The Knowledge Engineering Review, Vol. 31:2, 142–166. © Cambridge University Press, 2016
doi:10.1017/S0269888916000023

Classifying sanctions and designing a conceptual sanctioning process model for socio-technical systems

LUIS G. NARDIN[1,2], TINA BALKE-VISSER[3], NIRAV AJMERI[4], ANUP K. KALIA[4], JAIME S. SICHMAN[1] and MUNINDAR P. SINGH[4]

[1]*Computer Engineering Department, University of São Paulo, São Paulo, SP 05508-970, Brazil;*
e-mail: luis.nardin@usp.br, jaime.sichman@poli.usp.br;
[2]*Institute of Cognitive Sciences and Technologies, CNR, 00185 Rome, Italy;*
[3]*Centre for Research in Social Simulation, University of Surrey, Guildford GU2 7XH, UK;*
e-mail: t.balke@surrey.ac.uk;
[4]*Department of Computer Science, North Carolina State University, Raleigh, NC 27695-8206, USA;*
e-mail: najmeri@ncsu.edu, akkalia@ncsu.edu, m.singh@ieee.org

Abstract

We understand a socio-technical system (STS) as a cyber-physical system in which two or more autonomous parties interact via or about technical elements, including the parties' resources and actions. As information technology begins to pervade every corner of human life, STSs are becoming ever more common, and the challenge of *governing* STSs is becoming increasingly important. We advocate a normative basis for governance, wherein *norms* represent the standards of correct behaviour that each party in an STS expects from others. A major benefit of focussing on norms is that they provide a socially realistic view of interaction among autonomous parties that abstracts low-level implementation details. Overlaid on norms is the notion of a *sanction* as a negative or positive reaction to potentially any violation of or compliance with an expectation. Although norms have been well studied as regards governance for STSs, sanctions have not. Our understanding and usage of norms is inadequate for the purposes of governance unless we incorporate a comprehensive representation of sanctions.

We address the aforementioned gap by proposing (i) a sanction typology that reflects the relevant features of sanctions, and (ii) a conceptual sanctioning process model providing a functional structure for sanctioning in STS. We demonstrate our contributions via a motivating scenario from the domain of renewable energy trading.

1 Introduction

We understand a socio-technical system (STS) as a cyber-physical system that incorporates the interactions of multiple autonomous participants whose interests are at best imperfectly aligned (Singh, 2013). Traditional examples of STSs include the Internet as a whole, the global financial system, telecommunication networks, next-generation power grids, environmental systems, and regional and global transportation systems. Our conception additionally supports viewing even small and potentially transient systems, such as scientific collaborations, data sharing systems, and neighbourhood power cooperatives, as STSs. Thus, STSs represent a perspective on systems that take into account the social and technical aspects together (Houwing *et al.*, 2006; Fiadeiro, 2008). In particular, STSs consider the social interactions among the autonomous participants, and the technical interactions between these participants and the relevant technical elements.

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press

The success of an STS relies upon effective *governance*, which pertains to how the aforementioned interactions are controlled, especially with a view to achieving relevant participant objectives, both technical (e.g. performance) and social (e.g. fairness of access to common resources) (Balke & Villatoro, 2012).

Governance is achieved by norms established among the participants and sanctioning occurring with respect to such norms (Singh *et al.*, 2013). A norm in our conception captures an expectation of one party, Alice, that another party, Bob, will behave in a certain manner. An example of this would be: Alice expects Bob to conserve power by switching off the office space heater when leaving the office. In essence, Alice holds Bob accountable for the given norm. Even if the participants in an STS are peers, in general, they play different roles with distinct privileges and liabilities, expressed via distinct norms that apply to those roles (Singh, 2013).

A participant can potentially (1) *comply* with a norm by behaving as expected (e.g. turning the heater off), or (2) *violate* a norm by failing to behave as expected (e.g. leaving the heater on when leaving the office).

We define a *sanction* as a reaction triggered due to the violation of or the compliance with a norm, whose primary aim is promoting norm compliance. A sanction can be negative or positive. Thus, sanctioning provides a foundation for how participants in an STS may seek to influence each other's decision making and steer the STS towards their preferred direction. Although norms have been studied in regards to governance for STSs (Fiadeiro, 2008; Weigand, 2009; Jones *et al.*, 2013; Singh, 2013), sanctions have largely been neglected.

Etymologically, the term *sanction* has its origins in two roots, the Latin words *sanctionem* and *sanctus*, that date back to the 14th and 15th centuries, respectively. The former means the 'act of decreeing', and the latter, which *sanctionem* apparently derives from, means 'to decree, confirm, ratify, or make sacred' (Harper, 2010). More recently, however, the term sanction has also assumed a different connotation, that is, the imposition of a penalty for disobeying a norm or granting a reward in case of complying with a norm. The *American Heritage Dictionary* (Picket, 2011) recognises the following meanings of the term sanction:

(i) to give official authorisation or approval to, as and when a legislature sanctions a presidential action;
(ii) to encourage or tolerate by indicating approval;
(iii) to penalise, as for violating a moral principle or international law.

These meanings are reflected in the literature on sanctions with the computing literature emphasising (iii).

This paper develops a sanction typology that incorporates insights gleaned from diverse disciplines—not surprisingly, there is variation in the definition of sanctions not only across disciplines, but also within them. In addition, this paper introduces a conceptual sanctioning process model that provides a functional structure for sanctioning in STS.

Section 2 introduces a motivating scenario with several situations to which sanctions may apply. Sections 3 and 4, respectively, survey the conceptions of sanctions in multi-agent systems (MAS), and law and social sciences. Based on these, Section 5 describes a comprehensive sanction typology. Section 6 introduces a conceptual sanctioning process model that includes the main capabilities involved in enforcing norms in an STS. Section 7 demonstrates our sanction typology and conceptual sanctioning process model over our motivation scenario. Section 8 concludes with a summary of our findings and some ideas for future research.

2 Motivating scenario: renewable energy trading

To demonstrate our ideas, we consider a scenario based on the *smart grid*, understood as an electrical power grid that supports bi-directional flows of electricity and information between all network nodes, such as power plants and appliances. The smart grid enables real-time market transactions and it interfaces the interaction between people, buildings, industrial plants, generation facilities, and the electrical network (Vu *et al.*, 1997; Department of Energy, 2003).

This scenario is partially inspired by the Power Trading Agent Competition (PowerTAC, 2010), which is a competitive simulation that models transactions among the members of a power grid. PowerTAC serves as an STS because it involves multiple self-interested stakeholders collaborating with respect to

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press

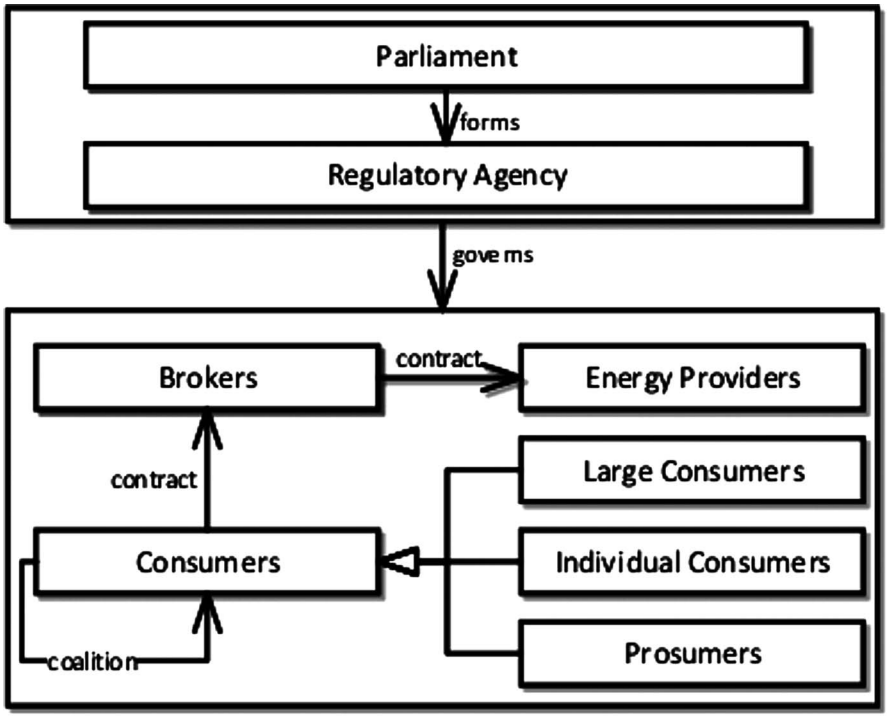


Figure 1 Power grid system motivating scenario

their computational and physical resources. Mah *et al.* (2012) discussed the governance challenges arising in power grids.

Figure 1 shows the main entities in our motivating scenario.

An *energy provider* generates (a large amount of) energy with high stability. A *consumer* may be one of the following kinds: (i) *large consumer* (e.g. a factory or an amusement park that consumes a large amount of energy); (ii) *individual consumer* (e.g. a house or a small office that consumes a small amount of energy); (iii) *prosumer* (e.g. a house with solar panels or a farm with wind generators that generates and consumes small amounts of energy, and whose generation is quite unpredictable, particularly due to the vagaries of the weather). In certain occasions, two or more consumers may form a *coalition*, thus acting as one single consumer, to buy or to sell energy.

A *broker* mediates energy transactions between energy providers or prosumers, and consumers. The *regulatory agency* is a distinguished authority that promulgates and enforces norms on the dealings between providers, consumers, and brokers. The *parliament* is the entity that constitutes the regulatory agency.

The regulatory agency formally governs the interactions among energy providers, brokers, and consumers. In addition, these entities can monitor each others' behaviour with respect to the established norms, and sanction each other.

For concreteness, consider three neighbours (John, Joseph, and Mary) connected to the same power network, each of whose monthly energy consumption is around 1000 kWh. Each of them installed solar panels with a capacity of around 400 kWh per month, characterising them as prosumers. They entered into separate energy buying contracts with a broker, which in turn has a buying contract with an energy provider. The broker may also buy renewable energy generated by prosumers at a price of $0.05 per kWh for a minimum of 1000 kWh per month, or at $0.02 per kWh otherwise. The broker has a selling contract with a factory (large consumer). We refer to John, Joseph, Mary, and the factory jointly as the broker's *consumers*.

The norms ruling this scenario establish that (i) the seller is expected to (uninterruptedly) supply the committed amount of energy to the buyer; (ii) a coalition member is expected to (uninterruptedly) supply the amount of energy agreed with the coalition; and (iii) the buyer is expected to pay for the amount of energy supplied by the seller.

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press

Based on such smart grid scenario, consider the following possible situations in which sanctions may apply:

Situation 1: energy provider failure

Due to a human error, the energy provider fails to fulfil its commitment of supplying energy without interruption to its consumers, which in turn causes the brokers that negotiated the energy supply to also fail to fulfil their commitments to their consumers. Unsatisfied with the service provided, consumers may decide to do nothing as the failure did not last long, or to take one or more of the following actions:

S1.1: blame themselves for selecting the service from this broker;
S1.2: take legal action against the broker;
S1.3: disseminate negative evaluations about the broker; or
S1.4: switch to another broker.

Subsequently, the broker may sanction the energy provider as its credibility and finances suffer due to the energy provider's fault. An option would be to simply switch to another provider; however, switching is impossible due to the fact that this energy provider is the only energy provider in the region capable of supplying the required amount of energy. The broker's choices, therefore, are limited to reactions stipulated in its contract with the provider. Thus, the broker may decide to sue the energy provider (*S1.5*).

Furthermore, the regulatory agency observing the consumers who did not receive adequate power decides to evaluate the broker and energy provider's liabilities and responsibilities in order to determine the sanctions to impose on them. The possible sanctions are:

S1.6: fine the energy provider between 1 and 5% of its monthly profit; or
S1.7: suspend the broker from signing new contracts for a period of up to 30 days.

Situation 2: coalition formation

John, Joseph, and Mary decide to take a vacation at the same time. They know that they can sell their spare energy to their broker. Joseph, however, realises that their broker buys renewable energy at a higher price from prosumers who can generate more than 1000 kWh per month. He suggests they form a coalition to which they would each contribute at least 350 kWh for 1 month. John and Mary agree with his proposal. As they would profit from Joseph's initiative, they may react by:

S2.1: thanking Joseph; or
S2.2: disseminating Joseph's good reputation due to his initiative.

Situation 3: coalition failure

Upon returning from their vacation, they notice Mary's solar panel malfunctioned because she did not follow the manufacturer's service recommendations. As their coalition failed to generate energy exceeding 1000 kWh, they received only a reduced price from the broker, as specified in their contract. John and Joseph may decide to do nothing as they understand that hardware failures are difficult to anticipate and Mary has a history of being conscientious, or they (and Mary) may react according to one or more of the ways:

S3.1: Mary blames herself for the solar panel malfunctioning;
S3.2: John and Joseph suggest that Mary have her solar panel serviced on a regular basis;
S3.3: John and Joseph reduce their trust in Mary as a partner;
S3.4: John and Joseph request compensation from Mary; or
S3.5: John and Joseph tell others that Mary is an unreliable partner.

Situation 4: coalition success

During next year's vacation, John, Joseph, and Mary again form a coalition to sell energy to the same broker. However, because of unforeseen circumstances (John's mother suffered a heart attack), John cancels his vacation and returns home accompanied with his mother, who requires special care and

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press

equipment that consumes a lot of energy. Conscious that he will not be able to supply the amount of energy he committed to, John requests his friend George to replace him in the coalition. George agrees to John's request, and Joseph, Mary, and George together generate more than 1000 kWh of energy, thus meeting their contracted threshold for receiving the higher price. Hence, Joseph and Mary may react by:

S4.1: thanking George for coming to their rescue;
S4.2: praising George to others;
S4.3: praising John to others as he had proposed a successful alternative to handle his contingency; or
S4.4: deciding not to form a coalition with John in the future, even though they recognise that John's behaviour was justified.

Situation 5: broker failure

In order to meet unanticipated market demands, a factory decides to operate an additional shift. Thus, it requests additional energy from the broker; the broker agrees to provide additional energy at a higher price. As the energy supplied by the energy provider is limited, the broker redirects the energy supplied to John, Joseph, and Mary to the factory. Unhappy with the service provided by the broker, the latter may react similarly to the options listed in Situation 1 (S1.1–S1.4). In contrast, the large consumer on obtaining the increased energy supply may:

S5.1: increase its trust in the broker as a service provider; or
S5.2: tell others about the willingness of the broker to meet increased demand.

The main features that the situations in the foregoing scenario bring out are as follows:

1. *Sanctions are loosely coupled to norms with multiple categories of sanctions being reasonable due to the violation of or compliance with a norm*. Situations 1 and 3 illustrate this feature as the affected parties (i.e. the parties affected by the norm violation or compliance) are not forced to apply a pre-established sanction to the violating party, if any. Yet, these situations describe a list of options available (i.e. loosely coupled to norms). In addition, the available sanctions are of different types, such as legal action, ostracism, or disseminating praise or criticism (i.e. availability of multiple categories of sanctions).
2. *A sanctioning party can consider a variety of factors in determining whether and which sanctions to apply*. Situation 3 illustrates this feature as John and Joseph take into account not only Mary's fault, but her history as an energy supplier (i.e. Mary's reputation) and what caused her to violate her commitment (i.e. hardware malfunction), in order to decide whether to sanction her or not. Deciding on applying a sanction, they may take into account the same factors to decide which of the available sanctions to apply.

These features reduce to the following three requirements of a sanctioning process for STSs:

R1: support for multiple categories of sanctions;
R2: potential association of multiple sanctions with a norm violation or compliance; and
R3: reasoning about the most adequate sanction to apply depending on different factors.

Implementing a sanctioning process that fulfils these requirements primarily demands a way of distinguishing sanctions according to their features. A typology of sanctions can help map out the space of possibilities, thereby enabling us to distinguish sanctions and to group them into categories (R1). These distinctions enable agents to link various sanctions as reactions to the same violation or compliance to a norm (R2), which enable them to evaluate the sanctions efficacy for each kind of violation or compliance. The efficacy information contributes to the reasoning about the most adequate sanction to apply (R3), and possibly help to improve the general level of compliance in the system.

In the next two sections, we review the existing literature on sanctions in MAS, and in law and social sciences, respectively, with a view of proposing a typology of sanctions and a conceptual sanctioning

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press

process model that supports scenarios of the above kind and meet the identified three requirements of a sanctioning process for STSs.

3 Sanctions in multi-agent systems

In MAS, sanctions are a form of *social control* used for achieving *social order* (Castelfranchi, 2000), which is akin to our notion of governance for STS. Social control and order occur in MAS via two main complementary approaches, namely, trust and reputation, and normative systems, each of which we discuss next.

3.1 Trust and reputation

Trust and reputation are a means to discourage unwanted behaviours and to foster desired ones among agents in MAS. Because trust functions as a decision criterion for an agent to engage in social activities, any action that potentially affects the trust placed in a party can serve as a sanction on that party. Trust and reputation reflect the idea of indirect sanctioning in which agents instead of acting directly against others (e.g. imposition of fines), use information about the past behaviour of others to evaluate how they might perform in the future and decide whether to interact with them. A positive performance history thereby would ordinarily lead to higher trust that the agent will perform well in the future, whereas a negative history results in the opposite. Thus, a sanction would arise indirectly via future actions.

Dellarocas (2006) recognises two functions for reputation: (i) the *sanctioning* role in which reputation is used for deterring moral hazards present in agreement settings in which each party may gain from acting in an antisocial manner (e.g. the eBay reputation mechanism that enables the evaluation of the seller's features promoting honest trade); and the *signalling* role in which reputation is used for reducing information asymmetries among interacting parties (e.g. the TripAdvisor rating system that makes visible the quality of the offered products and services indicating what to expect).

Due to the importance of trust and reputation for MAS (Castelfranchi & Falcone, 1998), several models have been proposed in the literature in the last two decades. The following non-exhaustive list is representative of trust and reputation models available in the MAS literature: Histos and Sporas (Zacharia & Maes, 2000), Mui, Mohtashemi and Halberstadt (Mui *et al.*, 2002), ReGreT (Sabater-Mir & Sierra, 2002), Repage (Conte & Paolucci, 2002; Sabater-Mir *et al.*, 2006), FIRE (Huynh *et al.*, 2004), Wang & Singh (Wang & Singh, 2010), L.I.A.R. (Vercouter & Muller, 2010), and BDI + Repage (Pinyol *et al.*, 2012). Further information about computational trust and reputation models can be found in Pinyol and Sabater-Mir (2013) and Hendrikx *et al.* (2015).

Ways to model trust and reputation include quantitative, for example, Wang and Singh (2010), and cognitive, for example, Conte and Paolucci (2002), approaches. The latter helps to distinguish an agent A's *image* (i.e. beliefs another individual has about A) from its *reputation* (i.e. beliefs others collectively have about A). Thus, image is personalised, while reputation is an impersonal evaluation produced by sharing information about agent A.

Image refers to the idea that the agent reacts to directly acquired beliefs when judging potential future interactions. Thus, in case of repeated interactions, gained beliefs can be used to identify agents that out-performed or under-performed, and respectively favour or disfavour their selection as a transaction partner. As a result, for example, when cheating another agent in one transaction, the cheater should consider the possibility that doing so might result in a negative image held by those cheated, thereby hurting future prospects for transacting. The corresponding sanction is indirect and delayed.

Rodrigues and Luck (2007) propose a model for building others' image based on Piaget's theory of exchange values (Piaget, 1995). Exchange values represent the gains and losses of agents in each direct interaction with others. These direct experiences are evaluated in terms of successful and unsuccessful interactions. The successfulness of an interaction is defined in terms of the balance between gains and losses: a successful interaction represents a situation in which the gains are equivalent or greater than the losses, and an unsuccessful interaction the opposite.

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press

In Kalia *et al.* (2014), image about others is learned based on a probabilistic trust model that estimates agents' trust parameters from positive, negative, and neutral interactions governed by commitments (i.e. a social relationship between two agents giving a high-level description of what one agent expects from the other).

Reputation presumes information sharing, but otherwise functions somewhat like image. Reputation is a general evaluation about a target, especially the target's ability to perform specific tasks, shared across some population. In contrast to image, in which agents act upon their own experiences, reputation requires the sharing of information. Such sharing can lead to a larger set of agents acquiring an evaluation about a target. Similar to image, reputation is an indirect sanction, but due to the inherent sharing involved, it takes the form of social control. It is worth noting that the underlying assumption of information sharing renders reputation mechanisms vulnerable to the lack of innate incentives for rational agents to report reliable and trustworthy information. Heitz *et al.* (2010) analyse various incentive mechanisms and identify that feedback reporting would be improved by rewarding those who share information. Strategies to overcome the effect of dishonest reports include (i) calculating reputation based on different information, and (ii) normalising the reported information based on the recommender's trustworthiness and (iii) behavioural stability.

Besides malicious or inaccurate reports, reputation may still not be an accurate predictor of an agent's future behaviour if the interactions' context are not taken into account. In Pinyol and Sabater-Mir (2013), for instance, this context dependence refers to the granularity of reputation information. Miles and Griffiths (2015) propose a reputation assessment method that uses past interaction context to determine its relevance to evaluate a reputation.

3.2 Normative multi-agent systems

The field of normative MAS (nMAS) reflects the idea of normative action (Habermas, 1984), which considers agents as members of a group with an expectation that they respect the norms of that group. Because agents are autonomous and pursue their own goals, norm internalisation is one possible explanation of why agents comply with norms even in situations they would be better-off violating them. A norm is internalised whenever its maintenance has become independent of external support events, such as reinforcement through sanctions (Arionfreed, 1968).

Andrighetto *et al.* (2010) characterise norm internalisation as a multi-step cognitive process that leads from externally enforced norms to norm-corresponding agent goals. This process is dependent on *norm salience*, that is, the level of importance an agent assumes a norm has within its social group in a given context. The more salient the norm is, the more it is internalised, and vice versa. Norm salience varies depending on individual and social factors, such as others actions and reactions intending to promote compliance with the norms.

Norm enforcement mechanisms thus play an important function in norm internalisation and in directing the system towards an expected path via the process of *enforcement*. Sanctions are an important mechanism for enforcing norms.

Balke and Villatoro (2012) propose a sanctioning process model composed of four phases: violation detection, sanctioning determination, sanctioning application, and assimilation. Each phase distinguishes the roles of the agents involved, in which Balke and Villatoro classify the sanctioning approaches and analyse popular nMAS frameworks with respect to their sanctioning ideas.

As pointed out by Balke (2009) with respect to sanctioning, the nMAS literature builds on traditional areas such as cognitive science, economics, and sociology. Importantly, some nMASs rely upon an enforcement mechanism that assumes that actors can be controlled and non-compliant actions can be prevented, that is, a violation is not possible. Jones and Sergot (1993) term such a mechanism *regimentation*, as do Grossi *et al.* (2007); others call it control-based enforcement (Pinninck, 2010: 14). Minsky (1991) distinguishes two modes of regimentation, namely, by interception (i.e. controlling the messages an agent is able to send), and by compilation (i.e. controlling the mental states of an agent). These mechanisms violate the autonomy and heterogeneity of agents, respectively (Singh & Huhns, 2005).

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press

Jones and Sergot (1993) term the complementary mechanism *regulation* wherein violations may occur, yet whenever a violation is detected, reactions (i.e. sanctions) may be applied to the violator. Others term this mechanism *incentive-based enforcement* (Pinninck, 2010: 16).

Pasquier *et al.* (2005) propose a sanction typology along three dimensions: (1) *direction*, which specifies the content of a sanction: *positive* or *negative*, respectively, representing rewards or penalties; (2) *type*, which specifies the nature of a sanction: *automatic*, *material*, *social*, or *psychological*; and (3) *style*, which specifies the target agent's awareness of the application of the sanction, and may be *implicit* or *explicit*. Pasquier *et al*. bring up the important point of it being an agent's decision about whether and which sanction to apply.

Cardoso and Oliveira (2009, 2011) synthesise Pasquier *et al.*'s dimensions into two broad categories of sanctions: (i) *direct or material*, which have an immediate effect on the (material) resources of a target agent, for example, by imposing fines; and (ii) *indirect or social*, which may have a future effect on the agents' interactions, for example, by changing the agent's reputation. In the first work, they propose a centralised norm enforcement mechanism for commitments, that is, agreements binding two or more parties describing their mutual expectations, to the degree that to renege on the commitments is costly. The mechanism uses only direct material sanctions implemented through fines as a deterrent. The main idea behind Cardoso and Oliveira's sanctioning mechanism is to base the severity of fines on statistics regarding violation: the severity of a fine is increased or decreased depending upon whether the number of violations is, respectively, greater or smaller than a specified threshold. Their approach relies upon a centralised entity that tracks commitments among agents and evaluates their violation or compliance. In effect, the centralised entity restricts agents' autonomy by determining sanctions and their severity, and imposes them without regard to any subjective or contextual distinction.

Centeno *et al.*'s (2011, 2013) mechanism resembles Cardoso and Oliveira's approach, but accommodates contextual information (e.g. the state of the system) to adapt sanctions to particular agents and situations. In particular, the mechanism identifies the appropriate actions an agent should perform, given the current state of the system, and applies sanctions to induce the agent to perform such actions. As in electronic institutions (Esteva *et al.*, 2000, 2001), each external agent is associated with an institutional component for sanctioning, which adapts policies to promote norm compliance by agents. Similarly, Campos *et al.* (2013) propose an adaptation mechanism that modifies norm violation penalties according to agents' behaviours through the use of case-based reasoning (Aamodt & Plaza, 1994).

The foregoing mechanisms, albeit adaptable, require *a priori* knowledge not only about the global utility function, but also about whether the system is gaining or losing utility. The need for a global utility function renders these approaches non-viable for STSs.

Relaxing the centralised monitoring characteristic of the previous architectures, Daskalopulu *et al.* (2002) introduce an architecture of contract performance monitoring with arbitration. Contractual party agents hold a state diagram representation of the contract in terms of obligations. Whenever they disagree about the obligation fulfilment, they present evidence supporting their view of what happened to an arbitrator agent, which undertakes to produce a resolution to the conflict. The arbitrator reasons about the evidence using Subjective Logic (Jøsang, 2001) and proposes a solution resetting the agreement to its normal course. If there is no solution, the agreement is terminated and litigation may ensue to establish liability and award damages.

Modgil *et al.* (2009) propose a general architecture for norm-governed systems that relies upon infrastructure agents for monitoring and sanctioning. The architecture comprises observer agents responsible for inspecting specific actions of agents and determining whether a norm violation has happened (Faci *et al.*, 2008). If so, they report any violated norms to manager agents, who apply pre-specified sanctions to the violators.

Criado *et al.* (2013) propose MaNEA, an architecture for enforcing norms, in which enforcer infrastructure agents monitor and sanction (i.e. punish or reward) application agents due to a norm violation or compliance. Importantly, each norm is associated with specific penalty or reward sanctions. Hence, the norm enforcers are not autonomous but they are forced to act as specified and cannot select the most appropriate sanction for a given situation.

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press

To overcome limitations of centralised and infrastructural approaches, some works support second-party and third-party sanctioning, in which an agent who is affected by or observes a violation is responsible for identifying and sanctioning the violating agent, respectively. Pinninck *et al.* (2010) propose a distributed mechanism in which non-compliant agents can be ostracised from the society. In Pinninck *et al.*'s approach, agents monitor and disseminate information about each other as a way to build a reputation measure, which is used in the decision process to ostracise recurrent non-compliers (i.e. non-reputable agents).

López and Luck (2003) introduce a distributed norm enforcement mechanism in which the violation of or compliance with a norm results in triggering an *enforcement norm*. The enforcement norm specifies the penalty or reward to apply due to, respectively, the violation of or the compliance with the norm, as well as the application criteria and the role of the agent responsible for applying it. Despite enabling the agents to monitor and to sanction against other agents, this mechanism pre-establishes the sanctions to be applied.

In contrast, adaptive sanctioning techniques enable agents to dynamically adapt the strength of a sanction. Whereas Villatoro *et al.*'s (2011) technique adapts the strength of the sanction based on the number of defectors, Mahmoud *et al.*'s (2012a) technique adapts it according to characteristics of the violation, such as magnitude and frequency. Mahmoud *et al.* (2012b) identify that, due to lack of information, these previous adaptive techniques fail to stop agents from violating norms in partially observable environments. Hence, they introduce reputation as a means to enrich agents' knowledge about others and to adapt the strength of a sanction. The drawback of these techniques is their limitation to the use of a specific type of sanction, namely material sanction.

Giardini *et al.* (2010) propose a cognitive model with distinct kinds of sanctioning behaviours. Andrighetto and Villatoro (2011) create a mechanism that takes into account Giardini *et al.*'s cognitive model to evaluate two distinct enforcing strategies, namely *Punishment* and *Sanction*. In the Punishment strategy, a sanction corresponds only to the imposition of a cost on the target (i.e. material sanction). In addition to imposing economic costs, the Sanction strategy has a norm-signalling component that influences the target by signalling about the existence of the norm, thereby indicating that it should be respected. Andrighetto and Villatoro show that the Sanction strategy is more effective in promoting compliance with the norm because in addition to inflicting a cost on the violator, it signals that the norm is relevant to other members of the social group.

Another set of works addresses the evaluation of different sanctioning strategies, for example, the performance of punishment and reputation in the public goods game (Helbing *et al.*, 2010; Giardini *et al.*, 2014). Giardini *et al.* compare the effectiveness of two sanctioning mechanisms, punishment and reputation in isolation, in reducing defection in the public goods game. Helbing *et al.* analyse various strategies and the levels of punishment that improve cooperation in the public goods game. In these works, however, the agents cannot autonomously choose the sanction they deem more appropriate to the situation.

3.3 Remarks

The proposals presented in Sections 3.1 and 3.2 suffer from drawbacks that render them unsuitable for supporting the requirements for STSs identified in Section 2.

Even though they involve multiple categories of sanctions (R1), such as reputation, ostracism, and material sanction, each approach uses a single category, established at design time. For instance, Pinninck *et al.*'s (2010) approach uses only reputation, and Pasquier *et al.* (2005) and Cardoso and Oliveira's (2011) approaches use only material sanctions. Hence, the approaches do not consider multiple categories of sanctions simultaneously (thus failing R2) and do not support the enforcer's decision making (thus failing R3). López and Luck's (2003) and Criado *et al.*'s (2013) mechanisms can support multiple categories of sanctions (R1). However, they model sanctioning as an automatic reaction, which limits agents' decision making and disregards context (thus failing R2 and R3). Villatoro *et al.*'s (2011) and Mahmoud *et al.*'s (2012a) approaches enable agents to adjust their sanctions (thus satisfying R3), but are limited to material sanctions (thus failing R1 and R2). Even Mahmoud *et al.* (2012) apply reputation only as a means to adjust the material sanction (i.e. as extra information) and not as a sanctioning mechanism by itself.

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press

In summary, existing MAS approaches to social control do not conveniently address a situation from our motivating scenario: waiving a sanction, in which the affected coalition members may decide not to sanction the violating agent (possible outcome of Situations 1 and 3) even though there is a set of possible sanctions linked to the norm violation.

Mechanism design (Hurwicz & Reiter, 2008), another MAS technique, may be considered a successful alternative way to build governance in STSs. The idea is that by designing a system in a certain way, one can encourage the participants to behave well, as respecting the STS requirements is beneficial to them. Hence, the particular design encourages truthfulness without the need of reputation or specific norms. Some authors, for example, Broersen *et al.* (2013), even state that normative systems are comparable to mechanism design. We consider that the implicit rules followed by the agents subject to a mechanism may be seen as equivalent to a set of norms inducing the agents to behave in a certain way; the difference is that the latter are explicit and the former do not take the notion of sanctions into account: the fact that the agent does not behave accordingly will have an effect exclusively on it (i.e. it will not gain a higher utility).

4 Sanctions in law and social sciences

Recognising the inadequacy of existing MAS mechanisms for completely modelling STSs, we turn to law and social sciences to develop a deeper understanding of sanctions.

4.1 Law

Legal positivism is nowadays dominant among the various legal theories (Patterson, 2010). It assumes that the existence and content of law depends on social facts and not on considerations such as morality (Gardner, 2001; Patterson, 2010: Ch. 14). In this tradition, law is an instrument of social order, but one that emanates from the state and is enforced through legal sanctions by recognised law enforcement institutions. Legal sanctions are reactions that seek to induce individuals to comply with legal rules (Garner, 2010). Generally, sanctions may be negative or positive, yet the law frequently considers negative sanctions as the only means to enforce obedience (Schwartz & Orleans, 1967). An example of a positive legal sanction is the Earned Income Credit in the US tax code that allows eligible taxpayers to subtract the amount of the credit from the total they owe the state. These credits are granted, among other reasons, to encourage certain individuals' behaviours, such as investing in renewable energy production (Energy Policy Act, 1992).

Some legal theorists, for example, Ellickson (1991), Posner and Rasmusen (1999), Posner (2000), Meares *et al.* (2004), oppose the interpretation of sanctions as enforced only by the state. They argue that non-legal forms of regulation (i.e. those applied by peers), such as gossip, disapproval, and ostracism, remain important. Although not in line with our definition of sanctions, yet highlighting the importance of non-legal sanctions, Posner and Rasmusen (1999) identify six types of sanctions for violating norms: (i) *Automatic*—the sanction is the direct consequence of the violator's action; (ii) *Guilt*—the violator feels bad about knowing that he has behaved in an inappropriate way, without others coming to know about it; (iii) *Shame*—the violator feels bad because he perceives his action has reduced the others' evaluation about himself; (iv) *Informational*—the violator unintentionally provides information about himself that he would like others not to know; (v) *Bilateral costly*—punishment inflicted on the violator by a second-party or third-party; (vi) *Multilateral costly*—punishment inflicted on the violator by a second-party or multiple third-parties.

The question *What justifies the application of sanctions against people?* is far from settled. According to Hart (1968: 1–27), an answer should address three distinct concerns: (i) What justifies the creation and maintenance of a sanctioning system? (ii) Who may be sanctioned? (iii) How should the appropriate sanction be determined? Existing theories differ in how they address these three concerns (Davis, 2009). The *consequentialist* theory justifies sanctioning by reference to its consequences, which is the discouragement of future misbehaviour. A form of consequentialism is *utilitarianism*, which views sanctioning as a cost-effective means to prevent future misbehaviours (Beccaria & Ingraham, 1819;

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press

Bentham, 1823; Mill, 1871). Commonly supported consequentialist mechanisms include the following (Cavadino & Dignan, 2002: Ch. 2):

- *Deterrence* involves causing fear among potential violators. It can be *individual*, applying to an individual, or *general*, applying to anyone who observes a violator being sanctioned (Nagin, 1998). Respective examples are (i) an energy broker being levied a fine for violating a commitment to provide a certain amount of energy, presumably leading it to create internal controls to avoid such violations, and (ii) brokers who observe another broker being penalised may develop controls to avoid such violations themselves.
- *Incapacitation* prevents future misbehaviour temporarily or permanently. For example, imprisonment incapacitates would be perpetrators by restricting their movements. In our scenario, a broker's trading account may be temporarily suspended.
- *Reform* improves a violator's character or behaviour to make it less likely to violate the norm in the future. For instance, requiring an energy provider to train its employees better would reduce the risk of future power interruptions.

In contrast to the consequentialist theory, the *retributive* theory (Kant, 1999) seeks to sanction an offender proportionally to the magnitude of his misbehaviour, and does not consider the possible future consequences of the sanctioning. Thus, in case of an energy blackout caused by an energy provider, the penalty would be calculated based on the aggregate damage that such an interruption of energy caused to the affected consumers.

4.2 Sociology

Radcliffe-Brown (1934) may have been the first sociologist to define sanctions. He defines them as a society's (or a 'considerable number' of its members) reaction to an approved or disapproved behaviour. Gibbs (1966), however, states that not all reactions to a behaviour count as sanctions and defines a set of criteria under which it counts as such. A sanction (i) requires a *referential*, typically a social norm; (ii) is applied by at least one *enforcer*; (iii) is associated with a *prescription*; (iv) specifies its *enforcer*'s capability; and (v) specifies whether it is to be *perceived* to be a sanction by its target.

In general, sanctions are used to ensure the compliance of individuals to desirable social norms, that is, prescribed behaviours shared and enforced by a community (Bicchieri, 2006). Sanctions therefore include not only legal penalties, but also informal rewards such as esteem from community members.

Radcliffe-Brown (1934) proposed an early classification of sanctions. A sanction may be *positive* or *negative*. It may also be *diffuse* (i.e. individual action) or *organised* (i.e. applied according to a social tradition and recognised procedure). For example, a legal sanction would be *negative* and *organised* as it is enforced by a recognised authority.

Morris (1956) proposes a classification of sanctions that includes six dimensions: *reward–punishment* ('more reward than punishment' to 'more punishment than reward'), *severity* ('light, unimportant' to 'heavy, important'), *enforcing agency* ('specialised, designated responsibility' to 'general, universal responsibility'), *extent of enforcement* ('lax, intermittent' to 'rigorous, uniform'), *source of authority* ('rational, expedient, instrumental' to 'divine, inherent, absolute, autonomous'), and *degree of internalisation* ('little, external enforcement, required' to 'great, self-enforcement, sufficient').

Gibbs (1966) proposes an alternative classification of sanctions based on four dimensions:

- *Type*: whether *internal* or *external* with respect to the individual who enforces it (Mill, 1871: Ch. 3). An *internal* sanction comes from the individual's own mind, and involves feelings resulting from personal morals, and whether or not the individual regrets a prior action. An *external* sanction reflects disapproval from others, such as peers or governmental institutions (i.e. police and judiciary).
- *Direction*: a *positive* sanction is a reward granted for compliance with a norm; a *negative* sanction is a punishment inflicted because of the violation of a norm.
- *Source*: a *formal* sanction is applied by a recognised social institution and an *informal* sanction by a peer.

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press

- *Effect*: a *preventive* sanction has the purpose of influencing behaviour to promote compliance or to prevent violation. The *inducement* of individuals to comply is a form of a preventive sanction. A *deterrent* is a sanction applied before compliance or violation. Examples are sanctions based on the *hedonic* conception, which involve physical or moral pain, or positive stimulation.

More recently, Clinard and Meier (2008) propose a simpler classification of sanctions based on two dimensions. *Direction* can be *positive* or *negative*. *Source* can be *informal* or unofficial, *formal* or official.

Although sociology emphasises informal sanctions, it recognises the need for multiple forms of sanctions to coexist for effective social control, and that institutionalised sanctions can be more effective for social control than informal ones (Meier, 1982; Miethe & Lu, 2005).

4.3 Psychology

Psychology sees sanctioning as essential for the maintenance of social life (Carlsmith, 2006). Indeed, sanctions are studied in psychology from the perspectives both of the sanctioner and the sanctionee. Regarding sanctioners, the primary psychological approach emphasises understanding individuals' motivations and justifications for punishing (Carlsmith *et al.*, 2002; Gabriel & Oswald, 2007; Petersen *et al.*, 2012). Regarding sanctionees, it involves modifying an individual's behaviour as a consequence of the sanction, either a punishment or a reward (e.g. Skinner's (1938) operand conditioning).

Recalling the distinction between *deterrence* and *retribution* (see Section 4.1), Carlsmith (2006) conducted experiments from which he concluded that individuals' sentencing decisions are affected primarily by retribution, even though they express preferences for utilitarian goals (deterrence) when legislating. That is, individuals relate sanctions and their severity to the harm they perceive from a violation: a more serious misbehaviour calling for a more severe sanction.

Extending the idea of proportionality, Petersen *et al.* (2012) argue that individuals base their decisions about the sanction and its severity on two factors: the seriousness of an offence and the offender's long-term value as an associate. These factors depend upon environmental cues, such as the offender's violation history, status (in-group or out-group), past contributions, expression of remorse, and kinship with the individual judging. According to experimental results, an individual's decision on whether to sanction depends upon the offender's value to them and not only on the seriousness of the offence. In contrast, the seriousness of the offence determines the intensity of the sanction applied. Therefore, an individual may apply a rehabilitative sanction to an offender when the former perceives the latter to hold some social worth.

4.4 Remarks

Law and the social sciences in general recognise the need for multiple categories of sanctions for maintaining social order. In particular, *informal* and *formal* sanctions coexist in human societies, as demonstrated by the situations of Section 2 and emphasised in sociology (Section 4.2) and psychology (Section 4.3).

Psychological studies show that humans usually reason about multiple factors before reacting to a violation. Interestingly, people reason differently depending upon whether they are creating legislation (promote deterrence, anticipating a potential violation) or reacting to a violation (engage in retribution). An individual would benefit from knowing about applicable sanctions, their usual consequences, and how others sanction in similar situations.

Because STSs involve humans, it makes sense that enforcement mechanisms applied to an STS inherit characteristics observed in pure human systems. The main characteristic we observed in the social sciences literature, that is, fields that study human systems, was a greater flexibility in the decisions to sanction. In addition, this greater flexibility also corroborates the requirements exposed by the motivating scenario in Section 2, in which multiple sanctions are available and multiple sanctioning decision factors influence the sanctioning decision.

Thus, advantages in using more flexible sanctioning mechanisms in STSs reside in the facts that (i) humans are used to a variety of sanctions and would find different types of sanctions natural in

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press

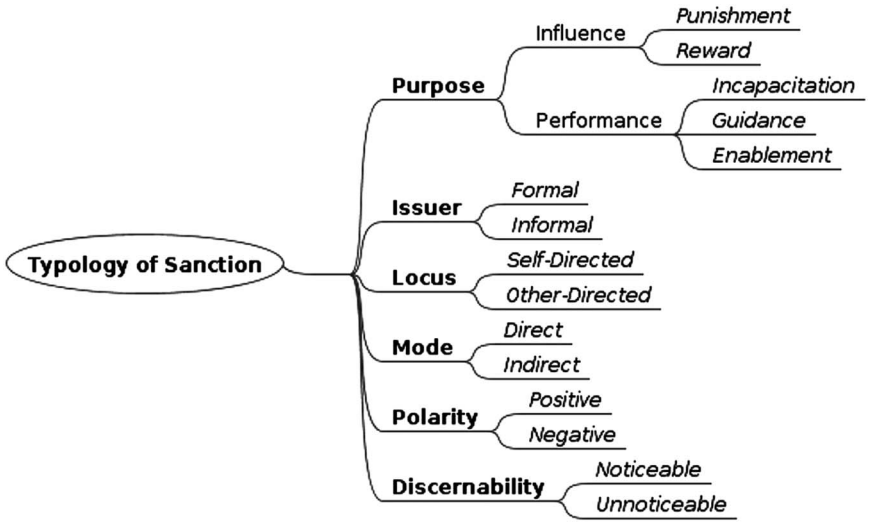


Figure 2 Dimensions of the proposed sanction typology

different circumstances, and (ii) sanctions of differing difficulty or costs may help achieve the same end result.

Existing enforcement mechanisms in nMAS do not support such flexibility, which motivates us to propose a general conceptual sanctioning process model in Section 6.

5 Sanction typology

The foregoing analysis of the literature illustrates the rich variety of concepts that come together in sanctioning. This situation leads us to propose a *typology*, that is, a systematic classification (Picket, 2011). A typology does not explain behaviour, but highlights distinctions that can feature in a theory as independent and dependent variables (Bailey, 1994).

Below, we describe a typology that lays the foundation for a comprehensive notion of sanctions as a possible means to prevent antisocial interactions (i.e. interactions in which one gains at others' expense, usually by breaking the norms) in an STS (Whitworth, 2006). Sanctions are applied to STSs to discourage interacting parties from taking advantage of others by reducing their possible gains in behaving antisocially. Details of how to discourage these kinds of behaviours are further discussed in this section and in Section 7.

We observed that existing sanction typologies (i) use distinct terms for the same concept; (ii) use the same term to describe distinct concepts; and (iii) incorporate disparate dimensions, which could be consolidated. Our proposed typology seeks to advance the understanding of sanctioning in STSs.

5.1 Dimensions

We now outline a sanction typology composed of six dimensions, as depicted in Figure 2, based on law and the social sciences literature but extended to accommodate STSs. These dimensions are *Purpose*, *Issuer*, *Locus*, *Mode*, *Polarity*, and *Discernability*.

We define the terms *source*, *target*, *sender*, and *receiver* used to describe some of these dimensions. Source and target are related to the content of the sanction. Source refers to the agent who generates the sanction (i.e. the affected agent or a third-party), and target indicates the agent whom the sanction is directed to. Sender and receiver refer to the agents that apply and receive the sanction, respectively. Thus, source and target relate to the content of the sanction, whereas sender and receiver relate to the individuals applying and processing the sanction.

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press

To illustrate the distinction between these terms, suppose a situation involving agents A, B, and C, wherein agent A sanctions agent C by informing agent B that agent C is not trustworthy. In this setting, agent A is the source and the sender of the sanction as it generates and applies the sanction, agent B is the receiver as it receives and processes the sanction, whereas agent C is the target due to the fact that it is the one the content of the sanction refers to.

5.1.1 Purpose

Purpose specifies the expected effect that the sanction is assumed to have on the social environment. Drawing from the literature on sanctions, we identify five possible purposes, organised into two aspects or regions of the dimension.

1. The *influence* aspect deals with incentives (negative or positive) and ranges over two purposes subject to violation or compliance by a target: *punishment* seeks to penalise the target to prevent future norm violations (e.g. the imposition of a fine upon the energy provider due to its failure to supply the contracted amount of energy (S1.6)); *reward* seeks to promote and motivate targets towards compliant behaviour (e.g. John and Mary thanking Joseph for his profitable coalition formation idea (S2.1); or, the factory informing others of the willingness of the broker to meet increased demand (S5.2)).
2. The *performance* aspect deals with capabilities and ranges over three purposes closely tied to the target's behaviour. First, incapacitation seeks to restrict the target's actions rendering norm violation impossible for a bounded period, differing from regimentation in which it is always impossible (e.g. suspension of the broker from signing new contracts for a period of up to 30 days (S1.7)). Second, *guidance* seeks to change a target's behaviour by instructing the target how to comply with the norm (e.g. John and Joseph suggesting that Mary have her solar panel serviced on a regular basis (S3.2)). Third, *enablement* seeks to provide an opportunity, and potentially the means, through which the target may comply with the norm (e.g. enable the broker to trade energy 24 hours a day without interruption instead of only 8 hours due to its good performance last year). Whereas enablement creates the conditions for repeating the sanctioned behaviour, reward provides an incentive for the target to repeat the sanctioned action.

5.1.2 Issuer

The *Issuer* specifies whether the sanction's issuer or enforcer is a recognised authority. *Formal* sanctions are established, and generally also enforced by recognised authorities, such as governmental institutions. Formal sanctions may be imposed not only by the state, but also by suitably empowered institutions, such as regulatory agencies (e.g. Federal Energy Regulatory Commission), or traders (e.g. eBay and Amazon). A specific example are the penalties specified in a trading contract in which an affected party may pay a reduced energy price due to a failure in the supply.

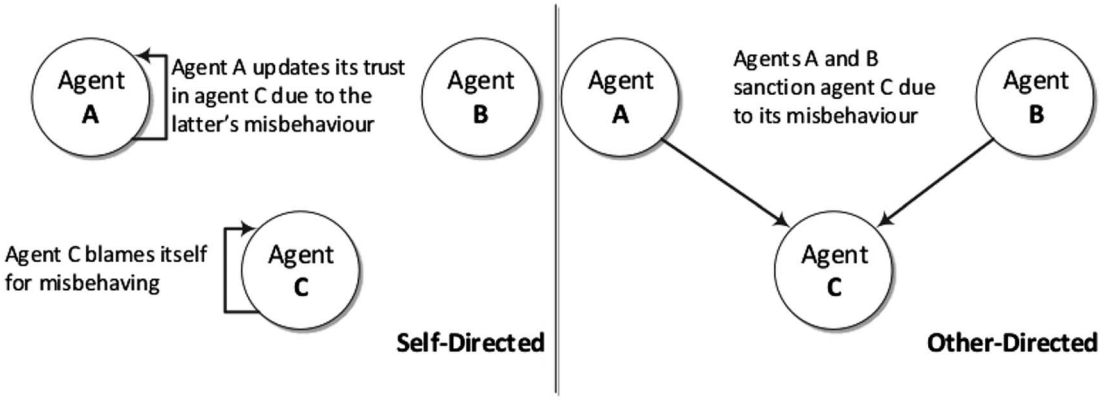


Figure 3 On the left, agent A updates its trust about agent C due to the latter's misbehaviour, and agent C reacts to its own misbehaviour by blaming itself (*sender = receiver*). On the right, agents A and B sanction agent C for its misbehaviour (*sender ≠ receiver*)

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press

Informal sanctions are established or enforced unofficially by members of the society, and need not be specified in a formal code. Examples include ridicule, ostracism, praise, and damage to or promotion of reputation (e.g. the disseminating of negative ratings about a broker that has failed to fulfil its contract agreements (S1.3)).

In law, formal sanctions include fines, mandatory social service, and imprisonment; there are no informal sanctions despite the fact that the former may facilitate the latter (Baker & Choi, 2014). In sociology, formal sanctions include not only fines and imprisonment, but also awards and bonuses, whereas informal sanctions include ridicule, ostracism, and praise.

5.1.3 Locus

The *Locus* refers to the recipient of a sanction. It determines whether a sanction is *self-directed* (i.e. the *sender* is the *receiver*) or *other-directed* (i.e. the *sender* is not the *receiver*) with respect to the individual who applies it (Figure 3). Locus does not refer to the target of the sanction, but to its recipient, even though in some cases they may coincide.

A *self-directed* sanction is directed towards and affects only its sender (e.g. Mary blames herself for the solar panel malfunctioning (S3.1)). A self-directed sanction can also refer to an action performed by another individual, which corresponds to a situation in which an individual sanctions himself because of an action by another (e.g. vicarious shame as when someone becomes ashamed due to football fans from his country misbehaving; or, when John and Joseph reduce their trust in Mary as a partner (S3.3)).

Other-directed sanctions correspond to a penalty or reward applied on another individual or group. It presumes an external action performed by the sanctioner towards the sanctionee. A classic example is the imposition of a fine due to misbehaviour or the grant of an award due to compliance (e.g. John and Joseph request compensation to Mary (S3.4); or, the consumers taking legal actions against the broker (S1.2)).

In law, other-directed sanctions include suspensions and fines, and there are no self-directed sanctions. In sociology, self-directed sanctions include guilt and trust, and other-directed sanctions include gossip and praise.

5.1.4 Mode

The *Mode* indicates how a sanction affects its target (Figure 4).

A *direct* sanction affects its target directly and immediately (e.g. the levying of a fine; or, the consumers blaming themselves for selecting the service from a mistrustful broker (S1.1)).

An *indirect* sanction affects its target indirectly, potentially influencing the future actions of others that will then affect the target (e.g. damaging the target's reputation, which would discourage others from transacting with the target; or, the dissemination of a positive opinion about Joseph by John and Mary for his initiative in forming a coalition (S2.2)).

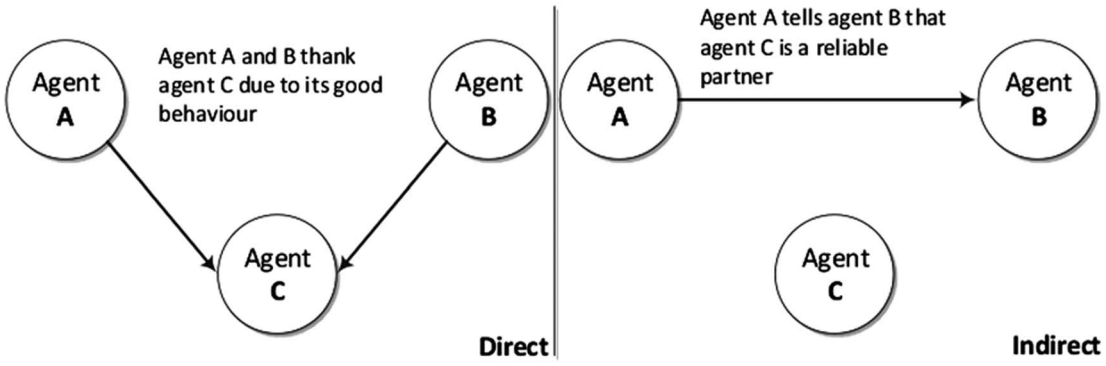


Figure 4 On the left, agents A and B directly affect agent C by thanking it for its support in previous activities (*target = receiver*). On the right, agent A indirectly affects agent C by informing agent B that agent C is a reliable partner (*target ≠ receiver*)

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press

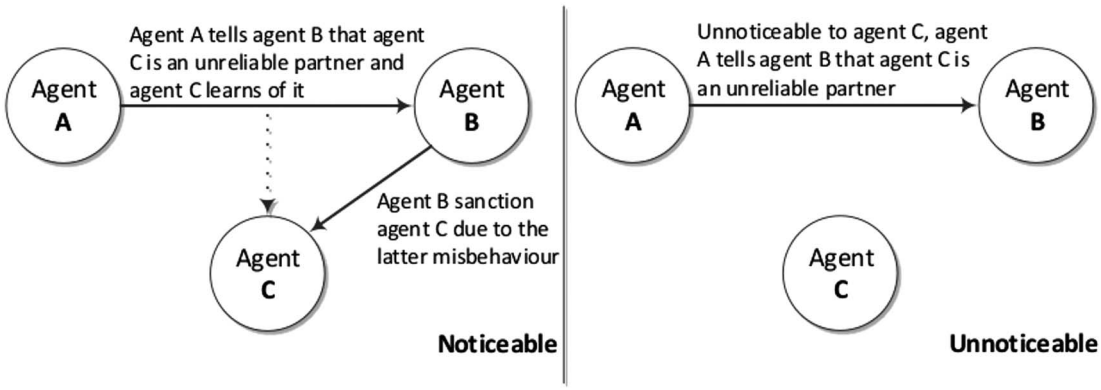


Figure 5 On the left, the sanctions are noticeable because agent C comes to know about the sanctions agents A and B are applying to it. On the right, otherwise, agent C does not notice the sanction, thus the sanction is unnoticeable

Table 1 Typologies to dimensions mapping

	Typology					
Dimension	Radcliffe-Brown (1934)	Morris (1956)	Gibbs (1966)	Pasquier *et al.* (2005)	Clinard and Meier (2008)	Cardoso and Oliveira (2011)
Purpose			✓			
Issuer	✓	✓	✓		✓	
Locus		✓	✓			
Mode						✓
Polarity	✓	✓	✓	✓	✓	
Discernability				✓		

✓ indicates the dimensions proposed in our typology that the referred existing sanction typology (identified in the header row) is capable of expressing.

5.1.5 Polarity

The *Polarity* of a sanction relates to its content. *Positive* indicates a reward (e.g. Joseph and Mary praising George to others as George successfully replaced John in the coalition (S4.2)). *Negative* indicates a penalty (e.g. John and Joseph requesting compensation from Mary for her non-fulfilment of the coalition agreement).

The law primarily considers negative sanctions, as applied in cases of violation. However, it considers positive sanctions for individuals who report fraud or help catch wanted criminals. Sociology and psychology consider both negative and positive sanctions more evenly than the law.

5.1.6 Discernability

Discernability indicates how perceptible a sanction is to its target (Figure 5).

A *noticeable* sanction, be it a penalty or a reward, is one that forces a target to notice it (e.g. Joseph and Mary thanking George for his successful help for the coalition to reach 1000 kWh (S4.1)); an *unnoticeable* sanction, such as badmouthing someone behind his or her back, is not easily noticeable (e.g. John and Joseph reduce their trust in Mary as a partner (S3.3)). A target would not easily be able to associate an unnoticeable sanction with the action that caused it.

5.2 Remarks

We now compare our typology's expressiveness with existing sanction typologies, as introduced in Sections 3 and 4. To this end, we adopt Jensen's (2002) *powerfulness* criterion, which states that a

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press

typology is more powerful than others if it creates categories that allow a more complete theoretical explanation of a set of empirical findings.

We now evaluate the dimensions of our typology with respect to governance in STSs, as exemplified by the smart grid scenario introduced above. Table 1 summarises the result of our comparison, which shows the relative advantages of our typology for STSs.

Our Purpose dimension accommodates purposes defined in the social sciences literature, thus going beyond Gibbs' (1966) conception of inducement and hedonic purposes. Our Purpose dimension provides sufficient granularity for an STS participant to select a sanction that aligns with his or her goals.

The typologies proposed in sociology, but not those in MAS, include the Issuer dimension. This dimension suits STSs well because they have aspects of both formal structure and informal relationships. A sanctioning agent can select a suitable issuer depending on the visibility or the seriousness of the sanction it wishes to apply, given its dealings with the target and with other agents.

The Locus dimension extends previous typologies by expanding self-directed sanctions based on another agent's behaviour. Doing so presents the possibility for one agent to sanction itself and thus alter either its behaviour or, more importantly, its associations with other agents as a result. For example, if John is embarrassed by his neighbours not conserving power, he may move out of the neighbourhood.

The Discernability dimension was introduced as the Style dimension in Pasquier *et al.*'s (2005) typology. A power company would noticeably sanction a consumer for non-payment via a fine or limiting the consumer's consumption for punishment purposes. However, some situations call for an unnoticeable sanction. For example, a consumer may not wish to noticeably sanction a neighbour who fails to keep her commitment to supply power for their coalition, possibly to avoid retaliation.

The Mode dimension is valuable for STSs as they involve interactions among autonomous participants. A participant, especially a regulatory agency, can apply direct sanctions. An ordinary participant can additionally apply an indirect sanction.

The Polarity dimension is common to the typologies we reviewed, except in Cardoso and Oliveira (2011). It applies to STSs because positive and negative sanctions generally apply equally to regulating interactions among autonomous parties.

To illustrate the use of the proposed typology, we classify the types of sanctions proposed by Posner and Rasmusen (1999) in Table 2.

Our typology excludes the so-called Automatic and Informational sanctions because we do not consider them to be sanctions. The Automatic sanction, for instance, is assumed to be any consequence resulting from a norm violation, even though the consequence does not intend to promote compliance with the norm. The Informational sanction is conveying undesirable information, but is not a sanction according to our definition as it is not a reaction.

The other types of sanctions that we can classify using our typology form two groups. In the first group, an individual punishes himself emotionally for what he has done (Guilt and Shame). In the second group, a second or a third-party reacts to an action (Bilateral and Multilateral costly).

Table 2 Classification of the types of sanctions proposed in Posner and Rasmusen (1999)

	Dimension					
Sanction	Purpose	Issuer	Locus	Mode	Polarity	Discernability
Automatic	—	—	—	—	—	—
Guilt	Punishment	Informal	Self-directed	Direct	Negative	Unnoticeable
Shame	Punishment	Informal	Self-directed	Direct	Negative	Unnoticeable
Informational	—	—	—	—	—	—
Bilateral costly	Punishment	Informal	Other-directed	Direct	Negative	Unnoticeable
Multilateral costly	Punishment	Informal	Other-directed	Direct	Negative	Unnoticeable

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press

6 Conceptual sanctioning process model

We now adopt the foregoing static model of sanctions as a foundation for a conceptual sanctioning process model. This conceptual model provides a functional structure for sanctioning in STS, showing its main capabilities and their relationships.

We begin from a sanctioning process for nMAS as proposed by Balke and Villatoro (2012). Their proposed process is composed of four stages: (i) *violation detection* involves monitoring agents to check whether other agents comply with the norms; (ii) *sanctioning determination* evaluates the violation of or compliance with norms and determines a sanction; (iii) if so, *sanctioning application* takes over; (iv) *assimilation* involves monitoring the sanction application to determine its efficacy. We extend Balke and Villatoro's (2012) model by associating specific capabilities with these stages.

Figure 6 depicts our conceptual sanctioning process model, illustrating the aforementioned stages being enacted by five capabilities (active entities: *Detector*, *Evaluator*, *Executor*, *Controller*, and *Legislator*) using two resources (passive entities: the data repositories *De Jure* and *De Facto*). Note that these capabilities and resources may occur in multiple ways, including in a fully centralised or a fully decentralised manner. Unlike Balke and Villatoro, our sanctioning process incorporates both norm violation and compliance, respecting the general notion of sanction we motivated above. (Please note that capitalisation matters in the text below: De Jure and De Facto refer to the repositories; de jure and de facto are modifiers as in 'de jure norms'.)

The De Jure repository stores norms and sanctions (as specifications) as well as links between them, that is, which sanctions apply to what norm violation or compliance: the relationship between norms and sanctions can be many to many. These norms and sanctions are initially given, but the Legislator entity may include, remove, or change specifications and relations at run-time.

The De Facto repository stores information about the sanctions as applied, and relevant information such as the observed violations, which can be used to assess the value and efficacy of different sanctions in achieving their purpose in specific contexts.

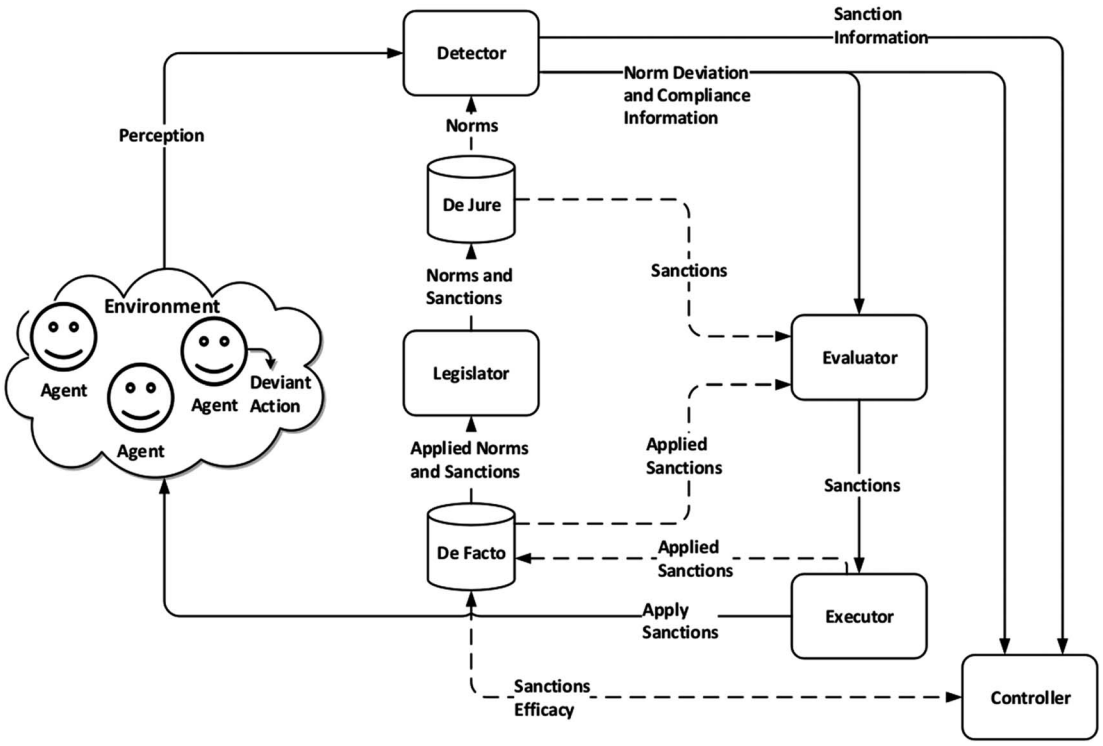


Figure 6 Conceptual sanctioning process model

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press

A significant benefit of our model is that it supports storing conflicting information in De Jure and De Facto. Particularly, a sanction (and the underlying norm) specified in De Jure may not be apparent in De Facto, indicating the well-known idea of a discrepancy between what is conceived and what is realized.

An *agent* represents an entity capable of performing actions in its environment and, importantly, of interacting with other agents. In our model, an agent's function is to represent the interest and perspective of a social entity in a given STS. An agent stands in for any social entity. Specific capabilities that agents have are as indicated in the model. Specifically, a Detector perceives the environment and detects a norm violation or compliance, and sanctions applied by other agents. In general, the environment would be only partially observable because of (i) its size and complexity, including the number of participants; (ii) the impossibility of identifying the Executor of an action; and (iii) the confidentiality of some communications.

Assuming the Detector perceives an action, it determines whether the agent who performed that action is governed by a de jure norm (e.g. given its capabilities in the STS) and, if so, whether the action violates or complies with any norm. Note that we limit the Detector to work based on de jure norms, the idea being that a violation or a compliance being detected is given de jure status.

The Evaluator, in addition, obtains information from De Jure and De Facto in order to determine whether to apply a sanction and, if so, which sanction. De Facto captures previous behaviours reported by the Controller and any sanctions applied in those cases, be it by the Evaluator or by other agents. The Evaluator's reasoning could incorporate the magnitude of the violation and an assessment of the success of previous sanctions with respect to their purposes. Importantly, De Facto is not necessarily a unitary entity. Hence, the Evaluator may access a portion of De Facto that captures not only the experiences shared among some members of the STS, but also personal experiences of the Evaluator. In Situation 3 of the motivation scenario, for instance, John and Joseph have the capability of Evaluators and recognise that Mary has violated her commitment to them. Assessing the situation, John disregards her fault as he had good prior experiences with her and assumes this as an exception. Joseph, however, who does not have any prior experience with her bases his decision on what sanction others usually apply to such a situation (de facto), and decides to tell others that Mary is an unreliable partner.

The Executor has the power to execute a sanction. In general, a formal sanction may require a more specific kind of Executor than an informal sanction. For example, imprisonment must be executed by the police even though the Evaluator is a judge, whereas ostracism may be executed by the same individual who serves as Evaluator. In Situation 3, Joseph as Executor tells other that Mary is an unreliable partner.

The Controller records in De Facto the sanctions applied by itself and other agents, and monitors the outcomes of applying a sanction, including the future behaviour of the target, such as to evaluate the efficacy of the sanction. Joseph, for instance, after telling others that Mary is an unreliable partner, may use his Controller's capability to monitor the environment and to identify whether the sanction applied has prevented her from forming partnership with others. If so, Joseph increases the sanction value to that kind of fault, otherwise he decreases such value.

The Legislator updates de jure norms and sanctions based on an assessment of De Jure and De Facto along with the environment. The updates could be motivated by reducing misalignments between de facto and de jure norms and sanctions.

7 Demonstration on the motivating scenario

Given an STS with its members' mutual expectations expressed as norms and sanctions, both in De Jure and De Facto, let us now demonstrate our sanction typology and conceptual sanctioning process model using our motivating scenario. As noted in the scenario, an *affected party* is one affected by a norm violation or compliance; a *third-party* is one that observes a norm violation or compliance, and albeit not affected, reacts to it; an *enforcer* is one that applies the sanction. The affected parties and third-parties can potentially choose among multiple sanctions for reacting to each situation. The enforcer would thus apply such sanctions on a (*sanction*) *target*. Table 3 classifies the possible sanctions identified in the motivation scenario in Section 2 according to our proposed typology.

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press

Table 3 Classification of the possible sanctions identified in the motivating scenario situations

	Role			Dimension					
Sanction	Affected Party or Third-Party	Sanction target	Sanction receiver	Purpose	Issuer	Locus	Mode	Polarity	Discernability
S1.1	John/Joseph/Mary	John/Joseph/Mary	John/Joseph/Mary	Punishment	Informal	Self-directed	Direct	Negative	Noticeable
S1.2	John/Joseph/Mary	Broker	Regulatory agency	Punishment	Formal	Other-directed	Indirect	Negative	Noticeable
S1.3	John/Joseph/Mary	Broker	Other consumers	Punishment	Informal	Other-directed	Indirect	Negative	Unnoticeable
S1.4	John/Joseph/Mary	Broker	Broker	Punishment	Informal	Other-directed	Direct	Negative	Noticeable
S1.5	Broker	Energy provider	Regulatory agency	Punishment	Formal	Other-directed	Indirect	Negative	Noticeable
S1.6	Regulatory agency	Energy provider	Energy provider	Punishment	Formal	Other-directed	Direct	Negative	Noticeable
S1.7	Regulatory agency	Broker	Broker	Incapacitation	Formal	Other-directed	Direct	Negative	Noticeable
S2.1	John/Mary	Joseph	Joseph	Reward	Informal	Other-directed	Direct	Positive	Noticeable
S2.2	John/Mary	Joseph	Other consumers	Reward	Informal	Other-directed	Indirect	Positive	Unnoticeable
S3.1	Mary	Mary	Mary	Punishment	Informal	Self-directed	Direct	Negative	Noticeable
S3.2	John/Joseph	Mary	Mary	Guidance	Informal	Other-directed	Direct	Positive	Noticeable
S3.3	John/Joseph	Mary	John/Joseph	Punishment	Informal	Self-directed	Indirect	Negative	Unnoticeable
S3.4	John/Joseph	Mary	Mary	Punishment	Formal	Other-directed	Direct	Negative	Noticeable
S3.5	John/Joseph	Mary	Other consumers	Punishment	Informal	Other-directed	Indirect	Negative	Unnoticeable
S4.1	Mary/Joseph	George	George	Reward	Informal	Other-directed	Direct	Positive	Noticeable
S4.2	Mary/Joseph	George	Other consumers	Reward	Informal	Other-directed	Indirect	Positive	Unnoticeable
S4.3	Mary/Joseph	John	Other consumers	Reward	Informal	Other-directed	Indirect	Positive	Unnoticeable
S4.4	Mary/Joseph	John	Mary/Joseph	Incapacitation	Informal	Self-directed	Indirect	Negative	Unnoticeable
S5.1	Big consumer	Broker	Big consumer	Reward	Informal	Self-directed	Indirect	Positive	Unnoticeable
S5.2	Big consumer	Broker	Other consumers	Reward	Informal	Other-directed	Indirect	Positive	Unnoticeable

Sanction S1.1 is classified as a self-directed locus (the sanction sender and receiver are the same individual); direct mode; negative polarity (negative emotions); noticeable; and of an informal issuer (there is no formal rule for guilt). Sanction S3.1 may be treated similarly.

Even though Sanctions S3.3, S4.4, and S5.1 are classified as having a self-directed locus (sanction sender and receiver are the same individual), because their contents refer to another agent's behaviour potentially not aware of its lowered trust, these sanctions have unnoticeable discernability and indirect mode. This happens because whereas the Locus dimension refers to the affected or third-party, the Discernability and Mode dimensions refer to the target.

Being legal, Sanctions S1.2, S1.5, S1.6, S1.7, and S3.4 have a formal issuer. In contrast, Sanction S1.4 has an informal issuer because it is applied by consumers, who have the right to change service providers at any time.

Sanctions S1.3, S2.2, S3.5, S4.2, S4.3, and S5.2 involve disseminating reputation (informal and other-directed) differing only in their polarity. Disseminating evaluations about the target can affect the target's reputation and thereby influence future decisions by others (other-directed locus), but it is unnoticeable (the target is usually unaware of it) and of indirect mode. Sanctions S2.1, S3.2, and S4.1 are noticeable and direct as they are communicated directly to the target.

To demonstrate our conceptual sanctioning process model, we now expand Situation 1 of our scenario. In this situation, each participant (John, Joseph, Mary, the broker, and the regulatory agency) have the Detector capability. For John, Joseph, and Mary, detection is easy as the failure affects them directly. As Evaluators, they select one or more sanctions from among the four available ones (S1.1–S1.4). Specifically, they can choose Sanction S1.1 if the severity of the failure is not high and they decide not to pursue a legal remedy. Or, if the failure causes significant harm, they might apply sanction S1.2. In addition or instead, John, Joseph, and Mary may report a negative rating about the broker as a service provider (S1.3). Moreover, depending on the frequency of such failures, they may take their business to another broker (S1.4) as the last resort to obtain better service. In these cases, the same individual serves as the Executor.

In contrast, Sanction S1.3 involves another entity, empowered to evaluate and to apply legal sanctions. This dependence on another party can affect the efficacy of a sanction. For example, in many cases, a legal process may not produce the expected results in the expected time frame. Consequently, an affected party may decide not to file a lawsuit even if the success of winning is warranted.

The broker may be directly affected by the energy provider's failure. As Detector, the broker assesses the severity of the failure. Judging the impact of the failure on the broker is more complex than for individual consumers. For example, the failure may have affected 90% of the consumers who have contracted the broker, and the broker would need to sample its consumers to estimate the impact. As Evaluator, the broker may decide on whether to file a complaint against the energy provider with the regulatory agency (S1.5). If so, as an Executor, the broker files a complaint, which is treated as an event in its own right, despite a social event. As a Detector, the regulatory agency detects that a complaint is filed. In addition, the regulatory agency as a Detector may capture the complaints of multiple consumers or brokers. It determines whether the complaints are legitimate, whether there was a norm violation by the energy provider, and the severity of the violation. As an Evaluator, it determines its sanctions and its targets (the provider and possibly the brokers). Possible sanctions include levying a fine on the provider and suspending the broker. As an Executor, it applies these sanctions.

8 Conclusions and future work

The main contribution of this paper is an approach to sanctioning that expands our understanding of how norms can be used as a foundation for governing STSs. Our approach (i) supports a rich panoply of sanctions (ii) that are loosely coupled with norms, and which (iii) may be flexibly chosen dependent upon contextual, historical, and personal factors.

We established that existing MAS sanctioning process models are inadequate for meeting these requirements. We found useful insights regarding sanctions in the law, sociology, and psychology literatures. We consolidated and enhanced these insights to develop a new sanction typology along with a conceptual sanctioning process model. This typology and process model assist in evaluating and in selecting sanctions for a particular norm violation or compliance.

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press

Interesting directions for future work include (i) conducting experiments with humans to evaluate the proposed typology of sanctions; (ii) formalising the proposed conceptual sanctioning process model; (iii) developing suitable decision-making techniques with which to choose appropriate sanctions in STSs; and (iv) realising the conceptual sanctioning process model and incorporating it in tools for social simulation.

Acknowledgements

This work was partially supported by the University Global Partnership Network, a collaboration among North Carolina State University, University of São Paulo, and University of Surrey. L. G. Nardin was partially funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 315874. T. Balke-Visser was partially funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 288147 as well as by the EPSRC-funded Whole Systems Energy Modelling Consortium (wholeSEM). N. Ajmeri and M. P. Singh were partially supported by the US Department of Defense under the Science of Security Lablet grant. J. S. Sichman was partially supported by CNPq, Brazil, under grant agreement no. 303950/2013-7.

References

Aamodt, A. & Plaza, E. 1994. Case-based reasoning; foundational issues, methodological variations, and system approaches. *AI Communications* **7**(1), 39–59.

Andrighetto, G. & Villatoro, D. 2011. Beyond the carrot and stick approach to enforcement: an agent-based model. In *European Perspectives on Cognitive Sciences*, Kokinov, B., Karmiloff-Smith, A. & Nersessian, N. J. (eds), pp. 1–6. New Bulgarian University Press.

Andrighetto, G., Villatoro, D. & Conte, R. 2010. Norm internalization in artificial societies. *AI Communications* **23**(4), 325–339.

Arionfreed, J. M. 1968. The concept of internalization. In *Conduct and Conscience*, Aronfreed J. M. (ed.). Academic Press, 15–42.

Bailey, K. D. 1994. *Typologies and Taxonomies: An Introduction to Classification Techniques. Quantitative Applications in the Social Sciences*. Sage Publications.

Baker, S. & Choi, A. H. 2014. *Crowding In: How Formal Sanctions Can Facilitate Informal Sanctions*. Technical report no. 2014-01/2014-04, University of Virginia School of Law. `http://dx.doi.org/10.2139/ssrn.2374109`.

Balke, T. 2009. A taxonomy for ensuring institutional compliance in utility computing. In *Normative Multi-Agent Systems*, Boella G., Noriega P., Pigozzi G. & Verhagen H. (eds), Dagstuhl Seminar Proceedings **09121**, 1–17. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Balke, T. & Villatoro, D. 2012. Operationalization of the sanctioning process in utilitarian artificial societies. In *Coordination, Organizations, Institutions, and Norms in Agent System VII*, Cranefield S., Riemsdijk M., Vzquez-Salceda J. & Noriega P. (eds), Lecture Notes in Computer Science **7254**, 167–185. Springer.

Beccaria, M. & Ingraham, E. D. 1819. *An Essay on Crimes and Punishments*. Philip H. Nicklin.

Bentham, J. 1823. *An Introduction to the Principles of Morals and Legislation*. Clarendon Press.

Bicchieri, C. 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press.

Broersen, J., Cranefield, S., Elrakaiby, Y., Gabbay, D., Grossi, D., Lorini, E., Parent, X., van der Torre, L. W. N., Tummolini, L., Turrini, P. & Schwarzentruber, F. 2013. Normative reasoning and consequence. In *Normative Multi-Agent Systems*, Andrighetto G., Governatori G., Noriega P. & van der Torre L. W. N. (eds), Dagstuhl Follow-Ups **4**, 33–70. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Campos, J., Lopez-Sanchez, M., Salamó, M., Avila, P. & Rodrguez-Aguilar, J. A. 2013. Robust regulation adaptation in multi-agent systems. *ACM Transactions on Autonomous and Adaptive Systems* **8**(3), 13:1–13:27.

Cardoso, H. L. & Oliveira, E. 2009. Adaptive deterrence sanctions in a normative framework. In *Proceedings of the IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, **2**, 36–43, IEEE Computer Society.

Cardoso, H. L. & Oliveira, E. 2011. Social control in a normative framework: an adaptive deterrence approach. *Web Intelligence and Agent Systems* **9**(4), 363–375.

Carlsmith, K. M. 2006. The roles of retribution and utility in determining punishment. *Journal of Experimental Social Psychology* **42**(4), 437–451.

Carlsmith, K. M., Darley, J. M. & Robinson, P. H. 2002. Why do we punish?: deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology* **83**(2), 284–299.

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press

Castelfranchi, C. 2000. Engineering social order. In *Proceedings of the 1st International Workshop on Engineering Societies in the Agent World (ESAW)*, **1972**, 1–18, Springer.

Castelfranchi, C. & Falcone, R. 1998. Principles of trust in MAS: cognitive anatomy, social importance, and quantification. In *Proceedings of International Conference on Multi Agent Systems (ICMAS)*, 72–79. IEEE Computer Society.

Cavadino, M. & Dignan, J. 2002. *The Penal System: An Introduction*. Sage.

Centeno, R., Billhardt, H. & Hermoso, R. 2011. An adaptive sanctioning mechanism for open multi-agent systems regulated by norms. In *Proceedings of the 23rd International Conference on Tools with Artificial Intelligence, ICTAI, Boca Raton, FL, USA, November 7–9, 2011*, 523–530. IEEE Computer Society.

Centeno, R., Billhardt, H. & Hermoso, R. 2013. Persuading agents to act in the right way: an incentive-based approach. *Engineering Applications of Artificial Intelligence* **26**(1), 198–210.

Clinard, M. B. & Meier, R. F. 2008. *Sociology of Deviant Behavior*, 3rd edition. Thomson/Wadsworth.

Conte, R. & Paolucci, M. 2002. *Reputation in Artificial Societies: Social Beliefs for Social Order*. Springer.

Criado, N., Argente, E., Noriega, P. & Botti, V. 2013. Manea: a distributed architecture for enforcing norms in open MAS. *Engineering Applications of Artificial Intelligence* **26**(1), 76–95.

Daskalopulu, A., Dimitrakos, T. & Maibaum, T. 2002. Evidence-based electronic contract performance monitoring. *Group Decision and Negotiation* **11**(6), 469–485.

Davis, M. 2009. Punishment theory's golden half century: a survey of developments from (about) 1957 to 2007. *The Journal of Ethics* **13**(1), 73–100.

Dellarocas, C. 2006. Reputation mechanisms. In *Handbook on Economics and Information Systems*, Hendershott T. (ed.), **1**. Elsevier Science Publishers B. V., 629–660.

Department of Energy 2003. *Grid 2030: A National Vision for Electricity's Second 100 Years*. Technical report, US Department of Energy.

Ellickson, R. C. 1991. *Order Without Law: How Neighbors Settle Disputes*. Harvard University Press, ISBN 0-674-64169-8.

Energy Policy Act 1992. Pub. l. 102-486, 106 stat. 2776.

Esteva, M., Rodrguez-Aguilar, J. A., Arcos, J. L., Sierra, C. & Garcia, P. 2000. Institutionalizing open multi-agent systems. In *4th International Conference on Multi-Agent Systems (ICMAS), Boston*, 381–382. IEEE Computer Society.

Esteva, M., Rodríguez-Aguilar, J. A., Sierra, C., Garcia, P. & Arcos, J. L. 2001. On the formal specifications of electronic institutions. In *Agent Mediated Electronic Commerce: The European AgentLink Perspective*, Lecture Notes in Artificial Intelligence **1991**, 126–147. Springer.

Faci, N., Modgil, S., Oren, N., Meneguzzi, F., Miles, S. & Luck, M. 2008. Towards a monitoring framework for agent-based contract systems. In *Proceedings of the Twelfth International Workshop on Cooperative Information Agents*, Klusch M., Pechoucek M. & Polleres A. (eds), Lecture Notes in Artificial Intelligence **5180**, 292–305.

Fiadeiro, J. L. 2008. On the challenge of engineering socio-technical systems. In *Software-Intensive Systems and New Computing Paradigms*, Wirsing M., Banâtre J.-P., Hölzl M. & Rauschmayer A. (eds), Lecture Notes in Computer Science **5380**, 80–91. Springer.

Gabriel, U. & Oswald, M. E. 2007. Psychology of punishment. In *Encyclopedia of Law and Society: American and Global Perspectives*, D. S. Clark (ed.). Sage, 1252–1254.

Gardner, J. 2001. Legal positivism: $5^{1/2}$ myths. *American Journal of Jurisprudence* **46**, 199–227.

Garner, B. A. (ed.) 2010. *Black's Law Dictionary*, 9th edition. West Group.

Giardini, F., Andrighetto, G. & Conte, R. 2010. A cognitive model of punishment. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, Ohlsson S. & Catrambone R. (eds), 1282–1288. Cognitive Science Society.

Giardini, F., Paolucci, M., Villatoro, D. & Conte, R. 2014. Punishment and gossip: sustaining cooperation in a public goods game. In *Advances in Social Simulation*, Kamiński B. & Koloch G. (eds), Advances in Intelligent Systems and Computing **229**, 107–118. Springer.

Gibbs, J. P. 1966. Sanctions. *Social Problems* **14**(2), 147–159.

Grossi, D., Aldewereld, H. & Dignum, F. 2007. Ubi lex, ibi poena: designing norm enforcement in e-institutions. In *Coordination, Organizations, Institutions, and Norms in Agent Systems II*, Noriega P., Vázquez-Salceda J., Boella G., Boissier O., Dignum V., Fornara N. & Matson E. (eds), Lecture Notes in Computer Science **4386**, 101–114. Springer.

Habermas, J. 1984. *The Theory of Communicative Action: Reason and the Rationalisation of Society*, 1. Beacon Press.

Harper, D. 2010. *Online Etymology Dictionary*. `http://www.etymonline.com`.

Hart, H. L. A. 1968. *Punishment and Responsibility*. Oxford University Press.

Heitz, M., König, S. & Eymann, T. 2010. Reputation in multi agent systems and the incentives to provide feedback. In *Multiagent System Technologies*, Dix J. & Witteveen C. (eds), Lecture Notes in Computer Science **6251**, 40–51. Springer.

Helbing, D., Szolnoki, A., Perc, M. & Szabó, G. 2010. Punish, but not too hard: how costly punishment spreads in the spatial public goods game. *New Journal of Physics* **12**(8), 083005.

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press

Hendrikx, F., Bubendorfer, K. & Chard, R. 2015. Reputation systems: a survey and taxonomy. *Journal of Parallel and Distributed Computing* **75**, 184–197.

Houwing, M., Heijnen, P. W. & Bouwmans, I. 2006. Socio-technical complexity in energy infrastructures conceptual framework to study the impact of domestic level energy generation, storage and exchange. In *IEEE International Conference on Systems, Man and Cybernetics*, **2**, 906–911.

Hurwicz, L. & Reiter, S. 2008. *Designing Economic Mechanism*. Cambridge University Press.

Huynh, T. D., Jennings, N. R. & Shadbolt, N. 2004. FIRE: an integrated trust and reputation model for open multi-agent systems. In *Proceedings of the 16th European Conference on Artificial Intelligence*, 18–22. Valencia.

Jensen, G. 2002. Typologizing violence: a Blackian perspective. *International Journal of Sociology and Social Policy* **22**(7/8), 75–108.

Jones, A. J. I., Artikis, A. & Pitt, J. 2013. The design of intelligent socio-technical systems. *Artificial Intelligence Review* **39**(1), 5–20.

Jones, A. J. I. & Sergot, M. 1993. On the characterization of law and computer systems: the normative systems perspective. In *Deontic Logic in Computer Science*, Meyer J.-J. C. & Wieringa R. J. (eds). John Wiley & Sons, 275–307.

Jøsang, A. 2001. A logic for uncertain probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based* **9**(3), 279–311.

Kalia, A. K., Zhang, Z. & Singh, M. P. 2014. Estimating trust from agents' interactions via commitments. In *ECAI-21st European Conference on Artificial Intelligence, Prague*, Schaub T., Friedrich G. & O'Sullivan B. (eds), Frontiers in Artificial Intelligence and Applications **263**, 1043–1044. IOS Press.

Kant, I. 1999. *Metaphysical Elements of Justice: Part I of the Metaphysics of Morals*, 2nd edition, Classics Series. Hackett Publishing Company.

López, F. L. y. & Luck, M. 2003. Modelling norms for autonomous agents. In *Proceedings of the 4th Mexican International Conference on Computer Science*, Chávez E., Favela J., Mejía M. & Oliart A. (eds), 238–245. IEEE Computer Society.

Mah, D. N., van der Vleuten, J. M., Ip, J. C. & Hills, P. R. 2012. Governing the transition of socio-technical systems: a case study of the development of smart grids in Korea. *Energy Policy* **45**, 133–141.

Mahmoud, S., Griffiths, N., Keppens, J. & Luck, M. 2012a. Efficient norm emergence through experiential dynamic punishment. In *ECAI-20th European Conference on Artificial Intelligence, Montpellier, France*, L. D. Raedt, C. Bessière, D. Dubois, P. Doherty, P. Frasconi, F. Heintz & P. J. F. Lucas (eds), Frontiers in Artificial Intelligence and Applications **242**, 576–581. IOS Press.

Mahmoud, S., Villatoro, D., Keppens, J. & Luck, M. 2012b. Optimised reputation-based adaptive punishment for limited observability. In *Sixth IEEE International Conference on Self-Adaptive and Self-Organizing Systems, SASO 2012, Lyon, France, September 10–14, 2012*, 129–138. IEEE Computer Society.

Meares, T. L., Katyal, N. & Kahan, D. M. 2004. Updating the study of punishment. *Stanford Law Review* **56**, 1171–1210.

Meier, R. F. 1982. Perspectives on the concept of social control. *Annual Review of Sociology* **8**, 35–55.

Miethe, T. D. & Lu, H. 2005. *Punishment—A Comparative Historical Perspective*. Cambridge University Press.

Miles, S. & Griffiths, N. 2015. Accounting for circumstances in reputation assessment. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, Weiss G., Yolum P., Bordini R. H. & Elkind E. (eds), 1653–1654. ACM Press.

Mill, J. S. 1871. *Utilitarianism*, 4th edition. Longmans.

Minsky, N. H. 1991. Law-governed systems. *Software Engineering Journal* **6**(5), 285–302.

Modgil, S., Faci, N., Meneguzzi, F., Oren, N., Miles, S. & Luck, M. 2009. A framework for monitoring agent-based normative systems. In *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1, (AAMAS)*, 153–160. International Foundation for Autonomous Agents and Multiagent Systems.

Morris, R. T. 1956. A typology of norms. *American Sociological Review* **21**(5), 610–613.

Mui, L., Halberstadt, A. & Mohtashemi, M. 2002. Evaluating reputation in multi-agents systems. In *Trust, Reputation, and Security: Theories and Practice, AAMAS International Workshop, Bologna, Italy, July 15, 2002, Selected and Invited Papers*, Falcone R., Barber K. S., Korba L. & Singh M. P. (eds), 123–137. Springer.

Nagin, D. 1998. Deterrence and incapacitation. In *The Handbook of Crime and Punishment*, Tonry M. (ed.). Oxford University Press, 345–368.

Pasquier, P., Flores, R. A. & Chaib-draa, B. 2005. Modelling flexible social commitments and their enforcement. In *Proceedings of the 5th International Conference on Engineering Societies in the Agents World (ESAW)*, Gleizes M. P., Omicini A. & Zambonelli F. (eds), Lecture Notes in Computer Science **3451**, 139–151. Springer.

Patterson, D. (ed.) 2010. *A Companion to Philosophy of Law and Legal Theory*, 2nd edition. Wiley-Blackwell.

Petersen, M. B., Sell, A., Tooby, J. & Cosmides, L. 2012. To punish or repair? Evolutionary psychology and lay intuitions about modern criminal justice. *Evolution and Human Behavior* **33**(6), 682–695.

Piaget, J. 1995. *Sociological Studies*. Routledge.

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press

Picket, J. P. (ed.) 2011. *The American Heritage Dictionary of the English Language*, 5th edition. Houghton Mifflin Harcourt.

Pinninck, A. P. d. 2010. *Techniques for Peer Enforcement in Multiagent Networks*. PhD thesis, Universitat Autónoma de Barcelona.

Pinninck, A. P. d., Sierra, C. & Schorlemmer, M. 2010. A multiagent network for peer norm enforcement. *Journal of Autonomous Agents and Multi-Agent Systems* **21**(3), 397–424.

Pinyol, I. & Sabater-Mir, J. 2013. Computational trust and reputation models for open multi-agent systems: a review. *Artificial Intelligence Review* **40**(1), 1–25.

Pinyol, I., Sabater-Mir, J., Dellunde, P. & Paolucci, M. 2012. Reputation-based decisions for logic-based cognitive agents. *Autonomous Agents and Multi-Agent Systems* **24**(1), 175–216.

Posner, E. A. 2000. *Law and Social Norms*. Harvard University Press.

Posner, R. A. & Rasmusen, E. B. 1999. Creating and enforcing norms, with special reference to sanctions. *International Review of Law and Economics* **19**(3), 369–382.

PowerTAC. 2010. Power Trading Agent Competition. `http://www.powertac.org`.

Radcliffe-Brown, A. R. 1934. Social sanction. In *Encyclopedia of the Social Sciences*, E. R. A. Seligman (ed.), **XIII**. Macmillan Publishers, 531–534.

Rodrigues, M. R. & Luck, M. 2007. Cooperative interactions: an exchange values model. In *Coordination, Organizations, Institutions, and Norms in Agent Systems II*, Noriega P., Vázquez-Salceda J., Boella G., Boissier O., Dignum V., Fornara N. & Matson E. (eds), Lecture Notes in Computer Science **4386**, 356–371. Springer.

Sabater-Mir, J., Paolucci, M. & Conte, R. 2006. Repage: REPutation and imAGE among limited autonomous partners. *Journal of Artificial Societies and Social Simulation* **9**(2), 3.

Sabater-Mir, J. & Sierra, C. 2002. Social ReGreT, a reputation model based on social relations. *ACM SIGecom Exchanges* **3**(1), 44–56.

Schwartz, R. D. & Orleans, S. 1967. On legal sanctions. *The University of Chicago Law Review* **34**(2), 274–300.

Singh, M. P. 2013. Norms as a basis for governing sociotechnical systems. *ACM Transactions on Intelligent Systems and Technology* **5**(1), 21:1–21:23.

Singh, M. P., Arrott, M., Balke, T., Chopra, A. K., Christiaanse, R., Cranefield, S., Dignum, F., Eynard, D., Farcas, E., Fornara, N., Gandon, F., Governatori, G., Dam, H. K., Hulstijn, J., Krüger, I., Lam, H.-P., Meisinger, M., Noriega, P., Savarimuthu, B. T. R., Tadanki, K., Verhagen, H. & Villata, S. 2013. The uses of norms. In *Normative Multi-Agent Systems*, Andrighetto G., Governatori G., Noriega P. & van der Torre L. W. N. (eds), Dagstuhl Follow-Ups **4**, 191–229. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Singh, M. P. & Huhns, M. N. 2005. *Service-Oriented Computing: Semantics, Processes, Agents*. John Wiley & Sons.

Skinner, B. F. 1938. *The Behavior of Organisms: An Experimental Analysis*. Appleton-Century.

Vercouter, L. & Muller, G. 2010. L.I.A.R. achieving social control in open and decentralized multiagent systems. *Applied Artificial Intelligence* **24**(8), 723–768.

Villatoro, D., Andrighetto, G., Sabater-Mir, J. & Conte, R. 2011. Dynamic sanctioning for robust and cost-efficient norm compliance. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, 414–419. AAAI Press.

Vu, K., Begouic, M. M. & Novosel, D. 1997. Grids get smart protection and control. *IEEE Computer Applications in Power* **10**(4), 40–44.

Wang, Y. & Singh, M. P. 2010. Evidence-based trust: a mathematical model geared for multiagent systems. *ACM Transactions on Autonomous and Adaptive Systems* **5**(4), 14:1–14:28.

Weigand, H. 2009. Using communication norms in socio-technical systems. In *Handbook of Research on Socio-Technical Design and Social Networking Systems*, Whitworth, B. & de Moor, A. (eds). IGI Global, 224–235.

Whitworth, B. 2006. Social-technical systems. In *Encyclopedia of Human Computer Interaction*, C. Ghaoui (ed.). Idea Group Reference, 533–541.

Zacharia, G. & Maes, P. 2000. Trust management through reputation mechanisms. *Journal of Applied Artificial Intelligence* **14**(9), 881–907.

https://doi.org/10.1017/S0269888916000023 Published online by Cambridge University Press