On the stability of adaptive service level agreements. *eTransactions on Network and System Management*, 2(1):13–21, 2006.

[15] R. Axelrod. *The Complexity of Cooperation: Agent-based Models of Competition and Collaboration.* Princeton Studies in Complexity, Princeton, 1997.

[16] J.A. Bergstra and M. Düwell. Accusation theory. *Transmathematica*, Dec. 2021.

[17] S. Galam. *Sociophysics.* Springer, 2012.

[18] M. Jusup, P. Holme, K. Kanazawa, M. Takayasu, I. Romić, Z. Wang, S. Geček, T. Lipić, B. Podobnik, L. Wang, W. Luo, T. Klanjšček, J. Fan, S. Boccaletti, and M. Perc. Social physics. *Physics Reports*, 948:1–148, feb 2022.