The ethics of controllability as influenceability

Emiliano Lorini, Giovanni Sartor

To be published in Jurix-2021 Proceedings, IOS Press

The Ethics of Controllability as Influenceability

Emiliano Lorini [a] Giovanni Sartor [b]
[a] *IRIT-CNRS, Toulouse University, France*
[b] *CIRSFID - Alma AI, University of Bologna, Italy;*
European University Institute, Florence, Italy

Abstract. We present a logical analysis of influence and control over the actions of others, and address consequential causal and normative responsibilities. We first account for the way in which influence can be exercised over the behaviour of autonomous agents. On this basis we determine the conditions under which influence leads to control on the implementation of positive and negative values. We finally define notions of causal and normative responsibility for the action of others. Our logical framework is based on STIT logic and is complemented with a series of examples illustrating the application. Our analysis applies to interactions between humans as well as to those involving autonomous artificial agents.

Keywords. logic of action, influence, responsibility, control, values

1. Introduction

As AI systems become more and more pervasive, autonomous, and powerful, their actions increasingly affect human values and interests. AI agents, however, are not alone, they rather participate in ecosystems in which multiple agents, humans and artificial ones influence one another. To understand the significance of agents' actions, and allocate responsibilities concerning their outcomes, we need to put such actions in the context of the influence patterns determining the performance of their actions as well as their indirect effects. More precisely we need to address two parallel issues. The first issue concerns incoming influences, i.e., determining to what extent and in what ways other agents have or may have influenced the agent's behaviour. The second issue concerns outcoming influences, i.e., determining to what extent and in what ways the agent has or may have influenced other agents. As an example involving both issues, consider for instance the case in which a bot is used to spread fake news, hate content, misleading ads, or to maximise engagement (in whatever way). In such cases the bot's owner influences the bot to influence the behaviour of the receivers of the bot's messages. When an agent has the ability to exercise influence over the behaviour of another agent we may say that the first agent has control over the second one. Control may cover all actions of the controlled agent, or only a subset of them.

For instance, relative to the actions by the online bot spreading unlawful or unethical content not only the user, but also the platform operator has some control, since the platform operator too could have blocked, or restricted, the operation of the bot, preventing

its activity. The platform operator also has control over the users of the bot, which he can exercise, for instance, by threatening sanctions (such as the exclusion from the platform) against the use of bots for such a purpose.

Control may have a normative significance: the controller may be praised or blamed when making the controllee perform good or bad actions, respectively. In some cases, the controller can also be blamed for not preventing the controllee from performing bad actions. In fact, when agent i has the capacity to exercise control in such a way as to prevent j from behaving badly, or to ensure that j behaves well, then it may make sense to consider i to be accountable for j's failures and possibly to subject i to sanction for such failures. For instance, if the platform's owners have the possibility to exercise control over the messages exchanged over their platforms, then it may make sense to blame and sanction them because of the harm resulting from such messages, even though they do not take the initiative to send such messages. Note however that failure to exercise control does not necessarily entail blame and sanction for controllers. The exercise of control entails disadvantages, regarding both the autonomy of the controllee and the expected social outcomes. It may be the case that, all things considered, the exercise of control would entail social costs that exceed the benefits it may provide. For instance, assume that the platform's owner could prevent all unlawful behaviours on its platform only by stopping all messages. However, this would entail more harm than good. Therefore, the controller cannot be blamed for failing to take such an action.

The purpose of this paper it to provide a logical framework for modelling patterns of influence and control. Our framework can be useful for understanding contexts where agents do or do not, can or cannot, exercise influence over others, and therefore for determining how prise and blame, and consequently responsibilities, should be allocated to such agents. Consider again the case of online platforms. As an instance of a prohibition to promote bad behaviour, consider that ISPs may be ethically or legally responsible for inducing their users to engage in harmful behaviour, as in the case of websites that are devoted to enabling online revenge porn, the spreading of politically oriented fake news, or to the distribution of unauthorised copyright materials. As an instance of an obligation to prevent bad behaviour, consider that large platforms, even if they to do not actively engage in promoting such a behaviour may still be responsible for failing to terminate unethical or unlawful illegal activities of their users. It is true that US Digital Millennium Copyright Act and Communication Decency Act, or, to a lesser extent, the EU eCommerce directive, provide for immunities of ISPs relative to the unlawful behaviour by their users. However, such immunities have exceptions and there is a vast debate for limiting them, in particular in the context of the coming EU Digital Services Act [1]. A precondition for a clear understanding of the ethical and legal issues just mentioned is possessing precise notions of influence and control over others' behaviour, and of the connection between control and responsibility. We think that the notion of influence-based responsibility is important not only in the law, but also in the regulation of multi-agent systems. When agents become intelligent enough as to understand that an evil outcome can also be obtained through the action of others, it becomes necessary to prohibit not only harmful actions, but also bad influence, leading to the performance of harmful actions.

The issue of influence-based responsibility has so far not been addressed in the literature on the logic of actions and norms. This paper aims to cover this gap, by providing a logical analysis of the relationship between influence, control, and responsibility.

Our logical analysis will be based on the STIT logic of action [2], one of the most popular approaches to the study of agency. We will need to address an important logical and philosophical issue concerning the very definition of interpersonal influence. Namely, we shall consider how an agent i's influence inducing an agent j to perform an action φ may be consistent with j's choice to perform φ rather than not performing it. Clearly, the influence must not make it necessary for j to perform φ, as this would contradict the agency of j in realising φ. Preserving j's freedom of choice is indeed necessary for considering j the principal author of the unethical or illegal action for which also i may be liable. Consider, for instance the case of i managing a website devoted to revenge porn. If j publishes a video with such a content, he remains responsible for the unlawful behaviour consisting in publishing the video, but in addition i is also responsible for having enabled the publication. Or consider the case in which j asks k to publish the video in i's website. In this case k is responsible for publishing the video, j for having induced k to publish the video, and i for enabling the publication.

The article is organised as follows. Section 2 provides a gentle introduction to the STIT syntax and semantics. Section 3 discusses the concept of social influence from an informal perspective, while Section 4 addresses it from a formal point of view. Section 5 explores the connection between influence and control. Finally, Section 6 is devoted to the formalisation of the relationship between the concepts of causal and normative responsibility and the concepts of influence and control.

2. Background on STIT

STIT logic (the logic of *seeing to it that*) by Belnap et al. [2] is one of the most prominent formal accounts of agency. It is the logic of sentences of the form "the agent i sees to it that φ is true". Different semantics for STIT have been proposed in the literature (see, e.g., [2,3,4,5,6,7]). Following [6], here we adopt a Kripke-style semantics for STIT. The Kripke semantics of STIT is illustrated in Figure 1, where each moment m_1, m_2 and m_3 consists of a set of worlds represented by points. For example, moment m_1 consists of the set of worlds $\{w_1, w_2, w_3, w_4\}$. Moreover, for every moment there exists a set of histories passing through it, where a history is defined as a linearly ordered set of worlds. For example, the set of histories passing through moment m_1 is $\{h_1, h_2, h_3, h_4\}$. Finally, for every moment, there exists a partition which characterizes the set of available choices of agent 1 in this moment. For example, at moment m_1, agent 1 has two choices, namely $\{w_1, w_2\}$ and $\{w_3, w_4\}$. Note that an agent's set of choices at a certain moment can also be seen as a partition of the set of histories passing through this moment. For example, we can identify the choices available to agent 1 at m_1 with the two sets of histories $\{h_1, h_2\}$ and $\{h_3, h_4\}$.

In the Kripke semantics for STIT the concept of a world should be understood as a 'time point' and the equivalence class defining a moment should correspondingly be understood as a set of alternative concomitant 'time points'. In this sense, the concept of a moment captures a first aspect of indeterminism, as it represents the alternative ways the *present* could be. A second aspect of indeterminism is given by the fact that moments are related in a (tree-like) branching time structure. In this sense, the *future* could evolve in different ways from a given moment. In the Kripke semantics for STIT these two aspects of indeterminism are related, as illustrated in Figure 1. Indeed, if two distinct moments

m_2 and m_3 are in the future of moment m_1, then there are two distinct worlds in m_1 (w_1 and w_3) such that a successor of the former (w_5) is included in m_2 and a successor of the latter (w_7) is included in m_3.

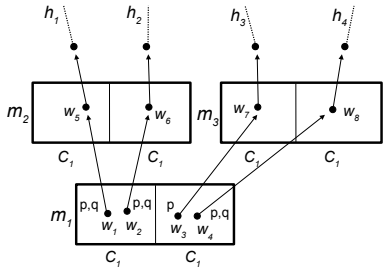

Figure 1. Illustration of Kripke semantics of STIT

The logic STIT allows us to talk about time. Specifically, existing STIT languages (see, e.g., [6,7]) include different kinds of future and past tense operators such the 'next' operator **X** (where $\mathbf{X}\varphi$ stands for "φ is going to be true in the next world") and the 'yesterday' operator **Y** (where $\mathbf{Y}\varphi$ stands for "φ was true in the previous world"). For example, the formula $\mathbf{X}\neg p$ is true at world w_1 in Figure 1. Indeed, at w_1 it is the case that p is going to be false in the next world (as p is false at world w_5). Moreover, the formula $\mathbf{Y}p$ is true at world w_5 since at world w_5 it is the case that p was true in the previous world (as p is true at world w_1).

The STIT language also includes an operator $\Box$ which allows us to represent those facts that are necessarily true, in the sense of being true at every point of a given moment or, equivalently, at every history passing through a given moment. For example, the formula $\Box p$ is true at world w_1 in Figure 1 since p is true at every point of moment m_1 including world w_1. The operator $\Diamond$ is the dual of $\Box$: it allows to represent those facts that are possibly true (i.e., true at some history passing through the actual moment).

Finally, the logic STIT provides for different concepts of agency, all characterized by the fact that an agent acts only if she sees to it that a certain state of affairs is the case. In this paper we shall use the deliberative STIT of [8] which is defined as follows: an agent i deliberately-sees-to-it that φ, denoted by formula $[i\ \mathbf{dstit}]\varphi$, at a certain world w if and only if: (i) for every world v, if w and v belong to the same choice of agent i then φ is true at v, and (ii) at w agent i could make a choice that does not necessarily ensure φ. Notice that the latter is equivalent to say that there exists a world v such that w and v belong to the same moment and φ is false at v. For example, in Figure 1, agent 1 deliberately sees to it that q at world w_1 because q is true both at world w_1 and at world w_2, while being false at world w_3. Deliberative STIT captures a fundamental aspect of agency, namely, the idea that for a state of affairs to be the consequence of an action (or for an action to be the cause of a state of affairs), it is not sufficient that the action ensures that the state of affairs holds, it is also required that, without the action, the state of affairs possibly would not hold (a similar idea is also included in the logic of "bringing it about" by Pörn, see in particular [9]). In this sense, $[i\ \mathbf{dstit}]\varphi$ at w is incompatible with the necessity of φ at w, since it requires that at w also $\neg\varphi$ was an open possibility.

3. The concept of social influence

Our analysis of social influence starts from a general view about the way rational agents make choices. Specifically, we assume that an agent might have several choices or alternatives *available* defining her *choice set* at a given moment, and that what the agent does is determined by her *actual* choice, which is in turn determined by the agent's *choice context* including her preferences and beliefs and the composition of her choice set. Our analysis of social influence expands this view by assuming that the agent's choice context determining the agent's actual choice might be determined by external causes. Specifically, the external conditions in which an agent finds herself or the other agents with whom the agent interacts may provide an input to the agent's decision-making process in such a way that a determinate action should follow. Note that here we only address the kind of influence that consists in *determining* the voluntary action of an agent by modifying her *choice context*, so that a different choice becomes preferable to the influencee on comparison to what would be her preferred option without this modification. This may happen, for instance: (a) by expanding the available choices (influence via choice set expansion), or (b)by restricting the available choices (influence via choice set restriction) or (c) by changing the payoffs associated to such choices, as when rewards or punishments are established (influence via payoff change).

To illustrate the concept of social influence, let us consider an example about influence via choice set restriction. The example is illustrated in Figure 2. It represents a situation where there are three objects to be purchased in an online marketplace, let us call them, for simplicity's sake, apple, banana and pear (they could be material objects, or stock items, etc.), and three bots acting for human buyers. The actions at issue consist in bringing about the purchase of the apple (*ap*), the banana (*ba*) or the pear (*pe*). Let us assume that agent 2 has certain preferences that remain constant along the tree structure. In particular, at all moments agent 2 prefers purchasing apples to bananas to pears. Let us also assume that 2 is rational, in the minimal sense that she acts in such a way as to achieve the outcome she prefers. Rational choices of agent 2 are depicted in grey. By choosing to purchase the apple at w_1, 1 generates a situation where, given its preferences, 2 will necessarily purchase the banana, rather than the pear. Indeed, although at moment m_2, 2 has two choices available, namely, the choice of purchasing the banana and the choice of purchasing the pear, only the former is rational, in the sense of being compatible with 2's preferences. In this sense, by deciding to purchase the apple at w_1 and removing this option from 2's choice set, 1 influences 2 to decide to purchase the banana at w_7.

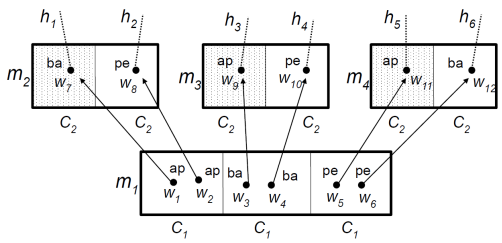

Figure 2. Example of influence via choice set restriction

As an example of influence via payoff change, consider the interaction between person 1 and bot 2, who is acting as an intermediary in an online marketplace, and has the goal of maximising the profits of its owner. Assume that 1 asks 2 to purchase for him an illegal drug (k) in exchange for a reward (r). The alternative choices for 2 are either purchasing the unlawful item or doing nothing. Let us assume that the bot has the goal of increasing its profits, so that 2 prefers purchasing the drug and getting the reward ($k \wedge r$) to not purchasing it and not being rewarded ($\neg k \wedge \neg r$). Moreover, the bot prefers the latter combination ($\neg k \wedge \neg r$) to purchasing the drug and not being rewarded ($k \wedge \neg r$), since the unlawful purchase of the drug involves the risk of a severe sanction. Under these assumptions, 1's promise of the reward at moment m_1 leads to a moment m_2 where agent 2 has the choice between producing $k \wedge r$ or $\neg k \wedge \neg r$, rather than to a moment m_3 where agent 2 would have had the choice between $k \wedge \neg r$ or $\neg k \wedge \neg r$ (no compensation been given for the action of purchasing the drug). Given 2's preferences, its rational choice at m_2 is to purchase the drug, while its rational choice at m_3 would have been not to purchase it. Thus we may say that at moment m_1 principal 1, by promising the reward, influences agent 2 to buy the unlawful item. An example of influence via choice set restriction is the case of an online platform disabling the posting of anonymous content. The option of refraining from publishing incitements to crime or terrorism becomes preferred by a user who starts to be deterred by the possibility of being identified and of facing the legal consequences of such a behaviour. These examples lead us to the following informal definition of social influence:

> An agent i influences another agent j *to perform a certain (voluntary) action* if and only if, i sees to it that every rational/preferred choice of j will lead j to perform the action.

We distinguish influence on action from influence on inaction.

> An agent i influences another agent j *to abstain from performing a certain (voluntary) action* if and only if, i sees to it that every rational/preferred choice of j will lead j to not perform the action.

For instance, the above example of a person who asks a bot to buy an unlawful drug fits the definition of influence on action, whereas the above example of the online platform disabling the posting of anonymous content fits the definition of influence on inaction. In the next two sections we move from an informal to a formal perspective on the concept of social influence and its relationship with the concept of controllability.

4. Formalization of social influence

In [10], we provided an analysis, based on STIT logic, of the concept of influence on action discussed in the previous section. To this aim, we extended STIT logic with special 'rational' STIT operators of the form $[i\ \mathbf{rdstit}]$. We are going to resume this analysis and extend it by the concept of influence on inaction.

The formula $[i\ \mathbf{rdstit}]\varphi$ has to be read "if agent i's current action is the result of a rational choice of i, then i deliberately sees to it that φ". We adopt a minimal concept of rationality, which is sufficient for our purpose: we assume that the choices of an agent

are ranked according to the agent's preferences, and an agent is rational as long as she implements her preferred choices. The $[i$ **rdstit**$]$ operator is interpreted relatively to STIT branching time structures, like the ones represented in Figure 2. Specifically, the formula $[i$ **rdstit**$]\varphi$ is true at a certain world w if and only if, *if* the actual choice to which world w_1 belongs is a rational choice of agent i *then*, at world w_1 agent i deliberately sees to it that φ, in the sense of deliberative STIT discussed in Section 2. For example, at the world w_7 in Figure 2, the formula $[2$ **rdstit**$]ba$ is true since the actual choice to which world w_7 belongs is a rational choice of agent 2 *and* at w_7 agent 2 deliberately sees to it that ba is the case. To capture the idea of influence on action, we introduce a social influence operator based on the concept of deliberative STIT (see [10]) defining it as follows:

$$[i \textbf{ inflAct } j]\varphi \stackrel{\text{def}}{=} [i \textbf{ dstit}]\mathbf{X}[j \textbf{ rdstit}]\varphi. \tag{1}$$

In other words, we shall say that an agent i influences another agent j to make φ true, denoted by $[i$ **inflAct** $j]\varphi$, if and only if, i deliberately sees to it that if agent j's current choice is rational then j is going to deliberately see to it that φ. The reason why the operator $[i$ **dstit**$]$ is followed by the temporal operator $\mathbf{X}$ is that influence requires that the influencer's choice precedes the influencee's action. On the contrary, we do not require that $[j$ **rdstit**$]$ is followed by $\mathbf{X}$ since in STIT the concept of action is simply captured by the deliberative STIT operator which does not need to be followed by temporal modalities.

In order to illustrate the meaning of the influence operator, let us go back to the example of Figure 2. Since agent 2 prefers bananas to pears, her only rational choice at moment m_2 is $\{w_7\}$. From this assumption, it follows that formula $[1$ **inflAct** $2]ba$ is true at world w_1. Indeed, at world w_1 agent 1 deliberately sees to it that, in the next world, if agent 2's choice is rational then 2 deliberately sees to it that ba is the case.

As emphasized above influence on action should be distinguished from influence on inaction. The latter concept is captured by the following abbreviation:

$$[i \textbf{ inflInact } j]\varphi \stackrel{\text{def}}{=} [i \textbf{ dstit}]\mathbf{X}(rat_j \rightarrow \neg[j \textbf{ dstit}]\varphi). \tag{2}$$

This means that an agent i influences another agent j to abstain from making φ true, denoted by $[i$ **inflInact** $j]\varphi$, if and only if, i's current choice guarantees that j will not be able to rationally see to it that φ. In other words, i's current choice will exclude the possibility that j will make φ true, if j will choose in conformity with his preferences. The constant symbol rat_j means that agent j's current choice is rational. It is an abbreviation, adopted for notational convenience, of $\neg[j$ **rdstit**$]\bot$, a formula that is satisfied only when j chooses rationally in the current word.

5. From influence to control

The notion of influence we discussed in the previous section is the key element of the notion of control on which the present analysis is focused.

In order to formalize this concept, we have to assume there are a set of conditional positive values (+values) $I^+ = \{(\varphi_1,\psi_1),\ldots,(\varphi_k,\psi_k)\}$ and a set of conditional negative values (-values) $I^- = \{(\varphi'_1,\psi'_1),\ldots,(\varphi'_h,\psi'_h)\}$. Specifically, $(\varphi,\psi) \in I^+$ means that the

occurrence of ψ should be promoted when φ is true and $(\varphi',\psi') \in I^-$ means that the occurrence of ψ should be hindered when φ is true.

A value i is active when its antecedent condition is satisfied. The following definition captures the concept of active value for $X \subseteq I^+$ and $Y \subseteq I^-$:

$$\textbf{Active}(X,Y) \stackrel{\text{def}}{=} \bigwedge_{(\varphi,\psi)\in X} \varphi \wedge \bigwedge_{(\varphi',\psi')\in Y} \varphi'. \tag{3}$$

Thus, for a conditional value (φ,ψ) in I^+ or I^- to be active, its triggering condition φ must be true.

As the following definition highlights, for an agent i to have control over another agent j, with respect to a set of +values X and a set of -values Y, i should be capable of influencing j to realise every active +value in X and to abstain from realising any active -value in Y. Thus, for $X \subseteq I^+$ and $Y \subseteq I^-$, positive control may be defined as follows:

$$\textbf{Ctrl}(i,j,X,Y) \stackrel{\text{def}}{=} \textbf{Active}(X,Y) \rightarrow \Diamond\Big(\bigwedge_{(\varphi,\psi)\in X} ([i \textbf{ inflAct } j]\psi) \wedge \bigwedge_{(\varphi',\psi')\in Y} ([i \textbf{ inflInact } j]\psi')\Big). \tag{4}$$

The previous notion of control is different from the notion of organizational control formalized by Grossi et al. [11]. While their notion is a primitive and is assigned to roles, our notion is assigned to agents and, more importantly, is defined from the more basic notions of action and capability.

Let us apply this notion to our online example. Let us assume that the following are the -values in I^-: publishing hate speech $(hate, publish)$ and spreading malware $(malware, spread)$. Let us also assume that the following is the unique +value in I^+: flagging or tagging fake news $(fake, tagged)$. Let us assume that the online platform under consideration has the capability of influencing the publisher to perform the good action and to abstain from the bad ones. The latter is achieved by giving to the publisher the right balance of incentives and disincentives. This assumption is formally expressed as follows:

$$\textbf{Ctrl}(i,j,\{(fake,tagged)\},\{(hate,publish),(malware,spread)\}). \tag{5}$$

6. From control to responsibility

We now move to the notion of control-responsibility, namely, the responsibility that results on failing to exercise control. An agent having control over an agent relative to the achievement of some +values X and some -values Y has secondary responsibility if he does not exercise the influence as needed, i.e., if the influencee fails to realise active +values, or to realise active -values:

$$\textbf{CtrlResp}(i,j,X,Y) \stackrel{\text{def}}{=} \textbf{Ctrl}(i,j,X,Y) \wedge \textbf{Active}(X,Y) \wedge \Big(\bigvee_{(\varphi,\psi)\in X} (\neg[i \textbf{ inflAct } j]\psi) \vee \bigvee_{(\varphi',\psi')\in Y} (\neg[i \textbf{ inflInact } j]\psi')\Big). \tag{6}$$

This notion of secondary responsibility, however, raises a problem. What if the potential infuencer cannot exercise influence relatively to all values at stake. For instance, what if the provider cannot prevent hate speech and fake news without also blocking useful content (being unable to distinguish in all cases hate speech and fakes from decent and sincere communication)?

A possible solution can come from what jurists call "proportionality", i.e., by considering the relative importance of the (sets of) values being implemented. Let us assume a strict preference ordering $\succ$ over pairs (X,Y) such that $X \in 2^{I^+}$ and $Y \in 2^{I^-}$. Let us write $(X',Y') \succ (X,Y)$ to mean that it is better to realise the +values X' and refrain from realising the -values Y', than to realise the +values X and to refrain from realising the -values Y.[1] Then, i would be normatively responsible for failing to exercise control over j relative to (X,Y) only if there is no (X',Y') that is preferable to (X,Y), such that if i exercises control over j relative to (X',Y'), i cannot at the same time exercise control over (X',Y'). The notion of exercising control can be defined as follows:

$$\mathbf{ExCtrl}(i,j,X,Y) \stackrel{\text{def}}{=} \mathbf{Active}(X,Y) \wedge \Big(\bigwedge_{(\varphi,\psi)\in X} ([i\ \mathbf{inflAct}\ j]\psi) \wedge \bigwedge_{(\varphi',\psi')\in Y} ([i\ \mathbf{inflInact}\ j]\psi') \Big). \tag{7}$$

Note that $\mathbf{ExCtrl}(i,j,X,Y)$ implies $\mathbf{Ctrl}(i,j,X,Y)$, since $\varphi \rightarrow \Diamond\varphi$ is valid.

This leads us to our final notion of normative responsibility. Agent i is normatively responsible for not exercising control over agent j relative to set of positive values X and the set of negative values Y if i is causally responsible for that, and her failure to exercise control is not due to the need to realise more important positive/negative values. The latter would be the case if there was a preferable pair (X',Y'), such that i exercises control over(X',Y') being unable to exercise control over all values in both pairs, i.e., over $(X \cup X', Y \cup Y')$:

$$\mathbf{NormCtrlResp}(i,j,X,Y) \stackrel{\text{def}}{=} \mathbf{CtrlResp}(i,j,X,Y) \wedge \neg \bigvee_{(X',Y')\succ(X,Y)} \Big(\mathbf{ExCtrl}(i,j,X',Y') \wedge \neg \mathbf{Ctrl}(i,j,X\cup X', Y\cup Y') \Big). \tag{8}$$

To illustrate the previous notion of normative responsibility, let us go back to the example of the online platform. Suppose all values in our example are active and that the provider can either (i) induce the publisher to tag fake news, or (ii) induce the publisher to abstain from spreading malware and from publishing hate speech. Finally, suppose it is impossible for the provider to jointly realise (i) and (ii), since it cannot prevent the publisher to publish hate speech unless it disables the tagging functionality. But (ceteris paribus) it is more important to realise (ii) than to realise (i). Then, the provider would not be considered normatively responsible for not guaranteeing (i) to be true if it guarantees (ii) to be true. The opposite would be the case if the preference was inverted.

[1] We can safely assume that the preference ordering $\succ$ is induced by a utility function $U : (I^+ \cup I^-) \longrightarrow \mathbb{R}^+$ measuring the degree of importance of a positive/negative value such that $(X',Y') \succ (X,Y)$ if and only if $\sum_{(\varphi',\psi')\in X'\cup Y'} U(\varphi,\psi) > \sum_{(\varphi,\psi)\in X\cup Y} U(\varphi,\psi)$.

7. Conclusion

In this paper we have modelled influence by using the framework provided by the STIT logic of action. On this basis we have defined a notion of control, as the possibility to exercise influence over another, to induce the influencee to realise positive values and refrain from realising negative values. An agent's failure to exercise control can be viewed as control-responsibility, namely as causal responsibility for the action of the potential influencee. We have then argued that causal responsibility for failing to exercise influence relative to certain values does not lead to normative control-responsibility where control has been exercised to achieve superior incompatible values. This has lead us to the final notion of normative control-responsibility.

The issues we have tackled are largely unexplored, as while primary responsibility has been addressed in STIT by [3,12], no account is yet available of secondary responsibility, understood as control-responsibility. [13] proposed to model social influence in the framework of the "bringing it about that" logic, but have not addressed the connection between influence and choice, focusing on influence through the exercise of normative powers. [14] modelled an agent's endorsement of the principal's goals as a source of responsibility for the principal, but did not consider how goal-alignment is put in place. This work is still preliminary and partial, as it only addresses some kinds of influence/control and some aspects of these notions. We plan to develop it further, covering both active and omissive influence, and addressing psychological influence, which changes the influencees' cognitive states (their beliefs and preferences) rather than the context of their action.

References

[1] Sartor G. The secondary liability of online intermediaries. In: Research Handbook on EU Media Law and Policy. Elgar; 2021. p. 141–65.

[2] Belnap N, Perloff M, Xu M, Bartha P. Facing the future: agents and choices in our indeterminist world. Oxford University Press; 2001.

[3] Broersen J. Deontic epistemic *stit* logic distinguishing modes of mens rea. Jan. 2011;9:137–52.

[4] Wolfl S. Propositional Q-Logic. Journal of Philosophical Logic. 2002;31:387–414.

[5] Lorini E, Schwarzentruber F. A logic for reasoning about counterfactual emotions. Artificial Intelligence. 2011;175:814–47.

[6] Lorini E. Temporal STIT logic and its application to normative reasoning. Journal of Applied Non-Classical Logics. 2013;vol. 23:pp. 372–399.

[7] Schwarzentruber F. Complexity Results of STIT Fragments. Studia Logica. 2012;100(5):1001–1045.

[8] Horty JF, Belnap N. The Deliberative STIT: A Study of Action, Omission, Ability, and Obligation. Journal of Philosophical Logic. 1995:583–644.

[9] Pörn I. On the Nature of Social Order. In: Fenstad JE, Frolov IT, Hilpinen R, editors. Logic, Metodology and Philosophy of Science. Vol. 8. North Holland; 1989. p. 553–67.

[10] Lorini E, Sartor G. A STIT Logic for Reasoning About Social Influence. Studia Logica. 2016;104(4):773–812.

[11] Grossi D, Royakkers LMM, Dignum F. Organizational structure and responsibility. Artificial Intelligence and Law. 2007;15(3):223–249.

[12] Lorini E, Longin D, Mayor E. A logical analysis of responsibility attribution: emotions, individuals and collectives. Journal of Logic and Computation. 2014;24(6):1313–1339.

[13] Santos FAA, Jones AJ, Carmo J. Action Concepts for Describing Organised Interaction. In: Thirtieth Annual Hawaii International Conference on System Sciences. IEEE Computer Society; 1997. p. 373–82.

[14] Smith C, Rotolo A, Sartor G. Reflex Responsibility of Agents. In: Proceeding of JURIX 2013: The Twenty-Sixth Annual Conference on Legal Knowledge and Information Systems. IOS; 2013. p. 135–44.