Social Value Orientation and Integral Emotions in Multi-Agent Systems

Daniel E. Collins(✉), Conor Houghton, and Nirav Ajmeri

Department of Computer Science, University of Bristol, Bristol, UK
{daniel.collins,conor.houghton,nirav.ajmeri}@bristol.ac.uk

Abstract. Human social behaviour is influenced by individual differences in social preferences. Social value orientation (SVO) is a measurable personality trait which indicates the relative importance an individual places on their own and on others' welfare when making decisions. SVO and other individual difference variables are strong predictors of human behaviour and social outcomes. However, there are transient changes in human behaviour associated with emotions that are not captured by individual differences alone. Integral emotions, the emotions which arise in direct response to a decision making scenario, have been linked to temporary shifts in decision making preferences. In this work, we investigated the effects of modifying social preferences according to transient integral emotions in multi-agent societies. We developed *Svoie*, a method for designing agents that make decisions based on established SVO policies, as well as alternative integral emotion policies in response to task outcomes. We conducted simulation experiments in a resource-sharing task environment, and compared societies of *Svoie* agents with societies of agents with fixed SVO policies. We find that societies of agents that adapt their behaviour through integral emotions achieved similar collective welfare to societies of agents with fixed SVO policies, but with significantly reduced inequality in welfare among agents with different SVO traits. We observed that by allowing agents to adapt their policy in response to task outcomes, our agent societies achieved reduced social inequality.

Keywords: Individual differences · Social decision making · Simulation

1 Introduction

Social value orientation (SVO) is a spectrum of personality traits that describes individual differences in social preferences, in terms of the relative value an agent places on its own welfare and the welfare of others when making decisions [33,34]. The SVO spectrum includes agents who are: *altruistic* or caring only for others, *cooperative* or caring both for self and others, and *selfish* or caring only for self. SVO is measurable in humans and considered to be relatively stable over time. Further, SVO has been found to be strongly correlated with patterns of social

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
N. Fornara et al. (Eds.): COINE 2023, LNAI 14002, pp. 118–138, 2023.
https://doi.org/10.1007/978-3-031-49133-7_7

behaviour through empirical study, such as the tendency to act cooperatively or individualistically [3,4].

Seminal works from social psychology provide a clear conceptual model of the influence of SVO on individual preferences in social interactions. A robust framework for agent simulation has been developed, the ring model [22,31], which defines utility functions for SVO traits that are now standard in multi-agent research. In a social dilemma, a rational agent would be expected to make decisions that maximise the utility associated with their individual preferences. However, humans are not rational agents, and they will not always seek optimal outcomes that would be expected for their stable characteristics. Through empirical studies, patterns of irrational decision making in humans have been linked to transient changes in affective state, emotions and mood states, resulting from changes in immediate circumstance or environment, or the consequence of longer-term contingencies or goals that interact with the current task. Emotions may serve an important role in adaptive decision making by motivating and guiding behaviour based on observations and judgements about the current context of the decision making environment and other within it.

Lerner et al. [30] outline two main categories of emotion, *incidental* and *integral*. *Incidental emotions* are task-unrelated emotions that arise in response to factors that are irrelevant to the current decision scenario, but which are nevertheless present during the decision making process. For example, a person who receives a frustrating message from a friend before an important meeting at work may be influenced by the unpleasant emotions during the meeting, even though they are task-unrelated, that could lead to impulsive decision making or unnecessary conflicts with colleagues. *Integral emotions* are task-related emotions that arise in direct response to the current decision, and are known to have a strong influence on behaviour. Integral emotions can be either *anticipated*, feelings about a potential future event or the possible outcome of an action, or *immediate*, feelings about a recent event or the observed outcome of an action. Our interest is in the latter, for example, the immediate integral emotion of feeling satisfied after performing well on an exam, and choosing to spend time helping others with their studies.

The "wounded pride" model of integral emotion [52] suggests that agents may react to unfair outcomes by feeling negative emotions, and acting spitefully, even when they know that it will result in a worse outcome for themselves on that specific task [40]. This is an example of how integral emotions can give rise to behaviour that is not explained by individual differences alone. Agents that adapt their policies based on integral emotion as in the wounded pride model may fare better than agents that only act based on SVOs, since some SVO policies may perform poorly on a given task compared to others. In this work, we investigated whether socially beneficial effects of altering social preferences according to integral emotions could be observed by modelling integral emotions in multi-agent societies with individual differences in SVOs.

Contributions. We developed *Svoie*, a method for designing agents that make decisions based on SVO and integral emotions. Our *Svoie* agents combine well-

established SVO decision making policies with a simple protocol for temporarily adopting alternative policies based on integral emotion. We define two alternative social-preference-based policies representing *positive* and *negative* emotions, that minimise or maximise payoff inequity respectively. These policies incorporate the wounded pride model of spiteful human decision making, and an idealised counter model for positive integral emotion. We model integral emotion as an internal state, that changes depending on the outcomes of recent decisions, and that defines the probability that an agent will adopt an integral-emotion-based policy in their next decision.

Findings. To evaluate *Svoie*, we conducted simulation experiments using a variant of the Colored Trails game [16,18], a resource-sharing task environment designed for studying social decision making. We generated societies of agents with heterogeneous SVOs, and simulated sequences of games between random pairs of agents in the society. We compare the distribution of payoffs accumulated by agents between *Svoie* and *Stable-SVO* societies, and evaluate societal outcomes in terms of collective welfare, a measure of the total payoff to all agents in a society, and welfare inequality, a measure of the variation of payoff between agents.

We investigated whether *Svoie* societies would have lower welfare inequality relative to *Stable-SVO* societies, by allowing agents to adapt their social preferences based on the frequency with which they are succeeding or failing to achieve their individual goals. We find that societies of *Svoie* agents exhibit significantly lower welfare inequality than *Stable-SVO* agents in societies with more than one SVO, with a small reduction in collective welfare.

Organisation. Section 2 describes preliminaries necessary to understand our contribution. Section 3 describes our method for modelling SVO and integral emotions in agents. Section 4 presents our experimental setup, results, and evaluation. Section 5 concludes with a discussion of future directions.

2 Preliminaries and Related Works

We now introduce the preliminaries necessary to understand our contributions.

2.1 Social Value Orientation

The SVO model describes a continuum of orientation types, reflecting the nature of social preferences in decision making [33,34]. SVOs are used in agent-based simulation to define agent decision making policies. SVO policies are typically implemented using the ring model of SVO [31]. In this model, an SVO utility function can be defined by any point on a unit circle, where the extent of preference for reward to self and to others is mapped to the x and y axes respectively. For example, this spectrum includes:

Altruistic Preference to take actions that increase the welfare of others, regardless of their own welfare.

Cooperative Preference to work with others to increases the welfare of themselves as well as others.

Selfish Preference to take actions that increase the welfare of themselves, regardless of the welfare of others.

These three SVOs types cover the positive quadrant of the ring model, in which SVO utility functions only consider positive preferences for reward to self, other or both. The complete spectrum of SVO traits also includes negative preferences, for example, competitive agents have a preference for increasing their own reward while also reducing the reward of others. Different SVO decision making policies are well defined, and give predictable differences in performance in simulated social task environments [22]. The relative performance of SVO policies depends on the nature of the task.

Social preferences have been explored in the context of developing autonomous agents for applications in various real-world domains such as cyber-security [26], and SVO has been utilised to simulate social behaviour in autonomous vehicle decision-making [7,12,42]. Multi-agent simulation incorporating SVO has been used alongside experimental data to better understand how individual differences can influence cognition and behaviour to benefit societies, e.g., through social cooperation [2] and adapting to changes in environment [47], and SVO has been used in the simulation of normative multi-agent systems to understand the emergence of prosocial and cooperative behaviour [46]. Related works have looked at agent-based modelling of other individual difference variables, such as Myers-Briggs personality types [6]. In this work, we aim to better understand the relationship between emotion and social preferences through agent-based simulation.

2.2 Integral Emotions

Integral emotions describe task-related emotions that are directly influenced by the current decision making process, for example, an individual may experience positive or negative integral emotions depending on whether they achieve their goal on a particular task [52].

Seminal works in psychology shed light on the influence of integral emotions on human behaviour through empirical studies using ultimatum games [19]. An ultimatum game between two agents, Alice and Bob, can be described as follows: Alice and Bob are in separate rooms. Alice is told that Bob has been given an amount of money, and has been asked to share some of this money with Alice. Bob can offer any portion of the money to Alice that they choose. Alice can either accept this offer, or reject it. If Alice rejects the offer, neither Alice nor Bob receive any of the money, hence Bob's offer is an ultimatum.

A key finding of early work on ultimatum games is that people often reject small amounts of money despite the fact that this results in a worse outcome for themselves—they are rejecting "free money". This finding has been replicated in numerous studies [51]. This may be thought of as a calculated spiteful behaviour, e.g., paying a cost in order to harm another. Emotional reactions like spite may

be considered in the context of social norms, pervasive expectations of certain behaviours within societies. Spiteful actions, in which a cost is paid to punish a perceived wrongdoer, may be adaptive behaviours which encourage cooperation norms, by enforcing sanctions in the form of punishments when cooperation norms are violated [39]. This could be extended to any norm related to how an individual expects that others should behave in a society, regardless of how they do. If an individual has a strong expectation for a particular norm, they may experience negative emotions when that norm is violated, and respond with spiteful actions. behaviour of this nature is common in online communities, for example, in commenting behaviours on the website Stack Overflow, [9].

The perspective of emotions as norm enforcing mechanisms is complicated by observations from ultimatum game experiments which show that spiteful behaviour may arise in the absence of any perceived social injustice, in the absence of any punishable perpetrator, and that once triggered, spiteful behaviour may be sustained and subsequently directed towards others arbitrarily. By altering the set-up of the ultimatum game, Straub and Murnigham [44] observed that participants sometimes rejected small offers even if they did not know the total amount of money from which the offer had been made, suggesting the rejection is not motivated by a sense of social inequity between participants. Further, they found that participants were just as likely to reject small offers when they did not know that the money had been split by another participant. They hypothesised that offers of small amounts of money were rejected because they evoked feelings of wounded pride, a direct emotional response to an unsatisfactory outcome. Pillutla et al. [40] conducted experiments using a sequence of ultimatum games between different pairs of participants, and found that participants who spitefully rejected a small offer would be more likely to take spiteful actions in subsequent games against new participants. In ultimatum games, individuals who receive an unsatisfactory offer may still try to act in retaliation, even if they cannot cause a disadvantage to the proposer of the unfair offer, suggesting that spiteful actions are a form of emotional release, or an expressions of internalised emotions [50]. The emotion may arise due to norm violation, but the resulting action may not be a calculated effort to enforce that same norm. More recently, Criado et al. [11] have explored role of emotions as motivators for norm compliant decision making towards the development of autonomous agents act in accordance with human norms.

These works describe a model of wounded pride, in which undesirable task outcomes can provoke a strong negative emotional response, which is expressed through subsequent non-cooperative behaviour. If an agent perceives that an outcome is unfair and unduly negative to them or contrary to an expectation of self-worth, feelings of wounded pride and anger are aroused which will influence their subsequent actions even if those actions cannot lead to a redress of the perceived wrong. In other words, when an individual experiences negative emotions in response to an unsatisfactory outcome, but cannot directly express these emotions to some perceived wrongdoer, they are nevertheless willing to retaliate by making sub-optimal decisions, which disadvantage others at some cost

to themselves. This mechanism may be beneficial in protecting altruistic agents from being repeatedly taken advantage of by self motivated agents. Conversely, we can conceive of a counter mechanism to wounded pride, wherein disproportionate success may elicit positive emotions, which in turn influence an agent to temporarily relax their preferences for high payoff and promote generosity. This aligns with ideas from social psychology on behaviour changes associated with positive emotion [21,48].

A common method of monitoring integral emotions in human studies is through self reporting of emotion valence, the degree of positive or negative feelings at a particular moment in time. This derives from the appraisal theory of emotion [36], which posits that human emotions are internal phenomena, constructed through the appraisal of external events and stimuli, for example by evaluating whether an event outcome aligns with personal goals or norm expectations. Valence has been used in autonomous agent research to define internal states related to emotions, for example, to define intrinsic rewards for guiding the behaviour of reinforcement learning agents [20], and as a component of comprehensive decision making architectures based on psychological theories [13]. These related works often make use of other components of appraisal theory, such as arousal and motivation. For simplicity, we will focus on the valence of integral emotion associated with task outcomes. A similar approach has been taken previously to investigate the relationship between emotions and behavioural norms [35].

2.3 Social Task Settings

Simulations of agent behaviour in game environments can be directly compared to human decision-making data on the same or similar tasks or used as an abstraction of complex real-world social decision-making scenarios. In stochastic games, random variations in the parameters of the game's setup and the agents involved in the game can give rise to a variety of different emergent scenarios. Sequences of stochastic games of varying complexity have been used to approximate complex real-world task environments for studying the influence of emotion and social factors on behaviour, both in empirical human studies and agent simulation [8,12]. There is a breadth of work in which stochastic games have been used to study the relationship between SVO and social behaviour [3]. Stochastic games have also been used to study how emotions influence behaviour. Bono et al. [5] use a stochastic resource-allocation game to study how emotions mediate SVO preferences in human decision making.

Colored Trails (CT) [16,18] is a research test bed designed for studying social factors in decision making. In CT, agents enter into a negotiation [27] and exchange resources to achieve their own individual goals. CT can be described in terms of generic elements of the task setting:

- Agents have individual goals they try to bring about.
- Agents have individual resources they can use to bring about their individual goals.

- Agents receive a reward upon achieving their goals.
- Individual circumstances of agents may vary, and therefore they may require different resources to bring about their goals compared to their peers.
- Agents may have insufficient resources to bring about their goals, or they may have surplus resources.
- Agents may negotiate an exchange of resources to help each other reach their goals.

CT is a highly flexible and expressive stochastic game, with various parameters that can be modified to customise the task environment. We chose to adopt CT as an environment for evaluating our agent societies, as it benefits from a clear task setting, and the random elements in the game's set-up allow agents to encounter different unique social tasks over a sequence of games [23].

3 Method

We now detail our implementation of the CT game environment, agent decision-making policies, and agent models.

3.1 Simulation Environment

We implemented a simplified version of CT as a simulation environment for studying *Svoie* and *Stable-SVO* agent societies, based on an existing Python implementation from Sloan and Ajmeri [43].

A game of CT is played between two agents, and it consists of two separate rounds. At the start of each game round, a new game-board is generated: a 4×4 grid of coloured tiles, where each tile is randomly assigned one of four possible colours (red, blue, green, yellow). Each agent is then placed on the game-board at separate random starting positions. A random goal position is then assigned on the game-board, which is not vertically or horizontally adjacent to either agent's starting positions. At the start of each round, each agent is allocated resources—a set of four randomly coloured chips—that agents can place to move to an adjacent position on the board where the chip colour matches the tile colour. The objective of the game is to get as close as possible to the goal position using the allocated resources. We assume agents have access to full information about the state of the game, e.g., the game-board, agent positions, goal position, and the resources of both agents.

Once per round, the agents may negotiate and exchange some or all of their resources to help each other reach the goal. During negotiation, one agent takes the role of *Proposer* and the other takes the role of *Responder*. The *Proposer* sends a proposal to the *Responder* comprising an offer, chips they will send from their own inventory, and a request, chips they want to receive from their opponents inventory. The *Responder* can then either accept the proposal, initiating the proposed exchange, or decline the proposal, meaning there is no-exchange and both players are left with their original allocated resources. Agents can then

use their resources to move as close as possible to the goal position, and receive a score, S, at the end of the round:

$$S = n + 1.5u(1 + g) \tag{1}$$

where n is the number of unused chips remaining in the agents inventory, u is the number of tile-chips used to create a path and g is equal to 0 or 1 depending on whether or not the agent reached the goal position respectively. This scoring function is taken from [24], and is designed to prioritise goal achievement strategies over strategies which seek to maximise score by gathering tiles, or creating long paths to arbitrary positions. Agents switch their negotiation roles between the two rounds of the game, so that each agent has one round as *Proposer* and one round as *Responder*.

By only allowing one offer and response per game round, CT becomes a more expressive form of the traditional ultimatum game discussed in Sect. 2. Here, the *Responder* can only choose between two possible outcomes: the ultimatum offer sent by the *Proposer*, or the no-exchange outcome determined by the randomised parameters of the game set-up. Random variations in individual circumstances and individual goals are encoded in CT through random variations in game-board set up, resource allocation, starting positions and goal positions.

Figure 1 shows a schematic example of one possible CT set-up, demonstrating how agents can cooperate to achieve a greater reward.

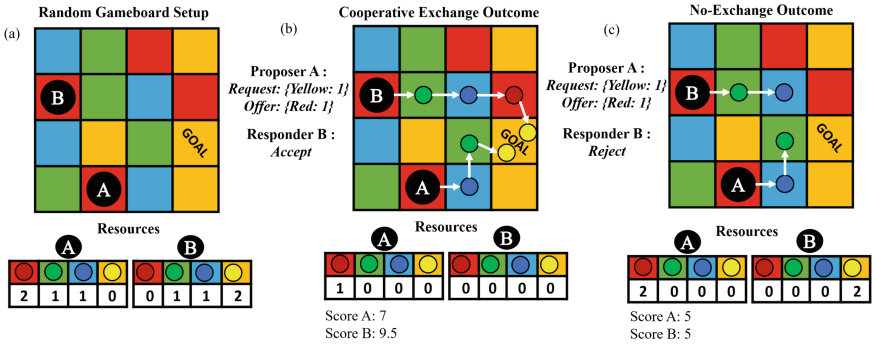


Fig. 1. Schematic example of one round of CT between agents A and B. **a** Random game-board setup parameters are generated at the start of the game: coloured tiles, agent positions, goal position and allocated resources. In CT, the resources are coloured chips that agents can use to move to an adjacent tile with the same colour. In this illustrated setup, neither agent can reach the goal using their initial resources. **b** A possible game outcome is shown, where B has agreed to A's mutually beneficial exchange proposal; A sends one red chip to B, and B sends one yellow chip to A. Agents then use their resources to reach the goal, and receive a score according to Eq. 1. **c** Alternatively, in the no-exchange outcome, B chooses to reject A's proposal, and agents must move as close to the goal as they can with their initial resources. Here, this results in a lower score for both agents.

3.2 Utility Functions for Social Preferences

To design agent decision-making protocol for the CT environment, social preferences and possible actions were mapped to quantitative utility functions. In each case, the agent perceives their environment, and uses the available information to select an action. An action is selected if it is expected to maximise the utility associated with the agents social preferences, a function of the game scores expected to result from an action, calculated using the scoring function in Eq. 1. The way in which an agent uses the utility function depends on whether it is acting as a *Proposer* or *Responder*.

Let, x, be an arbitrary exchange outcome, e.g., the resources that each agent possesses after the an exchange. If we assume that an agent will always use their chips optimally to achieve the highest possible score, each exchange outcome x maps directly to a pair of scores $S - P(x)$ and $S - R(x)$ for the *Proposer* and *Responder* respectively for a given game set-up. We can therefore define our utility functions in terms of x.

An agent acting as *Proposer* uses a chosen utility function as a ranking criteria to select a proposal. The *Proposer* calculates the utility associated with each exchange outcome, x, from the set of all possible exchange-outcomes, X, then selects the outcome with the greatest utility, and sends the corresponding proposal that would result in that outcome if accepted by the *Responder*. A *Responder* will accept a proposal only if it maximises a utility-based acceptance criteria relative to the no-trade outcome $\bar{x}$, the random set of resources possessed by each agent if no-exchange takes place. Here, the expected score for the no-trade outcome, $\bar{x}$, can be denoted $S - P(\bar{x})$ and $S - R(\bar{x})$. A proposal is only accepted if the utility of the proposed exchange is greater than the utility of the no-exchange outcome for the *Responder*.

Utility functions for socially oriented decision-making protocols are outlined for CT [15,17] based on different social preferences. We adapted these utility functions to describe agent protocols for out implementation of CT:

Individual Benefit the utility is the proposer score

$$U - r(x) = S - R(x) \tag{2}$$

or the responder score.

$$U - p(x) = S - P(x) \tag{3}$$

Aggregate Benefit the utility is the cooperative score, the sum of the proposer and responder scores.

$$U - c(x) = S - P(x) + S - R(x) \tag{4}$$

Outcome Fairness (Advantage of Outcome) the utility is the advantage achieved by the responder.

$$U - a(x) = S - R(x) - S - P(x) \tag{5}$$

Trade Fairness (Advantage of Trade) the utility is the advantage achieved by the responder, relative to rejection.

$$U - f(x, \bar{x}) = (S - R(x) - S - R(\bar{x})) - (S - P(x) - S - P(\bar{x})) \quad (6)$$

It is important to note that these functions are written from the perspective of the *Responder* so that they are positive when the action benefits the *Responder*. When used by the *Proposer*, the subscripts P and R are switched.

3.3 Agent Decision-Making Policies

In this section, we adapt the social-preference-based utility functions outlined in Sect. 3.2 to construct decision-making policies corresponding with altruistic, selfish and cooperative SVOs, and positive and negative integral emotions. We use these policies to develop baseline *Stable-SVO* agents, which always make decisions according to a fixed SVO-based policy, and *Svoie* agents, which act according to an SVO-based policy by default, but may temporarily adopt an integral-emotion-based policy in response to game outcomes in CT.

SVO Policies. Baseline *Stable-SVO* agents were created such that each agent has one of three possible SVO traits: selfish, altruistic or cooperative. Each SVO describes a fixed decision-making policy with a utility function reflecting social outcome preferences.

Selfish A selfish agent takes actions which maximise their own payoff.

- Proposal Ranking Criteria:

$$maximise U - p(x) \quad (7)$$

- Response Acceptance Criteria:

$$accept\ trade\ if\ and\ only\ if\colon U - p(x) > U - p(\bar{x}) \quad (8)$$

Cooperative A cooperative agent takes actions which maximise mutual payoff.

- Proposal Ranking Criteria:

$$maximise U - c(x) \quad (9)$$

- Response Acceptance Criteria:

$$accept\ trade\ if\ and\ only\ if\colon U - c(x) > U - c(\bar{x}) \quad (10)$$

Altruistic An altruistic agent takes actions which maximise payoff to others.

- Proposal Ranking Criteria:

$$maximise U - r(x) \quad (11)$$

- Response Acceptance Criteria:

$$accept\ trade\ if\ and\ only\ if\colon U - r(x) > U - r(\bar{x}) \quad (12)$$

Integral Emotion Policies. We devise two integral emotions policies to capture temporary changes in social preferences resulting from positive or negative integral emotions. Here, the integral emotion policies describe social outcome preferences that are not captured in *Stable-SVO* policies. The negative emotion policy, *competitive equity aversion*, is one which is expected to result in achieving a higher score with the largest margin of difference between the agent and its opponent ("Advantage of Outcome") or "unfair" proposal). Conversely, the positive emotion policy, *inequity aversion*, is one which will minimise the margin of difference between the resulting scores. These are distinct from SVO policies as they do not consider game score maximisation.

Positive Integral Emotion (Inequity Aversion). An agent with positive integral emotion valence takes "fair" actions that minimise the difference in payoff between themselves and others.

– Proposal Ranking Criteria:

$$minimise\, 1/(1 + |U - a(x)|) \tag{13}$$

– Response Acceptance Criteria:

$$accept\ trade\ if\ and\ only\ if\colon U - f(x, \bar{x}) < 0 \tag{14}$$

Negative Integral Emotion (Competitive Equity Aversion). An agent with negative integral emotion valence takes "unfair" actions that maximise the difference in payoff between themselves and others, and for which the payoff to themselves is greater than that to others.

– Proposal Ranking Criteria:

$$maximise:\ 1/(1 + |U - a(x)|) \tag{15}$$

– Response Acceptance Criteria:

$$accept\ trade\ if\ and\ only\ if\colon U - f(x, \bar{x}) > 0 \tag{16}$$

Internal Emotion State for *Svoie*. We adopted standard decision-making protocols for altruistic, cooperative, and selfish SVOs to form baseline *Stable-SVO* agents, where agents always make decisions which align with their SVO. We then introduced an integral emotion component to the *Stable-SVO* agents to produce a *Svoie* agent—an agent that has an SVO, as well as positive and negative integral emotion policies. We designed *Svoie* agents so that positive integral emotion would be associated with reaching the goal in a round of CT, and negative emotion with not reaching the goal. To encode integral emotion in *Svoie*, we define an internal state $E \in \{-1, -0.5, 0, 0.5, 1\}$ representing the current valence of the agent, e.g. the positiveness or negativeness of their integral emotion. This is an internal state that is updated based on goal achievement at the end of each game round. For simplicity, we allow E to take one of five

discrete states between –1 and 1, however, a higher granularity or continuous implementation could be used.

In the CT game, goal achievement results in a step increase in E and conversely, goal non-achievement results in a step decrease. We use E to define the probability that an agent selects an integral-emotion-based policy. $E = 0$ represents a neutral emotion state, in which the agent always defaults to its baseline SVO decision-making policy. When $E = 0.5$ or $E = -0.5$, the agent will have a 50% chance of selecting the positive or negative emotion policy respectively, and when $E = 1$ or $E = -1$, the agent will always select the associated emotion policy. In this way, agents can exhibit varying degrees of emotion-based behaviour over many repeat interactions depending on how frequently their decision-making policy causes them to achieve or miss their goals. The state E is designed to reflect the "appraisal theory" of emotion [36], that posits that human emotions are internal phenomena, constructed through the appraisal of external events and stimuli, for example, by evaluating whether an event outcome aligns with personal goals or expectations.

4 Experiments and Results

We conducted simulation experiments using CT (Sect. 3.1) as a task environment. We repeat our experiments using four different agent societies, which we define based on the proportions of agents with different SVO trait:

altr-coop Agent society with equal number of altruistic and cooperative agents
altr-self Agent society with equal number of altruistic and selfish agents
coop-self Agent society with equal number of cooperative and selfish agents
mixed Agent society with number of altruistic, cooperative and selfish agents

Each simulation is run over 1,000 time steps. At each time step, each agent in the society is paired with another agent at random, and each pair of agents plays two rounds of CT and receives a score. We compare simulations of *Svoie* agent societies to simulations of *Stable-SVO* societies.

Stable-SVO Agents follow fixed decision-making rules associated with their SVO.
Svoie Agents act the same as *Stable-SVO* initially, and have an SVO trait, but may deviate from their stable SVO trait based on game outcomes.

We define metrics and hypotheses in Sect. 4.1, for evaluating whether the integral emotion mechanism introduced in *Svoie* has a beneficial effect on societal outcomes at the end of the simulations.

4.1 Evaluation Metrics and Hypotheses

We define and compute *Individual Welfare*, *Collective Welfare* and *Welfare Inequality* for evaluating simulated *Svoie* and *Stable-SVO* agent societies.

Welfare measures the success of agents in maximising their score. We calculate the mean score achieved by individual agents and across samples of agents to evaluate welfare.

Inequality measures inequality of outcomes between members of an agent society. We assess inequalities over distributions using the Coefficient of Variation (CoV) measure [32]. Whereas Gini Coefficient is used in other research to measure inequality, we select CoV for its simplicity, and because the distributions of individual measures are observed to be approximately normal in preliminary runs.

1. ***Individual Welfare*** The mean score an individual agent achieves over all time steps in a simulation run.
2. ***Collective Welfare*** The mean score over a sample of agents.
3. ***Welfare Inequality*** The CoV of the distribution of individual welfare of agents in a sample. The magnitude of this measure is smaller for more equal societies.

We evaluate two hypotheses corresponding to the evaluation metrics for simulated agent societies.

H1 *Svoie* gives greater collective welfare than *Stable-SVO* over all agents in a society.

H2 *Svoie* gives lower welfare inequality than *Stable-SVO* over all agents in a society.

4.2 Simulation Setup

We simulated a sequence of CT games, described in Sect. 3.1, between random pairs of agents in each multi-agent society. At each time step, all agents are randomly paired, and each pair of agents plays two rounds of CT. Each simulation was performed over 1,000 time steps with a population size of 300 to account for random variations in game set-up and agent pairings at each time step. For each game round, we record the scores achieved by each agent. At the end of each simulation, we compute the metrics listed in Sect. 4.1. For *Svoie* agents, we initialised integral emotion to $E = 0$, so that all agents start by using the policy associated with their SVO trait.

The results presented are derived from the average of three repetitions for each simulation. We conducted tests to identify significant differences in our evaluation metrics between *Svoie* and *Stable-SVO*, across entire societies and specific samples of agents with a particular SVO trait. We use a two sample t-test, and report the means, μ, and p-values, p, and measure effect size as Cohen's d [10].

4.3 Evaluation

To evaluate hypotheses H1 on collective welfare, and H2 on welfare inequality, we compared *Svoie* to *Stable-SVO* for the four societies described in Sect. 4: *altr-coop*, *altr-self*, *coop-self* and *mixed*.

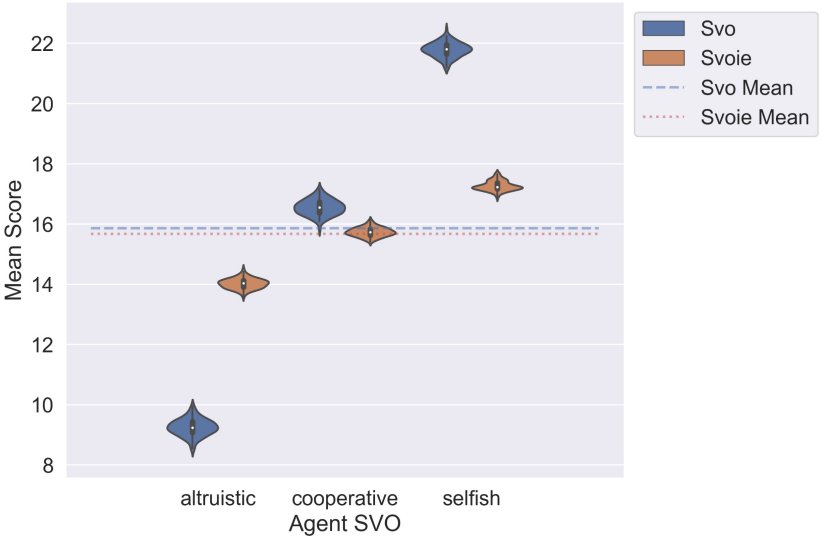
Table 1 compares population metrics measured for *Svoie* and *Stable-SVO* agent societies: (1) mean game score (Mean Score) achieved by all agents in a society, and in samples of agents with the same SVO, as a measure of the collective welfare achieved by those groups; (2) the coefficient of variation (CoV) of the distribution of mean welfare for individual agents in each group as a measure of welfare inequality. All results are calculated from three repeat runs.

Table 1. Comparison of the mean score and coefficient of variation in societies of *Stable-SVO* and *Svoie* agents with various combinations of SVO traits.

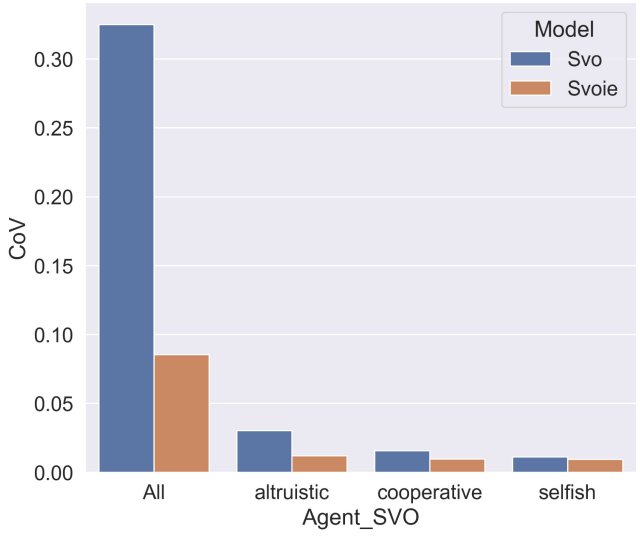
Configuration			*Stable-SVO*			*Svoie*		
Society	Sample SVO	Size	Mean score	Std	CoV	Mean score	Std	CoV
altr coop	all	300	16.299	2.435	0.149	15.754	0.807	0.051
	altr	150	13.877	0.206	0.015	14.966	0.169	0.011
	coop	150	18.720	0.221	0.012	16.541	0.177	0.011
altr self	all	300	15.257	7.358	0.481	15.456	1.676	0.108
	altr	150	7.917	0.279	0.035	13.791	0.153	0.011
	self	150	22.597	0.271	0.012	17.121	0.181	0.011
coop self	all	300	16.299	1.758	0.108	15.816	0.690	0.044
	coop	150	14.558	0.226	0.015	15.147	0.157	0.010
	self	150	18.040	0.219	0.012	16.486	0.169	0.010
mixed	all	300	15.863	5.149	0.324	15.664	1.332	0.085
	altr	100	9.269	0.271	0.029	14.016	0.170	0.012
	coop	100	16.527	0.243	0.015	15.731	0.160	0.010
	self	100	21.792	0.267	0.012	17.244	0.179	0.010

Our findings suggest that deviations from stable SVO traits in *Svoie* minimally impact collective welfare. We find that there is no significant difference in collective welfare in the mixed society, *Svoie* (μ = 15.664) and *Stable-SVO* (μ = 15.863), (p=0.5436, d=0.0497), or for the altr-self society, *Svoie* (μ = 15.456) and *Stable-SVO* (μ =15.257) (p=0.6739, d=0.034). However, *Svoie* yields lower collective welfare in both the altr-coop society, *Svoie* (μ = 15.754) and *Stable-SVO* (μ =16.299) (p<0.001, d=0.305), and in the the coop-self society, *Svoie* (μ=15.816) and *Stable-SVO* (μ = 16.299) (p<0.001, d=0.371), albeit with small effect size. Therefore, the societies of *Svoie* agents, which are more likely to seek fair or "inequity averse" actions in response to reaching goals and which are more likely to seek unfair "competitive equity averse" in response to missing goals, were found to perform roughly as well as societies of agents which only act according to their SVO.

Across all societies, we observed that *Svoie* agents significantly reduced welfare inequality compared to *Stable-SVO*, with a large effect size: (alt-coop: *Stable-SVO* μ = 16.299, *Svoie* μ = 15.754, p<0.001 d=84.106), (altr-self: *Stable-SVO*



(a) Mean welfare.



(b) Inequality.

Fig. 2. Comparison of welfare and inequality in societies of *Stable-SVO* and *Svoie* agents, with an equal number of agents with altruistic, cooperative and selfish SVO traits.

$\mu = 15.257$, *Svoie* $\mu = 15.456$, $p<0.001$, $d=39.543$), (self-coop *Stable-SVO* $\mu = 16.299$, *Svoie* $\mu = 15.816$, $p<0.001$, $d=73.007$), (*Stable-SVO* μ 15.863, *Svoie* $\mu = 15.664$, $p<0.001$, $d=32.384$). This is illustrated by the distributions of individual welfare (mean score) for samples of agents in the mixed society simulation, shown in Fig. 2a (a). We can see that the distributions of scores for each sample of agents with a particular SVO trait are further apart for the *Stable-SVO* simulations and closer together for the *Svoie* simulations, but the ordering of their performance is unchanged. For example, we observe that altruistic *Svoie* agents perform better than altruistic *Stable-SVO* agents, as they are likely to use unfair strategies in response to being taken advantage of, and selfish *Svoie* agents perform worse than selfish *Stable-SVO*, as they are likely to use fair strategies after taking advantage of others. Further, the width of the distributions of mean score for each SVO is reduced in the *Svoie* simulation, therefore welfare inequality within an individual SVO trait sample is reduced relative to *Stable-SVO* societies as well. This is reflected in the data shown in Table 1 which contains measurements of the mean score achieved by samples of agents with different SVO traits, and the coefficient of variation of the distributions of agent scores within those samples.

5 Limitations, Directions and Conclusions

We now discuss limitations and directions. Firstly, we model societies with heterogeneous SVO by generating populations of agents which can take one of either two or three distinct SVO traits, from altruistic, selfish and cooperative. In human societies, SVO varies continuously between individuals as described by the ring model [31]. Further, we assume an equal distribution of SVO traits in society, whereas in human societies, certain ranges of SVO are more common than others. Buckman et al. [7] implement a more realistic treatment of SVO in agent societies, by sampling agent traits from ranges of the SVO ring model found to be most prevalent in human society using relevant experimental data on SVO prevalence. We did not attempt to simulate realistic human societies, and were focused instead on modelling integral emotions alongside SVO to investigate how this would affect societal welfare and welfare inequality in a society of agents with different SVO policies. The three SVO policies we used in our work give different and non-overlapping distributions in welfare in our baseline simulations, and we therefore considered them to be appropriate for our purposes.

Secondly, we model integral emotion as the variable state E using several simplifying assumptions which prevent any direct comparison with integral emotion in real human behaviour. We only incorporate two integral-emotion-based policies, for positive and negative E respectively. These policies are based on human behaviours which have previously been associated with positive and negative emotions, however they do not follow any explicit model. Further, we assume only one environment trigger, goal-achievement or non-achievement, to be relevant for influencing emotion, whereas there is evidence that other factors influence emotion, e.g. fairness of outcomes [40,44], which could be utilised in the

CT environment. We also only allow E to vary over five possible states, and the extent to which E changes is constant and chosen arbitrarily, preventing any differences in sensitivity to emotional stimuli between agents. We implemented *Svoie* agents as a coarse-grained model of SVO and integral emotion in agent societies, and did not seek to accurately model human behaviour. In this context, we found that societies of *Svoie* agents had lower welfare inequality compared to baseline *Stable-SVO* agents, and that collective welfare was preserved. These results suggest that by introducing transient changes in decision-making, triggered by task relevant events, agents can adapt their otherwise stable policies depending on the society they operate within.

Agent-based modelling of social decision-making will always require simplifying approximations and assumptions, and cannot accurately capture all aspects of human behaviour, but they are nevertheless useful for studying specific aspects and edge cases [14]. Research on integral emotions (discussed in Sect. 2.2) present the foundational idea behind our contributions—in a social decision-making context, people may make seemingly irrational choices in reaction to recent task outcomes, which may primarily be motivated by strong task-related integral emotions rather than by fixed values, or a rational effort to punish or reward another person due to perceived social inequity. We conduct simulation experiments to investigate the effects at the society level that result from acting according to a simplified and idealised model of this type of behaviour, when compared to acting rationally according to fixed preferences. The limiting and simplifying assumptions of our agent model mean that we cannot predict whether the effects that we observe would extend to real human societies. However, this simulation method offers a useful tool for modelling dynamic behaviour, and better understanding existing models of human behaviour. Understanding the interplay between emotions and social preferences in human decision-making is important for the development of autonomous agents which can understand human social norms, and act in accordance with human moral and ethical principles [1,28,38,49].

There are many factors thought to exert a guiding influence on human behaviour, and models which seek to explain how these factors give rise to a variety of seemingly irrational patterns of behaviour observed in humans, such as predictable deviations from fixed preferences in games and other social contexts. Simulation methods have been applied in related works to study the possible adaptive and socially beneficial effects of different examples of these phenomena. For example, Kampik et al. [25] investigated the role of sympathy in cooperative behaviour, Sylwester et al. [45] have examined antisocial punishment, paying a cost to punish pro-social actors, as a form of social norm enforcement, and Köster et al. [29] demonstrate how the enforcement of arbitrary and inconsequential social norms may improve overall norm compliance in agent societies. Further, there is a rich body of existing work which explores the role of human factors on norm emergence in multi-agent systems [37,41].

In our chosen simulation task environment, CT, random variations allow differences between the scenarios encountered by agents in each game, however the average performance for any agent is predictable over many time steps. This

work could be extended by investigating how integral emotions influence societal outcomes across multiple task environments, to understand the implications of integral emotion for regulating behaviour in a changing environment. Here, the societal effects of emotions and individual differences could be studied in the context of simulating the emergence and spread of norms in multi-agent systems which benefit survival.

Acknowledgements. DC was supported by the UK Research and Innovation (UKRI) Centre for Doctoral Training in Interactive Artificial Intelligence Award (EP/S022937/1). CH is a Leverhulme Research Fellow (RF-2021-533). NA thanks the University of Bristol for support. DC thanks Phillip Sloan for help with Colored Trails implementation.

References

1. Ajmeri, N., Guo, H., Murukannaiah, P.K., Singh, M.P.: Elessar: ethics in norm-aware agents. In: Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), pp. 16–24. IFAAMAS, Auckland (2020). https://doi.org/10.5555/3398761.3398769
2. Andrighetto, G., Capraro, V., Guido, A., Szekely, A.: Cooperation, response time, and social value orientation: a meta-analysis. PsyArXiv (2020). https://doi.org/10.31234/osf.io/cbakz
3. Balliet, D., Parks, C., Joireman, J.: Social value orientation and cooperation in social dilemmas: a meta-analysis. Group Process. & Intergroup Relat. **12**(4), 533–547 (2009). https://doi.org/10.1177/1368430209105040
4. Bogaert, S., Boone, C., Declerck, C.: Social value orientation and cooperation in social dilemmas: a review and conceptual model. Br. J. Soc. Psychol. **47**(Pt 3), 453–480 (2008). https://doi.org/10.1348/014466607X244970
5. Bono, S.A., van der Schalk, J., Manstead, A.S.R.: The roles of social value orientation and anticipated emotions in intergroup resource allocation decisions. Front. Psychol. **11**, 1455 (2020). https://doi.org/10.3389/fpsyg.2020.01455
6. Braz, L.F., Sichman, J.S.: Using the Myers-Briggs Type Indicator (MBTI) for modeling multiagent systems. Revista de Informática Teórica e Aplicada **29**(1), 42–53 (2022). https://doi.org/10.22456/2175-2745.110015
7. Buckman, N., Pierson, A., Schwarting, W., Karaman, S., Rus, D.: Sharing is caring: socially-compliant autonomous intersection negotiation. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 6136–6143. IEEE, Macao (2019). https://doi.org/10.1109/IROS40897.2019.8967997
8. Cheng, K.L., Zuckerman, I., Nau, D., Golbeck, J.: The life game: cognitive strategies for repeated stochastic games. In: Proceedings of the 3rd IEEE Int'l Conference on Privacy, Security, Risk and Trust and 3rd IEEE Int'l Conference on Social Computing, pp. 95–102. IEEE, Boston (2011). https://doi.org/10.1109/PASSAT/SocialCom.2011.62
9. Cheriyan, J., Savarimuthu, B.T.R., Cranefield, S.: Norm Violation in Online Communities—A Study of Stack Overflow Comments. In: Aler Tubella, A., Cranefield, S., Frantz, C., Meneguzzi, F., Vasconcelos, W. (eds.) Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XIII,

vol. 12298, pp. 20–34. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-72376-7_2
10. Cohen, J.: Statistical Power Analysis for the Behavioral Sciences, 2nd edn. Lawrence Erlbaum Associates, Hillsdale, New Jersey (1988)
11. Criado, N., Argente, E., Noriega, P., Botti, V.: Human-inspired model for norm compliance decision making. Inf. Sci. **245**, 218–239 (2013). https://doi.org/10.1016/j.ins.2013.05.017
12. Crosato, L., Wei, C., Ho, E.S.L., Shum, H.P.H.: Human-centric autonomous driving in an AV-pedestrian interactive environment using SVO. In: Proceedings of the IEEE 2nd International Conference on Human-Machine Systems (ICHMS), pp. 1–6. IEEE, Magdeburg (2021). https://doi.org/10.1109/ICHMS53169.2021.9582640
13. Dias, J., Mascarenhas, S., Paiva, A.: FAtiMA Modular: towards an agent architecture with a generic appraisal framework. In: Bosse, T., Broekens, J., Dias, J., van der Zwaan, J. (eds.) Emotion Modeling. Lecture Notes in Computer Science, vol. 8750, pp. 44–56. Springer International Publishing, Cham (2014). https://doi.org/10.1007/978-3-319-12973-0_3
14. Dignum, F.: Should we make predictions based on social simulations? Int. J. Soc. Res. Methodol. **26**(2), 193–206 (2023). https://doi.org/10.1080/13645579.2022.2137925
15. Gal, Y., Grosz, B., Kraus, S., Pfeffer, A., Shieber, S.: Agent decision-making in open mixed networks. Artif. Intell. **174**(18), 1460–1480 (2010). https://doi.org/10.1016/j.artint.2010.09.002
16. Gal, Y., Grosz, B., Kraus, S., Pfeffer, A., M. Shieber, S.: Colored trails: A formalism for investigating decision-making in strategic environments. In: Proceedings of the 2005 IJCAI workshop on reasoning, representation, and learning in computer games. 19th International Joint Conference on Artificial Intelligence, IJCAI-05; Conference date: 30–07-2005 Through 05–08-2005, pp. 25–30 (2005)
17. Gal, Y., Pfeffer, A.: Modeling reciprocal behavior in human bilateral negotiation. In: Twenty-Second Conference on Artificial Intelligence (AAAI-07). 22nd National conference on Artificial intelligence, AAAI-07; Conference date: 22–07-2007 Through 26–07-2007, vol. 22, pp. 815–821. AAAI Press (2007)
18. Grosz, B.J., Kraus, S.: The influence of social dependencies on decision-making: Initial investigations with a new game. In: Proceedings of the 3rd International Joint Conference on Multi-agent Systems, AAMAS'04, pp. 782–789 (2004)
19. Harsanyi, J.C.: On the rationality postulates underlying the theory of cooperative games. J. Confl. Resolut. **5**(2), 179–196 (1961). https://doi.org/10.1177/002200276100500205
20. Huang, X., Wu, W., Qiao, H.: Computational modeling of emotion-motivated decisions for continuous control of mobile robots. IEEE Trans. Cogn. Dev. Syst. **13**(1), 31–44 (2021). https://doi.org/10.1109/TCDS.2019.2963545
21. Isen, A.M.: An influence of positive affect on decision making in complex situations: theoretical issues with practical implications. J. Consum. Psychol. **11**(2), 75–85 (2001). https://doi.org/10.1207/S15327663JCP1102_01
22. Joireman, J.A., Shelley, G.P., Teta, P.D., Wilding, J., Michael Kuhlman, D.: Computer Simulation of Social Value Orientation: Vitality, Satisfaction, and Emergent Game Structures. In: Liebrand, W.B.G., Messick, D.M. (eds.) Frontiers in Social Dilemmas Research, pp. 289–310. Springer, Berlin, Heidelberg (1996). https://doi.org/10.1007/978-3-642-85261-9_16
23. de Jong, S., Hennes, D., Tuyls, K., Gal, Y.: Metastrategies in the colored trails game. In: AAMAS (2011)

24. Kalia, A.K., Ajmeri, N., Chan, K.S., Cho, J.H., Adalı, S., Singh, M.P.: The interplay of emotions and norms in multiagent systems. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence, pp. 371–377. International Joint Conferences on Artificial Intelligence Organization, Macao (2019). https://doi.org/10.24963/ijcai.2019/53
25. Kampik, T., Nieves, J.C., Lindgren, H.: Explaining Sympathetic Actions of Rational Agents. In: Calvaresi, D., Najjar, A., Schumacher, M., Främling, K. (eds.) Explainable, Transparent Autonomous Agents and Multi-Agent Systems. Lecture Notes in Computer Science, vol. 11763, pp. 59–76. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-30391-4_4
26. Kianpour, M., Øverby, H., Kowalski, S.J., Frantz, C.: Social Preferences in Decision Making Under Cybersecurity Risks and Uncertainties. In: Moallem, A. (ed.) HCI for Cybersecurity, Privacy and Trust. Lecture Notes in Computer Science, vol. 11594, pp. 149–163. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-22351-9_10
27. Kraus, S.: Negotiation and cooperation in multi-agent environments. Artif. Intell. **94**(1), 79–97 (1997). https://doi.org/10.1016/S0004-3702(97)00025-8
28. Kuipers, B.: Human-like morality and ethics for robots. In: AAAI Workshop: AI, Ethics, and Society (2016)
29. Köster, R., Hadfield-Menell, D., Everett, R., Weidinger, L., Hadfield, G.K., Leibo, J.Z.: Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents. Proc. Natl. Acad. Sci. **119**(3), e2106028118 (2022). https://doi.org/10.1073/pnas.2106028118
30. Lerner, J.S., Li, Y., Valdesolo, P., Kassam, K.S.: Emotion and decision making. Annu. Rev. Psychol. **66**(1), 799–823 (2015). https://doi.org/10.1146/annurev-psych-010213-115043
31. Liebrand, W.B.G., McClintock, C.G.: The ring measure of social values: a computerized procedure for assessing individual differences in information processing and social value orientation. Eur. J. Pers. **2**(3), 217–230 (1988). https://doi.org/10.1002/per.2410020304
32. Maio, F.: Income inequality measures. J. Epidemiol. Community Health **61**, 849–52 (2007). https://doi.org/10.1136/jech.2006.052969
33. McClintock, C.G.: Social motivation-a set of propositions. Behav. Sci. **17**(5), 438–454 (1972). https://doi.org/10.1002/bs.3830170505
34. McClintock, C.G., Allison, S.T.: Social value orientation and helping behavior1. J. Appl. Soc. Psychol. **19**(4), 353–362 (1989). https://doi.org/10.1111/j.1559-1816.1989.tb00060.x
35. de Melo, C.M., Terada, K.: The interplay of emotion expressions and strategy in promoting cooperation in the iterated prisoner's dilemma. Sci. Rep. **10**(1), 14959 (2020). https://doi.org/10.1038/s41598-020-71919-6
36. Moors, A.: Appraisal Theory of Emotion. In: Zeigler-Hill, V., Shackelford, T.K. (eds.) Encyclopedia of Personality and Individual Differences, pp. 1–9. Springer International Publishing, Cham (2017). https://doi.org/10.1007/978-3-319-28099-8-493-1
37. Morris-Martin, A., De Vos, M., Padget, J.: Norm emergence in multiagent systems: a viewpoint paper. Auton. Agents Multi-Agent Syst. (JAAMAS) **33**(6), 706–749 (2019)
38. Murukannaiah, P.K., Ajmeri, N., Jonker, C.M., Singh, M.P.: New foundations of ethical multiagent systems. In: Proceedings of the 19th International Conference

on Autonomous Agents and Multiagent Systems (AAMAS), pp. 1706–1710. IFAAMAS, Auckland (2020). https://doi.org/10.5555/3398761.3398958. Blue Sky Ideas Track
39. Nardin, L.G., Balke-Visser, T., Ajmeri, N., Kalia, A.K., Sichman, J.S., Singh, M.P.: Classifying sanctions and designing a conceptual sanctioning process model for socio-technical systems. Knowl. Eng. Rev. (KER) **31**, 142–166 (2016)
40. Pillutla, M.M., Murnighan, J.K.: Unfairness, anger, and spite: emotional rejections of ultimatum offers. Organ. Behav. Hum. Decis. Process. **68**(3), 208–224 (1996)
41. Savarimuthu, B.T.R., Cranefield, S.: Norm creation, spreading and emergence: a survey of simulation models of norms in multi-agent systems. Multiagent Grid Syst. **7**(1), 21–54 (2011). https://doi.org/10.3233/MGS-2011-0167
42. Schwarting, W., Pierson, A., Alonso-Mora, J., Karaman, S., Rus, D.: Social behavior for autonomous vehicles. Proc. Natl. Acad. Sci. **116**(50), 24972–24978 (2019). https://doi.org/10.1073/pnas.1820676116
43. Sloan, P., Ajmeri, N.: Commitment-based negotiation semantics for accountability in multi-agent systems. In: Proceedings of the 10th International Workshop on Engineering Multi-Agent Systems (EMAS), pp. 1–25. Springer, Virtual (2022). https://doi.org/10.1007/s10472-023-09875-w
44. Straub, P.G., Murnighan, J.K.: An experimental investigation of ultimatum games: information, fairness, expectations, and lowest acceptable offers. J. Econ. Behav. & Organ. **27**(3), 345–364 (1995)
45. Sylwester, K., Herrmann, B., Bryson, J.J.: Homo homini lupus? Explaining antisocial punishment. J. Neurosci. Psychol. Econ. **6**(3), 167–188 (2013). https://doi.org/10.1037/npe0000009
46. Tzeng, S.T., Ajmeri, N., Singh, M.P.: Fleur: Social Values Orientation for Robust Norm Emergence. In: Ajmeri, N., Martin, A.M., Savarimuthu, B.T.R. (eds.) Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XV. Lecture Notes in Computer Science, vol. 13549, pp. 185–200. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-20845-4_12
47. Vilone, D., Realpe-Gómez, J., Andrighetto, G.: Evolutionary advantages of turning points in human cooperative behaviour. PLoS ONE **16**(2), e0246278 (2021). https://doi.org/10.1371/journal.pone.0246278
48. Västfjäll, D., Slovic, P., Burns, W.J., Erlandsson, A., Koppel, L., Asutay, E., Tinghög, G.: The arithmetic of emotion: integration of incidental and integral affect in judgments and decisions. Front. Psychol. **7** (2016). https://doi.org/10.3389/fpsyg.2016.00325
49. Woodgate, J., Ajmeri, N.: Macro ethics for governing equitable sociotechnical systems. In: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS), pp. 1824–1828. IFAAMAS, Online (2022). https://doi.org/10.5555/3535850.3536118. Blue Sky Ideas Track
50. Yamagishi, T., Horita, Y., Takagishi, H., Shinada, M., Tanida, S., Cook, K.S.: The private rejection of unfair offers and emotional commitment. Proc. Natl. Acad. Sci. **106**(28), 11520–11523 (2009). https://doi.org/10.1073/pnas.0900636106
51. Yamagishi, T., Li, Y., Takagishi, H., Matsumoto, Y., Kiyonari, T..: In Search of Homo economicus. Psychol. Sci. **25**(9), 1699–1711 (2014). https://doi.org/10.1177/0956797614538065
52. Zheng, Y., Yang, Z., Jin, C., Qi, Y., Liu, X.: The influence of emotion on fairness-related decision making: a critical review of theories and evidence. Front. Psychol. **8**, 1592 (2017). https://doi.org/10.3389/fpsyg.2017.01592