philosophies

MDPI

Review

Prismal View of Ethics

Sarah Isufi [1], Kristijan Poje [2], Igor Vukobratovic [3] and Mario Brcic [2,*]

[1] For Humanity, Thornwood, NY 10594, USA
[2] Faculty of Electrical Engineering and Computing, University of Zagreb, 10000 Zagreb, Croatia
[3] Independent Researcher, 10000 Zagreb, Croatia
* Correspondence: mario.brcic@fer.hr

Abstract: We shall have a hard look at ethics and try to extract insights in the form of abstract properties that might become tools. We want to connect ethics to games, talk about the performance of ethics, introduce curiosity into the interplay between competing and coordinating in well-performing ethics, and offer a view of possible developments that could unify increasing aggregates of entities. All this is under a long shadow cast by computational complexity that is quite negative about games. This analysis is the first step toward finding modeling aspects that might be used in AI ethics for integrating modern AI systems into human society.

Keywords: ethics; game theory; multiagent systems; AI ethics; moral innovation

check for updates

Citation: Isufi, S.; Poje, K.; Vukobratovic, I.; Brcic, M. Prismal View of Ethics. *Philosophies* **2022**, *7*, 134. https://doi.org/10.3390/philosophies7060134

Academic Editor: Marcin J. Schroeder

Received: 5 October 2022
Accepted: 23 November 2022
Published: 25 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Life is rich with challenges, decision-making, and questions we pose to ourselves. Decision-making occurs within a context whose characteristics we will refer to as The Setting. *Ethics* is a discipline concerned with good and wrong moral values and norms that can be right and wrong. Norms define standards of acceptable behavior by groups. Specific ethical systems, through their norms (computable conventions), constrain and partially solve the problem of life. The importance of ethics for society is paramount, as no social group can stay cohesive and in existence if there are no constraints on the behavior of individuals. For example, frequent, reasonless escalations and attacks with killing or injuring others would dissolve any group. Authors [1] refer to morality as pro- or anti-social norms with direct benefit or cost to others (e.g., theft, murder, generosity, sharing).

Significant technological and cultural advancements have occurred throughout the last millennia of human history. The speed with which these changes arrived was accelerating. However, it was still a pedestrian pace compared to the changes coming with more excellent connectivity (internet), stronger computation (Moore's and descendant laws), cognitively powerful non-human entities (artificial intelligence), and many other disruptive technologies made possible by those. Strong computation and algorithms introduce powerful, flexible, and fast-changing entities into society while the connectivity diffuses the effects of their actions to all corners of the world. All social groups will become paired with these artificial entities, and social adaptation and integration will, due to the speed of changes, be tested as never before. Technology that is the source of difficulties in the first place can, through its dual use, also be used to help alleviate the problem. Wittgenstein suggested a pragmatic view on language development through language games [2]. We wish to pursue a similar line of thought with ethics and investigate its properties from the computational perspective. We shall tease out different properties that might help in modeling, simulating, and potentially innovating ethical systems that will circumvent issues and deliver us to the good side of future history.

Cooperation was a topic of thorough research conducted and surveyed from the perspective of social [3,4] and natural sciences [5]. The former has approached the problem from the top through empirical studies on people. They face interpretation problems

because scarce results under-constrain the studied complex setting and leave a multitude of plausible interpretations. Social physicists have approached the problem bottom-up by researching the evolution of cooperation in a simplified utilitarian setting of social dilemma games with low strategic complexity. This narrow focus has enabled them to establish a richness of rigorous conclusions. However, their applicability to realistic cases is quite limited for several reasons. Simplifying assumptions that need to be made for computational reasons also limits the transfer of results to other situations. Cooperation in social dilemma games is only one form of a more general class of moral behavior [6]. Preferences of agents in some situations cannot be entirely explained just by the monetary outcomes of games, but following personal norms can offer a better explanation [7]. Moreover, Bowles [8] claims that incentives and social preferences are not separable, and the former affects the latter. Additionally, Broome [9] criticizes approaches that assume a single objective that affects each agent's decision-making. Although cooperation is much better understood, there are no conclusive answers to essential questions about cooperation and ethics.

Design for values (value-sensitive and ethically aligned) is an application of ethics that calls for responsible innovation in the face of accelerating progress that strains the existing social fabric [10,11]. We hold that with the increasing complexity of technology, we are hitting the limits of inference, such as unverifiability and limits to explainability [12], that make that well-intentioned proposition long-term infeasible in the current form due to cognitively superior agents with which value alignment is still a wide-open problem.

The contributions of this paper are:

- We offer a review of work in different fields related to investigating cooperation and ethics.
- We pivot from the existing practice by focusing on ethics as the first-class mechanism, teasing out its general properties to provide common ground for future interdisciplinary investigations. Additionally, we enrich the description with a computational perspective that relates to computational efficiency. The research in social physics has been narrowly focused and sometimes off-mark by focusing on problems that do not possess these properties. It aimed to show conditions and mechanisms under which cooperation emerges from unbiased and simplified non-cooperative agents. Such generality is a vital a priori requirement, and the found conditions may not even be aligned with our current situation. On the other hand, we accept and carefully describe the current position where humans have significant prosocial bias.
- We argue for more intentional moral innovation to prepare for coexistence with cognitively superior agents. Current ethics so far emerged collaterally has some deficient properties that make value alignment with advanced technologies even more challenging. We can even use technology for meet-in-the-middle approaches to value alignment. Based on computational complexity considerations, we provide a few pointers regarding how this can be made.

Section 2 deals with the basic properties of ethics. Section 3 considers modeling situations/decisions posing ethical dilemmas through game theory and multi-agent systems. Section 4 deals with ethics and its importance in group coordination. Furthermore, it elaborates algorithmic role and utility of the human values in group coordination. Evolutionary game theory as a modeling basis for ethics is described in Section 5. In Section 6, we dive into the advanced properties of ethics that deal with global inconsistencies and fine balance between competition and cooperation in groups. Conclusions are drawn in Section 7 and future directions are proposed in Section 8.

2. Basic Properties of Ethics

In the following section, we shall cover the views on the basic properties of ethics in the literature. We deal with the purpose of ethics as a group performance improver. We describe the setting in which it operates as uncertain and multicriterial. These, coupled with the societal scale make for a complex setting in which adaptability is crucial. Ethics changes, but so far emerged collaterally through cultural evolution on a longer time scale.

This seems to have worked well, but the shortening of timescales and greater societal perturbations due to rapid technological advances are bringing the effectiveness of such an unguided process into jeopardy. This drives our argument of the necessity for more intentional moral innovation and managing of the system later in the paper.

2.1. Main Purpose of Ethics

We argue that the primary purpose of ethics is achieving better group performance. Coordination is an essential aspect of ethics since it makes collectives more progressive under the cap of available resources to bring about a better outcome for the group. Cooperative societies with a clear division of labor progressed faster because the members were united under the same goal [13,14].

The cost-effectiveness of cooperation at all stages of social development is an essential item and a prerequisite for determining posterior ethics and moral rules. Despite the complexity of moral imperatives in the past and present ethics, many have a visible discourse about cooperation within the community. From the slave-owning societies of Greece to South America, from monarchies to republics, there is a rule of respect and cooperation with one's equals. The difference is in the definition of equality and which social, age and class groups fall into that definition, and which are outside it. Thus, cooperation is a plausible precondition for the emergence of ethics per se, no matter how it developed later, whether it included a larger or smaller group of people, the whole community, or just a selected few. The advancements of societies are a by-product of individual satisfaction, which comes from social evolution. To maximize social evolution, the freedom of individuals is required because only then the selection process has enough variability to maximize social fitness. Consequently, the freedom and struggle for survival yield altruism which, next to cooperation, serves as the backbone of every prosperous society [15].

The selection process is the main reason for increased altruism in society, illustrated in the following example. Parents who are not altruistic toward their children will have children with a lower survival rate. Over time, the altruistic population will increase, and individuals without those traits will decrease in the number [16]. This kin-based altruism [15] has a limited range of effects. Another mechanism for cooperation is reciprocity which appears in repeated interactions [17]. It somewhat increases the range of effects but is sustainable in dyads and tends to collapse in larger groups [4,18]. Kin-based altruism and long-term interactions are mechanisms through which natural selection on genes can produce cooperation [19]. However, they are insufficient to explain humans' high level of cooperation. Cultural products such as social norms and institutions maintained by mechanisms related to reputation, signaling, and punishment form longer-term cooperations within much larger groups and under a broader range of conditions [4,20].

2.2. Ethical Dynamics

It has already been mentioned how normative ethics emerges due to social dynamics and cultural trade-offs. Simply put, ethics limits the abilities of person A, so that person A cannot harm person B and vice versa. The trade-off is the willing acceptance of a restriction of action to increase welfare for all the factors involved and the general social structure. Therefore, this makes ethics prone to change with changing social standards and ultimately uncertain.

The basis of the claim is the existence of a social consensus on whether a rule will be accepted or not [21]. It is evident that ethics is not static, but it changes dynamically in response to environmental changes. One needs not look too far into the past to see remarkable changes in moral systems during the 20th century [22]. Today's growing technological innovation puts people in new situations that need new societal wisdom, for example, artificial intelligence, globalization, the rise of multinationals, and the metaverse.

The nature of human morality is confined to norms and conventions [23] describing the individual's behavior and posed rules to regulate individuals and groups. However,

society's degree to which individual differences are permitted is variable. The tolerance of society to the variability of individual diversity is crucial to maintaining a biological system that adapts to changes in the environment and throughout time. In general, the rules defining the morality of humans have evolved to transmit genes to succeeding generations. Various beliefs and behaviors have been developed to support this goal during various historical accidents, climate changes, and different structures of the gene pools [24].

Another essential aspect of the evolution of morals is that cultural bias and human values are not genetically predetermined, i.e., humans have multiple behavioral potentials. Despite inherited predispositions, humans have the emotional and cognitive abilities to be selfish and cooperative. Different circumstances and societies cause individuals to find their moral trajectory [25]. Cultural evolution is another driver of the development of human culture. Humans share information via language and media (e.g., music, writings) that enables the distribution of information and resources, thus providing mechanisms for cultural evolution. Another property of humans accelerating cultural evolution by freeing cultural information from conceptual limits is metarepresentation, i.e., thinking about how we think [26]. Unlike genetic evolution causing slow changes in societal culture, cultural evolution has a substantially faster rate [27].

2.3. Multicriteriality of The Setting

Utility theory, which is based on improving a single objective, has been criticized in economics due to the apparent incommensurability of options in reality [9]. Brcic and Yampolskiy [12] hypothesize that human decision-making is made in multicriterial space where the mood selects a subset of focal criteria. These focal criteria are heuristically optimized as near as possible to the Pareto front. Non-focal criteria are simultaneously kept within acceptable bounds. When there are multiagent interactions, we enter the multicriterial aspects of ethics, which tend to create ethical dilemmas. The trade-offs can be between the essential drives within an individual or between the benefits of the individual and society. These dilemmas cannot be elegantly resolved. We argue that this can be connected to the property that there are many competing criteria on which ethical decisions must be based, as well as decisions in which games to take part at a specific moment. The trolley problem [28] is one such problem where the criteria of "do no harm" and "reduce suffering" play against each other and cannot flatly be resolved without being wrong against some criteria. Ethical dilemmas constrain the achievement of perfect outcomes, so it is often impossible to respect multiple criteria simultaneously. This means inevitable trade-offs, defined in, e.g., fairness [29] and Social Choice Theory (SCT) [30], must be made where we choose the solution that achieves the maximal possible hypervolume indicator [12].

Consequentialist ethics is prone to dilemmas originating from multicriteriality whereby several criteria must be traded off in a consequential state. Deontological ethics can use norms to dissolve complicated, commonly occurring dilemmas into more specific coordination problems [31]. However, such systems introduce dilemmas through inconsistencies, as explained in Section 6.2.

2.4. Status of Uncertainty in Ethics

Ethical behavior is, first and foremost practical activity. Namely, epistemological limits (information and cognition) are not held against actors in the case of mistakes and bad outcomes; instead, they are used for discounting responsibility. Actors often do not possess sufficient information or necessary cognition to achieve omniscient and omnipotent (and yet, still subject to some limits) solutions. Courts recognize the same principle in most legal systems. For example, a person with temporary or permanent reduced cognitive ability will receive a more lenient sentence. The primary motivation behind this act is that mental impairment caused by mental illness or substance use diminishes the mental capacity to make rational decisions [32].

Another example of ethical uncertainty is caused by insufficient information. Hindsight bias indicates that human post-fact decisions are likely to be affected by knowing the

outcome of their actions. This means humans will reconstruct the entire thinking process leading them to an initial decision when they hear the outcome and change their final decision accordingly [33]. Hindsight is discounted from responsibility. For this reason, if a surgeon, for example, misinterprets a patient's diagnosis due to latent factors leading a patient to death, he will not be prosecuted. Had he known the actual diagnosis, he would have taken different actions.

2.5. Collateral Nature

There are universal moral rules, but there is no unified ethics [34] as there is a lot of variation between different moral systems in human culture [35]. Ethics has always awaited us; it was a forward handoff from a continuous stream of generations to their posterity. As the COVID-19 pandemic has demonstrated, new situations call for new solutions. Since new situations, especially significant ones, are inherently random, ethics has so far emerged collaterally, i.e., under no guidance by some human designer. Societies adjust ethics to technical progress, social conditions, and cultural standards. The question is: Can ethics that does not arise collaterally even be created? Can there be a system for predicting ethics or the best possible moral course for society?

When considering the origin of ethics, we argue that it has emerged collaterally but not randomly. Instead, several factors have influenced the development of ethics, including the neurobiological characteristics of each individual and the sociocultural environment in which the individual develops. Moreover, the essential elements determining the development of moral judgment and consequently functioning when resolving dilemmas are derived from cultural characteristics, spirituality, socioeconomic environment, life experiences, and correct neurological functioning [36].

Darwin's view on moral theory is based on conscience, i.e., social instinct. A social instinct is how an individual behaves in a group for that group's benefit. Individual behavior will result from adopted human values, influencing every decision that has consequences for the group. Consequently, the social instinct results from the group's evolution, increasing group fitness. Unlike other social animals, humans have developed intellect that allows reasoning when faced with dilemmas. However, such reasoning is inevitably constrained by social instinct and human values [37].

3. Ethics, Multiagency and Games

Game theory is a branch of science that deals with interactions between different actors, precisely the level of operation for ethics. Classical Game Theory (CGT) is based on rationality and just-in-time computation interleaved with acting with an unrealistic amount of information and computing. CGT enables simple interepisodic learning (memory) on the level of an individual. Evolutionary Game Theory (EGT) in the classical form is an application of game theory on evolving populations, and it does not require rationality. It is a form of evolutionary policy search where the genotype completely describes the lifetime behavior (phenotype). Hence, "learning" in EGT is populational and intergenerational. Multi-Agent Reinforcement Learning (MARL) [38,39] is a more modern framework than the previous two. It enables more complex and structured strategic learning on the level of individuals during their lifetime. It scales to more complex group dynamics and strategies than CGT, bringing about individual and lifetime learning compared to EGT.

It is plausible that ethics has arisen due to evolutionary processes that a game theory can model. Therefore, it can be represented by an evolutionary model containing a representation of the population's state and a dynamic set of laws influencing the state changes over time. Different mechanisms have been used to explain the rise of cooperation, norms, and ethics in societies: kinship altruism, direct reciprocity, indirect reciprocity, network reciprocity, group selection, and many others [6,40,41]. They have been analyzed from different perspectives, including biologists, political scientists, anthropologists, sociologists, social physicists, economists, etc. The following three concepts are crucial for our exposition. Nash equilibrium is a strategy profile from which deviation would not be profitable for any

player. Evolutionary Stable Strategy (ESS) is a refinement of the evolutionary stable Nash equilibrium. The population adopting it could not be invaded by mutant strategy through natural selection. Finally, correlated equilibrium is a generalization of Nash equilibrium that emerges in the presence of a correlation device.

3.1. Examples of Games

Many games are used in literature for theoretical analysis [5,7] and behavioral experiments [4]. Here, we give several examples with results obtained from them.

The Prisoner's Dilemma (PD) is one of the fundamental problems of game theory that shows remarkable property that can be connected to emergent ethics based on direct reciprocity [17]. This problem exemplifies pure competition, which is the most challenging environment for cooperation. Namely, in the case of a single-iteration PD game, the maximum benefit comes from selfish play, that is, from betraying a cooperating partner. However, when the problem is changed to a multi-iteration prisoner's dilemma, we can get cooperation between partners as stable and optimal behavior. By the folk theorem, iterated PD has an abundance of Nash equilibria, which solving process ends up sensitively depending on the specifics of the environment [5,42]—with both defection/extortion [43] and generosity being a possible dominant solution [44].

Cooperative behavior can also be observed in different contexts, such as where neighbors settle disputes in ways that are not achievable between strangers [45]. In the repeated play, selfishness is charged because the teammate has insight into the player's past moves, making it not profitable to be selfish through direct reciprocity. However, nowadays, we have tools such as Internet reputations and social media ratings that are publicly available, giving us insight into players' past moves without previously playing games.

The Stag Hunt (SH) problem in game theory originated from Rousseau's Discourse on Inequality as a prototype of the social contract. It describes the trade-off between safety and cooperation to achieve more significant individual gain [46]. Unlike the PD problem, where an individual's rationality and mutual benefit are conflicted, in the SH problem, the rational decision is nearly a product of beliefs about what the other player will do. If both players decide to employ the same strategy, stag hunting and hare hunting are the best options. However, if one player chooses to hunt stag, he risks the other player will not cooperate. On the other hand, a player choosing to hunt a hare is not faced with such a risk since the other player's actions do not influence his outcome, meaning rational players face a dilemma of mutual benefit and personal risk [47].

Fair division theory deals with procedures for dividing a bundle of goods among n players where each has equal rights to the goods. Comparing which procedure is the most equitable gives a fair insight into popular notions of equity [48]. The modern theories of fair division are used for various purposes, such as division of inheritance, divorce settlement, and frequency allocation in electronics. The most common division procedure is divide and choose, used for a fair division of continuous resources. Steinhaus describes it in an example of dividing a cake among two people where the first person cuts the cake into two pieces and the second person selects one of the pieces; the first person then receives the remaining piece [49]. Such a game is categorized in the field of mechanism design, where the setting of the game gives players an incentive to achieve the desired outcome [50]. However, the procedure proposed by Steinhaus does not always yield fairness in a complex scenario setting since a person might behave more greedily to acquire more of the goods he desires. A procedure that is considered fair implies that the allocation of the goods should be performed in a manner where no person prefers the other person's share [51].

EGT shows in several examples, e.g., PD, Hawk/dove, Stag/hare [15], the tendency that cooperation is a better approach in the long run (an iterated relational game). At the same time, selfishness tends to be better in the single-step (transactional version of the game). These results of repeated games depend on the settings of problems and the utilization of different mechanisms that support the emergence of cooperation [6].

3.2. Modelling Choices

Two main choices are given when representing the population: continuous or discrete models. Continuous (aggregative) models describe the population using global statistics. The distribution of the genotypes and phenotypes in the population represents the individual's inherited behavior and the influence of the environment on the individual, respectively. Since the population's state is described as frequency data, the differences between individuals are lost in such a model. On the other hand, the discrete (agent-based) models maintain each individual's genotype/phenotype information in addition to other properties such as the location in the social network and spatial position [52].

The fundamental difference between the two models is in computational complexity. Aggregative models can be expressed as a set of differential/difference equations, making it possible to find the solution analytically. On the other hand, solving problems solely using analytical techniques is not feasible with discrete (agent-based) models. Therefore, one must run a series of computer simulations and employ Monte Carlo methods to yield the solution (i.e., convergence behavior). However, despite being computationally less demanding and heavily utilized in solving multiplayer games, aggregative models cannot be utilized for modeling structured relations. Human interactions within society are represented as structured interactions, i.e., humans are constrained to the network of social relationships. That means interactions with close ones and their respective groups will significantly impact future behavior, unlike random strangers [53,54]. Therefore, utilizing aggregative models for modeling human interactions would be detrimental because structured interactions between individuals produce different outcomes compared to unstructured interactions [55].

The introduction of the structure in evolutionary game-theoretic models dramatically influenced the model's long-term behavior [5,52]. Embedding human-like social interaction structure into the structure of the agent-based models enables forecasting the less divergent long-term behavior, which resembles the actual human population. Therefore, such evolutionary game-theoretic models can account for a wide variety of human behaviors predicting the outcomes of many cooperative ethical dilemma games elaborated above, such as the Prisoner's dilemma, Stag Hunt, and fair division in the Nash bargaining game [5,56]. It can be observed that, ultimately, the structure of society heavily influences the evolution of social norms [52].

4. Ethics and Coordination

If we put actors into (limited) material circumstances, we can expect that better-performing actors gain an advantage. In such circumstances, moral and ethical rules arise spontaneously to enable cooperation since greater coordinated groups are more effective than individuals if they have a similar developmental basis [17]. It is argued that cooperation helped the human race survive in a discrepancy with competitiveness [57]. Cooperation has been the basic organizational unit of the development of civilization since the time of hunter-gatherers [27].

Traffic is an excellent example of written and unwritten rules of conduct [58,59]. It is in the interest of every driver to cross the road from A to B as quickly and safely as possible. By refusing to follow the written rules, the driver risks being stopped by the police and losing his driving license (which, in this case, means expulsion from the game or losing the opportunity to participate). Failure to follow the unwritten rules carries the risk of condemnation, i.e., lousy will by other players or their refusal to cooperate. Well-engineered traffic rules enable the transport system to work effectively and at increased performance.

The application of the Prisoner's Dilemma is the same here. If the driver of car X drives in an unknown place to which he will never return, selfish behavior, such as taking away the advantage of and not letting other vehicles through, will bring him maximum short-term benefit. However, when the driver of car X does the same in the community where he is known, such behavior will bring him a bad reputation. Such stigma will negatively impact future rides regarding legal penalties and consequences outside the

ride, e.g., degraded relations with community members. Whether a person has selfish or altruistic interests, both people know that in expectation, it is most profitable to follow the rules [60]. Violation of the rules can bring a one-time benefit, i.e., overtaking in the opposite lane over the full line, if the necessary conditions are met, and the person is not fined or physically punished for this procedure. If a person repeats this procedure, the chances of a positive outcome are reduced, and the person risks being excluded from traffic and being punished with some form of legal penalty, which means that in the long run, it is unprofitable to break the set rules consistently. In this case, it is opportune to follow the rules to get a satisfactory result, i.e., to reach the ride's goal. Legal codes of conduct in traffic are found on almost every part of the road regarding prohibitions, permits, or warnings making traffic an excellent example of legally enforceable and supervised ethics. Moreover, behavioral rules of individuals in society are another, yet more subtle, example where ethical rules are unwritten. On the other hand, laws are an example of written applied ethics; however, they under-define human interactions, further honed with unwritten (traditional, habitual) rules.

Another example of ethics (and law) are community standards and rules, for example, in online circumstances. Facebook uses agent-based models to simulate the effects of different rules [61]. Ethicists try out different rules and test for consequences in the system. This is a form of consequentialist exploration whereby deontology is made based on rules' consequences (consequentially derived). Moreover, consequentialism relies on the principle of inherent cognitive limits unattainable to limited agents, especially in real time. On the other hand, a simple set of rules is easy to follow, even in real-time, for a limited agent. Therefore, it makes sense to invest considerable effort in moral innovation to pre-calculate offline straightforward sets of rules that can be quickly followed under more strict limitations. The principle of offline pre-calculation of ethical rules in conditions with enough time and computation resources is similar to planning and acting under time constraints.

Algorithmic Role and Utility of Human Values in Coordination

In addition to moral and legal obligations, there is also the issue of human values. Coordination is non-trivial, even hard to achieve. Mathematical-computational models and their analysis can reinforce the previous statement [62,63]. For example, the problem of finding Nash equilibrium is PPAD-complete; hence solving it might take prohibitively long. This is the case in a single [63] and iterated settings of problems [64], despite the folk theorem and abundance of Nash equilibria in the latter. Additionally, Nash equilibrium is achieved by rational actors only if they share beliefs about how the game is played. The rational actor model has no inherent mechanisms to enforce shared beliefs, so complex Nash equilibria do not arise spontaneously between rational agents. On the other hand, there is a concept of correlated equilibrium that is an appropriate equilibrium concept for social theory [65]. It is a generalization of Nash equilibrium which includes the correlation device in the model that induces correlated beliefs between the agents. Correlation devices can take the form of shared playing history, selection of players, public signals (like group symbols), etc. [66]. Additionally, finding a correlated equilibrium is much easier than Nash equilibrium as it can be done in polynomial time for any number of players and strategies in a broad class of games by using linear programming, even though finding the optimal one is still NP-hard [67].

Values are legally and morally undefined items individuals elevate, value, and cultivate because of cultural and personal prejudices [68]. We hypothesize that shared values ingrained in us through culture are an emerging phenomenon that helps coordinate in a fast heuristic fashion. This is in line with results suggesting that moral judgments are driven at least partly by imprecise heuristics and emotions [69]. Authors in [7] have mathematically modeled moral preferences by augmenting single-objective utility function with a weighted (scalarized) term for following personal norms (in addition to monetary outcomes).

Human values are an emerging concept that allows for easier coordination among like-minded people within a community. Suppose a person makes judgments based on a pre-judgment created by the human values defined above. There is an increased chance that the foundation will lead the person to a different conclusion from someone with different human values. If we have correlated values, we have a similar basis for decision-making, hence heuristically aiming for a correlated equilibrium.

Norms as sets of rules and conventions could also be a correlating device if they are simple enough to follow [17]. They should at least be explainable and comprehensible [70–73]. However, following many rules is certainly computationally hard, as constraint satisfaction problems from computer science can attest [74]. Using continuous fields of values enables using approximate-continuous instead of combinatorial reasoning, making it a very effective mechanism that can be seen in today's deep neural networks. If we were to use combinatorial reasoning in complex and fast situations, we would be paralyzed in decision-making under our cognitive limits, and coordination would be rare [75]. Even worse would be trying to calculate Nash equilibrium on the fly, outside the realm of games with a choreographer.

From a philosophical and psychological point of view, human values can be represented as a mixture of clustered criteria individuals use to evaluate actions, people, and events. Moreover, the Values Theory identifies ten distinct value orientations common among people in all cultures. Those values are derived from the human condition's three universal requirements: individuals' biological needs, requisites of coordinated social interaction, and groups' survival and welfare needs [76]. Individuals communicate these ten values with the remainder of the group to pursue their goals. According to the Values Theory, these goals are described as trans-situational and of varying importance serving as guiding principles in people's lives [77]. Figure 1 depicts the ten values in a circular arrangement so that the distance and antagonism of their underlying motivation are inversely proportional, i.e., two close values share a similar motivation, and two opposite values have opposing motivations. Moreover, values can be divided into two planes: self-enhancement (pursuit of self-interest) versus self-transcendence (concern for the interests of others) and openness (independence and openness to new experiences) versus conservation (resistance to change).

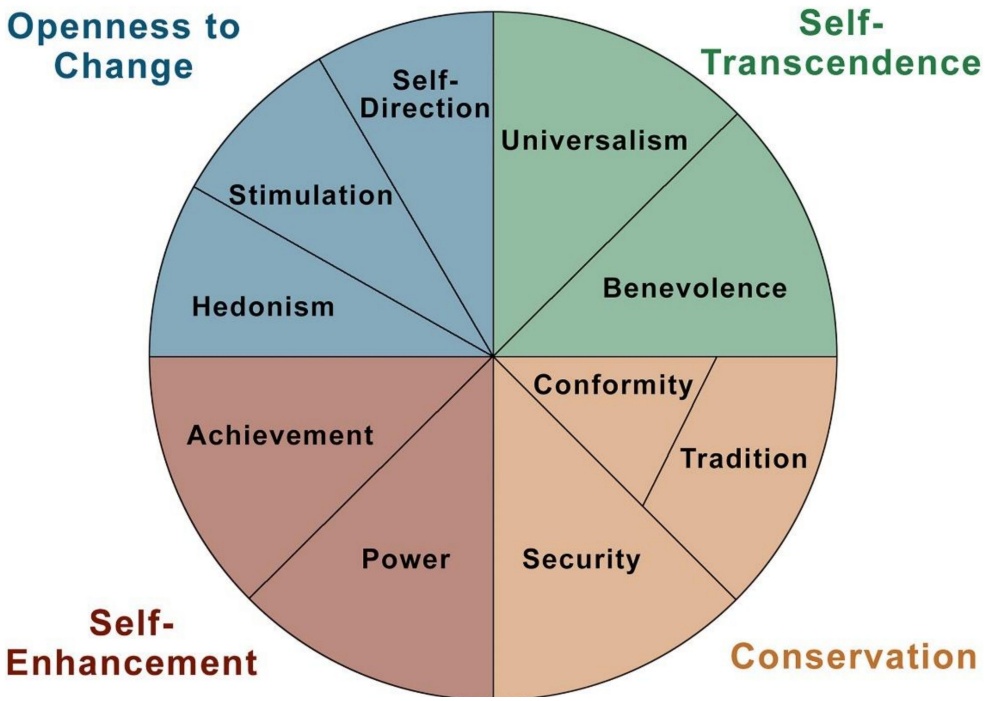

Figure 1. Ten motivational values and their relationships [76].

5. EGT as Modeling Basis

A commonly accepted hypothesis is that evolutionary processes shaped life on earth. Similarly, EGT can be used to determine the influence of external forces on heterogeneous ethics as its underlying component. Therefore, a possible explanation of those external forces is the necessity for cooperation as the basis of heterogeneous ethics on which all today's civilizations are built. It is argued that EGT is a realistic explanation of the material circumstances that preceded the creation of the first unwritten moral rules of cooperation [78,79].

EGT implements the three main pillars: (1) higher payoff strategies over time replace lower payoff strategies, also known as the "survival of the fittest"; (2) evolutionary change does not happen rapidly; (3) players' future actions are made reflexively without reasoning [80]. Biologists and mathematicians initially developed the evolutionary game to resolve open questions in evolutionary biology [81]. However, it has far-reaching implications in many areas, such as economics, ethics, industrial organization, policy analysis, and law. In general, evolutionary game models are suitable for systems where agents' behavior changes over time and interacts with the other agents influencing their behavior. However, the other agents must not collectively influence the behavior of an individual agent, and all decisions must be made reflexively [80]. One downside to original EGT is that players (agents) are born with a particular strategy that cannot be changed during their lifespan [82], making it unrealistic to model humans who evolve and change their strategy of social interactions throughout time. However, EGT, as it is currently defined, is suitable for modeling reptiles that do not have strong learning capabilities [83]. Simple, parametric forms of learning through memory and reputation mechanisms have been implemented in EGT, but it does not include richer lifetime learning due to computational complexity concerns. For modeling humans, extensions of EGT should be investigated to find concepts interpolated between stable evolutionary strategies and Nash equilibria since the first is reactive (without deliberation). In contrast, the second is unrealistically rational and computationally demanding. Correlated equilibrium is a good direction with a good balance of power and efficiency.

The usage of EGT as a basis for modeling morality has been extensively discussed and previously mentioned by Alexander in his The Structural Evolution of Morality, where he recognized the deficiencies of solely using EGT and proposed utilizing the combination of EGT, theory of bounded rationality, and research in psychology [52]. Although EGT enables the identification of behavior that maximizes the expected long-term utility, the motivation behind this behavior and the subsequent action that complies with the moral theory remains unexplained. To maximize an individual's lifetime utility, his actions must be bounded by rationality, requiring reliance on moral heuristics such as fair split and cooperation. Consequently, incorporating bounded rationality into one's actions complies with moral theory [84].

Authors in extended evolutionary synthesis propose improving systems focused on genetic evolution by considering the co-evolution of genome and culture. Cultural evolution alters the environment faced by genes, indirectly influencing natural selection. Adding social norms with the possibility of arbitration can substantially widen the range of successful cooperation [85]. This can explain the ultra-sociality of the human species. This co-evolution supposedly creates multiple equilibria, among which many are group-beneficial.

According to this line of thinking, in-group competition solves the free-rider problem with punishments, reputation, and signaling, which are mechanisms for large-scale cooperation. It sustains adherence to norms and settles the group into some correlated equilibrium. What is unique about these mechanisms is that they can sustain any costly behavior with or without communal benefit. They can sustain social norms that need not necessarily be cooperative [4].

Cultural evolution is a much faster and more innovative information processing system. Unlike genetic evolution, where there are two models for recombining traits, there are many more models simultaneously from which cultural traits interact. Additionally, transmission fidelity is much lower, and selection is strongly influenced by psychological

processes, which drives greater innovation [4]. As it is known, the success of strategies in a population is conditional on the populational distribution of other strategies, and these conditions can shift fast in changing. Using cultural learning, individuals can quickly adapt behavior to circumstances for which genetic learning is too slow by imitation learning and can keep cooperation from collapsing. Hence, culture may have created prolonged cooperation based on indirect reciprocity, which may have been just enough for genetic evolution to pick it up to develop supportive psychology to perpetuate it.

Cultural evolution is more likely to create inter-group competition since it is fast, noisy, and nonvertical compared to genetic evolution [86]. This competition puts groups against each other performance-wise, and it tends to lead to more prosocial norms and institutions. Competition at a lower level (of smaller groups) can help cooperation at higher levels (of greater collectives), and vice versa, stronger cooperation at a lower level can be detrimental to cooperation at a higher level [87]. Sometimes inter-group competition weakens kin bonds, reducing effectiveness at lower scales to promote effectiveness at higher scales [4].

EGT is somewhat successful in modeling social phenomena due to interactions between individuals trying to maximize utility. The emergence of altruism in an n-player prisoner's dilemma using EGT is proposed by [88]. Authors suggest that utilizing an EGT approach has been shown to help understand the inherited similarities between weak and strong altruism. The influence of social learning on human adaptability is discussed in [89]. By using the EGT approach to model the social learning of individuals through selective imitation, the authors supported the hypothesis. The development of social norms as an evolutionary process is another example of modeling social phenomena. Evolutionary psychologists argue that humans lack logical problem-solving skills [90]. Therefore, humans do not reason what is true or false when faced with reasoning; they match different patterns to a particular case. In [91], it was shown that human development is more consistent with cumulative cultural learners than with Machiavellian intelligence that tries to outmaneuver an opponent strategically. People will use previously learned reasoning that includes obligated, permitted, or forbidden actions. Social norms and, consequently, inheriting such reasoning can be justified using EGT [92]. Authors [31] describe how social norms, through sanctions, transform mixed-incentive games with social dilemmas where cooperative outcomes are unstable into easier coordination problems.

ESS conditioned on cues from public signals have been proven to be correlated equilibria of the game [93], and these equilibria can be found by repeated play [94]. Authors [58,66] model social norms that act as "choreographers" that induce correlated beliefs in agents, allowing them to coordinate on a correlated equilibrium of the game.

EGT and its extension to genetic-cultural co-evolution can model dynamics, progress, and limits. What is necessary is to incorporate cognition and more complex learning and strategies into cultural processes to make more precise dynamic change models. Additionally, a mixture of games should be modeled on a set of players, with uncertainty surrounding the specifics of the game played and outcomes. Such players would have evolving interests that depend on a selection model that mirrors the one in humans. Something along that line of thinking, but outside of ethical considerations, was done in machine learning to solve a large set of tasks with the same agent [95]. In social physics, some progress has been achieved in multi-games [96] and modeling more complex group dynamics with higher-order interactions [53].

6. Advanced Properties of Ethics

In the following section, we shall cover the views on the advanced properties of ethics in the literature. We deal with the structure underlying ethics as a patchwork of norms that are locally consistent. From this follows further issues of dilemmas through the global inconsistencies that jeopardize the coordination, especially in novel situations. Finally, we cover the importance of tension and balance between cooperation and competition in well-functioning societies. Although seemingly exclusively opposing forces, competition plays important role in innovation and cohesion within cooperation.

6.1. Social Norms as Behavioral Patterns

Ethics consist of behavioral patterns/regularities (social conventions, of which norms are a subset) that can be observed in resolutions of recurring coordination problems-situations [23,31] in a society of agents with similar capabilities. These patterns are emergent through time from the interactions in the environment. Under the assumption of evolutionary-guided changes (e.g., genetic-cultural co-evolution), all circumstances that often appeared in time were used for selective pressure [4]. For these reasons, it is expected that such norms would be locally consistent and good performing in frequent circumstances that led to their creation. Authors in [66] have shown that natural selection can be a blind choreographer that spontaneously creates beliefs and norms from stochastic events to serve as correlated equilibria without sophisticated knowledge or external enforcement. These beliefs and norms can be sustained using simpler and, later, more complex mechanisms [97].

Humans face various situations where certain decisions must be made during their lifetime. Such situations are simply a part of life, and we cannot avoid them. However, making decisions and confronting the resulting consequences is under our power. Throughout the evolution of humankind, individuals have been confronted with various decisions passed to and replicated by others over generations. Over time, the aggregation of these decisions led to the development of ethics, which can be compared to a patchwork, as depicted in Figure 2. Every patch in patchwork represents similar situations (episodic games) and belonging norms. However, the two neighboring patches are similar in problem space but have different norms that govern them. Additionally, the white patches represent the absence of norms in certain areas due to the absence of lived experience in that space. Examples of such white patches might involve significantly novel and impactful technology (such as super-intelligence). In that case, humans have to extrapolate norms from neighboring patches, i.e., similar ethical settings. The extrapolation, if it may be uniquely done in the first place, is not guaranteed good performance or relevance.

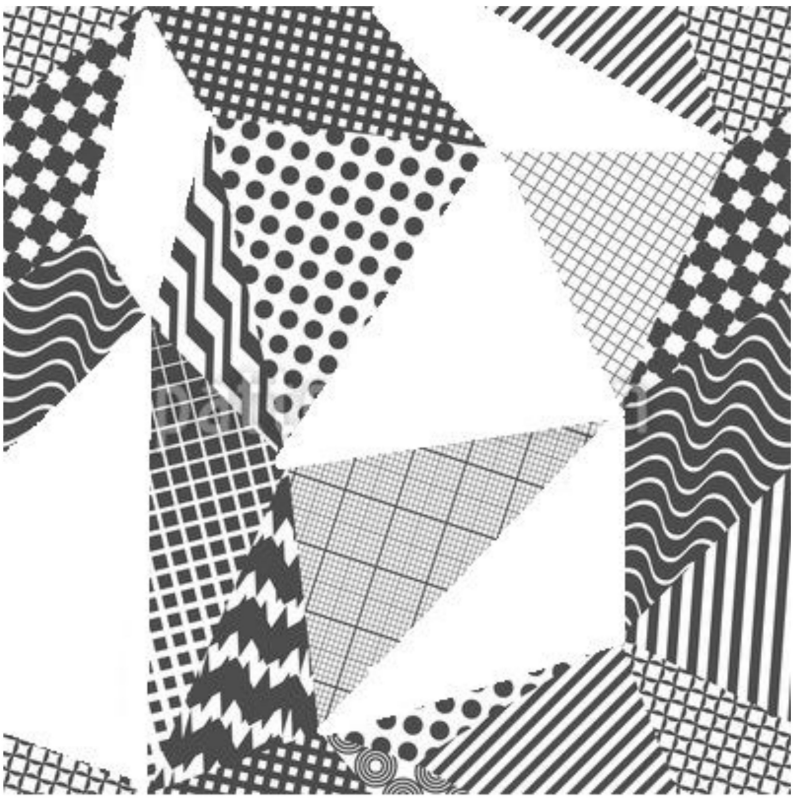

Figure 2. Ethical system—a patchwork of norms.

6.2. Consistency of Ethics

In addition to multicriteriality as a source of dilemmas for all normative ethics, deontological ethical systems may additionally experience dilemmas through inconsistencies. Two mechanisms yield inconsistency: norm confusion and faulty execution.

When extrapolating to substantially new situations from existing patterns, we might get norm confusion—inconsistencies between the different patches of locally consistent patterns. It is unclear which norm should be applied, and we get into a dilemma [31]. These inconsistencies are problematic for algorithmization and alignment with future AI systems. All existing ethics contain inconsistencies, with evident contradictions if looked at from a high-enough level. Such inconsistencies are not necessarily visible locally.

Faulty execution yields moral inconsistency in humans based purely on emotions associated with a particular case, i.e., moral dilemma, and not on formal inconsistencies. The root cause of such inconsistencies occurs when an individual, faced with dilemmas, treats the same moral cases differently. Moral learning is the process of learning from such mistakes and self-improvement, allowing the individual to maintain consistency with moral norms shared within a society. However, avoiding moral inconsistencies through moral learning is not always straightforward due to conflict with self-interest. Moreover, moral norms are generic, i.e., applied to a wide array of cases, and consequently, there will always be exceptions. For an individual (learning agent) to learn through moral problems on their own, one must think about moral problems from the other's perspective. For example, using Bayesian reasoning, one can derive a clear moral rule based on the judgments of other individuals [98].

To avoid biases when dealing with ethical decisions, philosopher John Rawls proposed the Veil of Ignorance as a tool for increasing personal consistency regarding some forms of faulty execution. Here, one should imagine sitting behind a veil of ignorance, keeping him away from his identity and personal circumstances. By being ignorant in such a manner, one can objectively make decisions. This would lead to a society that should help those who are socially or economically lacking behind [99] because robust optimization under total ignorance yields a maximin solution.

6.3. Cooperation vs. Competition

The question of the place of competition within well-functioning societies is open for investigation. It is argued that ethics based on cooperation brings more significant progress in the long run [100,101]. However, the relationship and the balance between the two are complex, even if the desired final goal is worldwide cooperation. Social physics exhibits a complex relationship between the emergence of the two that is very sensitive to the setting of the problem. Competition in society plays both an innovative and cohesive role in cooperation.

In addition to being the basis for the development of civilization, cooperation incorporates individual and social interests and helps create a balance among community members. On the micro-level, in civilized societies and everyday activities, cooperation with other community members is more profitable in the long run due to the installed norms and institutions. This makes personal goals faster and easier while achieving greater communal well-being. The opposite of cooperation is competitiveness, which in itself is not bad. Competition is one of the drivers of innovation, while cooperation is more effective at operational issues in repeated situations. It is good to be competitive with, for example, a past version of ourselves, set personal goals, and fight to achieve them. Additionally, competition is a cohesive element of cooperation.

According to the extended evolutionary synthesis, in-group competition is vital to solving the free-rider problem through punishment, reputation, and signaling mechanisms. Hence, it improves adherence to group norms. This efficiently leads to the correlated equilibrium.

Inter-group competition is important solely for correlated equilibrium selection, i.e., search. In line with theoretical results, searching for optimal correlative equilibrium is a

painfully slow process. It can be incomplete to remove group-damaging norms—especially when the latter are entangled with important cooperative norms. Additionally, the balance and distribution of competition and cooperation are sensitive, whereby competition on lower levels can favor cooperation at higher and stricter cooperation on lower levels can lead to collapse on a higher level. Such complex group dynamics can be modeled and tested on graphs [54] and hypergraphs [53].

In the long term, cooperation outweighs competition when relying on scarce resources. This is best described in Hardin's The Tragedy of the Commons [102], where each individual consumes resources at the expense of the others in a rivalrous fashion. If everyone acted solely upon their self-interest, the result would be a depletion of the common resources to everyone's detriment. The solution to the posed problem is the introduction of regulations by a higher authority or collective agreement, which leads to the correlated equilibrium [103]. Regulations could directly control the resource pool by excluding the individuals who excessively consume the resources or regulating consumption use. On the other hand, self-organized cooperative arrangements among individuals can rapidly overcome the problem (with a punishment mechanism for deviators). Here, the individuals share a common sense of collectivism, making their interest not to deplete all resources selfishly [102].

7. Conclusions

We have looked at ethics through an analytical prism to find some of its constitutive properties. The problem that ethics tries to solve is improving group performance in a setting that is multi-criteria, dynamic, and poised by uncertainties. Ethics operates on a large societal scale, making for a complex setting in which adaptability is crucial. Ethics emerged collaterally through cultural evolution on a longer time scale, meaning all changes have been slow and gradual. This seems to have worked well, but the shortening of timescales and greater societal perturbations due to rapid technological advances are jeopardizing the effectiveness of such an unguided process. Furthermore, current ethical systems are globally inconsistent, though they are locally consistent. This can lead to additional dilemmas that pose a further risk for the coordination, especially in novel situations. Then, we proceeded in the direction that could help with future work in guiding that process and reducing inherent risks—modeling and general computational/algorithmic issues. We must pick the appropriate model type and be wary of flaws in models of certain systems to remove them. Additionally, appropriate algorithmic approaches must be selected to circumvent the problems of computational complexity that could void the guiding efforts infeasible.

We argue that ethics is related to multi-agent interaction so that game theory can adequately model it, especially variants of evolutionary game theory. Moreover, correlated equilibrium is an important and appropriate concept that can be efficiently computationally found in the presence of a shared correlation device. Honed behavioral patterns—social norms—can play the role of a correlation device if they are simple enough to follow. However, following many rules is certainly computationally hard, as constraint satisfaction problems from computer science can attest [74]. Values can approximate complex ethical norms and thereby help the coordination by offering better correlation devices that reduce computational complexity.

Ethics is focused on cooperation, but it also depends on the competition for efficiency and adaptability. Moreover, the balance between competition and cooperation is delicate. The levels at which competition takes place significantly impact the level at which beneficial cooperation emerges, if at all. Mechanisms such as reputation, signaling, and punishment are elements of in-group competition that drive group cohesion. Social norms within-group competition play a crucial role as a correlation device that enables finding a correlated equilibrium into which a group may settle computationally efficiently. This is in stark contrast to the problem of finding Nash equilibrium which is PPAD-complete, and solving it might take a long time. However, there are no guarantees that the found correlated equilibrium benefits its group. Inter-group competition drives a slow search for better

equilibria. This slowness is in line with the results of the computational complexity theory. The optimal ethical system could be computationally found in principle, though at an impractically high computational cost.

8. Future Directions

Brcic and Yampolskiy [12] argue that ethics should stop being collateral and that through modeling, we can take more control over the process of ethical developments to obtain codifiable, more consistent, and adaptable ethics. The aim would be to make alignment easier within vast aggregates of agents (humanity, AI, inforgs [104]) through co-evolution between different constituents guided by a set of meta-principles. We propose agent-based models of genetic and cultural co-evolution similar to [105]. However, agents in these models should also be equipped and amplified with basic cognition and more complex learning and strategies as idealizations to which we may strive in simulations according to available computational and algorithmic resources. This considerable computational effort can be invested in moral innovation to pre-calculate offline straightforward sets of rules, norms, and values that can be quickly and stably followed under more strict real-time limitations. More generally, findings from this research direction might benefit decentralized computing on an unprecedented scale and heterogeneity. If this is possible and practical, it remains to be seen.

Author Contributions: Conceptualization, S.I. and M.B.; methodology, S.I.; validation, M.B.; investigation, S.I. and I.V.; writing—original draft preparation, S.I., I.V. and K.P.; writing—review and editing, K.P. and M.B.; supervision, M.B.; project administration, S.I.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Purzycki, B.G.; Lang, M.; Henrich, J.; Norenzayan, A. Guiding the Evolution of the Evolutionary Sciences of Religion: A Discussion. *Relig. Brain Behav.* **2022**, *12*, 226–232. [CrossRef]
2. Wittgenstein, L. Philosophical Investigations. Schulte, J., Ed.; Wiley-Blackwell: Chichester, UK; Malden, MA, USA, 1953; ISBN 978-1-4051-5928-9.
3. Bowles, S.; Gintis, H. *A Cooperative Species: Human Reciprocity and Its Evolution*, 2nd ed.; Princeton University Press: Princeton, NJ, USA, 2013; ISBN 978-0-691-15816-7.
4. Henrich, J.; Muthukrishna, M. The Origins and Psychology of Human Cooperation. *Annu. Rev. Psychol.* **2021**, *72*, 207–240. [CrossRef] [PubMed]
5. Jusup, M.; Holme, P.; Kanazawa, K.; Takayasu, M.; Romić, I.; Wang, Z.; Geček, S.; Lipić, T.; Podobnik, B.; Wang, L.; et al. Social Physics. *Phys. Rep.* **2022**, *948*, 1–148. [CrossRef]
6. Capraro, V.; Perc, M. Grand Challenges in Social Physics: In Pursuit of Moral Behavior. *Front. Phys.* **2018**, *6*, 107. [CrossRef]
7. Capraro, V.; Perc, M. Mathematical Foundations of Moral Preferences. *J. R. Soc. Interface* **2021**, *18*, 20200880. [CrossRef]
8. Bowles, S. *The Moral Economy: Why Good Incentives Are No Substitute for Good Citizens*; Yale University Press: New Haven, London, UK, 2016; ISBN 978-0-300-16380-3.
9. Broome, J. *Ethics out of Economics*; Cambridge University Press: Cambridge, UK, 1999; ISBN 978-0-521-64275-0.
10. Helbing, D. *Next Civilization: Digital Democracy and Socio-Ecological Finance—How to Avoid Dystopia and Upgrade Society by Digital Means*, 2nd ed.; Springer: Cham, Switzerland, 2021; ISBN 978-3-030-62329-6.
11. van den Hoven, J.; Miller, S.; Pogge, T. (Eds.) *Designing in Ethics*; Cambridge University Press: Cambridge, UK, 2019; ISBN 978-0-521-13525-2.
12. Brcic, M.; Yampolskiy, R.V. Impossibility Results in AI: A Survey. *arXiv* **2021**, arXiv:2109.00484. *preprint*
13. Becker, G.S.; Murphy, K.M. The Division of Labor, Coordination Costs, and Knowledge. *Q. J. Econ.* **1992**, *107*, 1137–1160. [CrossRef]
14. Ricardo, D. *On the Principles of Political Economy and Taxation*; Liberty Fund, Inc.: Indianapolis, IN, USA, 2004; ISBN 978-0865979659.
15. Dawkins, R. *The Selfish Gene*; Oxford University Press: New York, NY, USA, 1976; ISBN 978-0-19-857519-1.
16. Thompson, P. Evolutionary Ethics: Its Origin and Contemporary Face. *Zygon®* **1999**, *34*, 473–484. [CrossRef]

17. Axelrod, R.M. *The Evolution of Cooperation*; Basic Books, Inc.: New York, NY, USA, 1984; ISBN 978-0-465-00564-2.
18. Boyd, R.; Richerson, P.J. *Culture and the Evolutionary Process*; University of Chicago Press: Chicago, IL, USA, 1988; ISBN 978-0-226-06933-3.
19. Lehmann, L.; Keller, L. Synergy, Partner Choice and Frequency Dependence: Their Integration into Inclusive Fitness Theory and Their Interpretation in Terms of Direct and Indirect Fitness Effects. *J. Evol. Biol.* **2006**, *19*, 1426–1436. [CrossRef]
20. Wu, S.A.; Wang, R.E.; Evans, J.A.; Tenenbaum, J.B.; Parkes, D.C.; Kleiman-Weiner, M. Too Many Cooks: Bayesian Inference for Coordinating Multi-Agent Collaboration. *Top. Cogn. Sci.* **2021**, *13*, 414–432. [CrossRef]
21. Assaad, L. The Structural Evolution of Cooperation: Can Evolutionary Game Theory Teach Us About Morality? *Rerum Causae* **2021**, *12*.
22. Wheeler, M.A.; McGrath, M.J.; Haslam, N. Twentieth Century Morality: The Rise and Fall of Moral Concepts from 1900 to 2007. *PLoS ONE* **2019**, *14*, e0212267. [CrossRef] [PubMed]
23. Lewis, D. *Convention: A Philosophical Study*; John Wiley & Sons: Hoboken, NJ, USA, 1969; ISBN 978-0-470-69296-7.
24. Petrinovich, L.F. *Human Evolution, Reproduction, and Morality*; MIT Press: Cambridge, MA, USA, 1998; ISBN 978-0-262-66143-0.
25. Allchin, D. The Evolution of Morality. *Evol. Educ. Outreach* **2009**, *2*, 590–601. [CrossRef]
26. Distin, K. *Cultural Evolution*; Cambridge University Press: Cambridge, UK, 2011; ISBN 978-0-521-18971-2.
27. Diamond, J.; Renfrew, C. Guns, Germs, and Steel: The Fates of Human Societies. *Nature* **1997**, *386*, 339.
28. Foot, P. The Problem of Abortion and the Doctrine of the Double Effect. *Oxf. Rev.* **1967**, *5*, 5–15.
29. Kleinberg, J.; Mullainathan, S.; Raghavan, M. Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv* **2016**, arXiv:1609.05807. *preprint*
30. Arrow, K.J. A Difficulty in the Concept of Social Welfare. *J. Polit. Econ.* **1950**, *58*, 328–346. [CrossRef]
31. Bicchieri, C. *The Grammar of Society: The Nature and Dynamics of Social Norms*; Cambridge University Press: Cambridge, UK, 2005; ISBN 978-0-521-57372-6.
32. Bernard, K.N.; Gibson, M.L. Professional Misconduct by Mentally Imparied Attorneys: Is There a Better Way to Treat an Old Problem Current Developments 2003–2004. *Georget. J. Leg. Ethics* **2003**, *17*, 619–636.
33. Sligo, F.; Stirton, N. Does Hindsight Bias Change Perceptions of Business Ethics? *J. Bus. Ethics* **1998**, *17*, 111–124. [CrossRef]
34. Kinnier, R.T.; Kernes, J.L.; Dautheribes, T.M. A Short List of Universal Moral Values. *Couns. Values* **2000**, *45*, 4–16. [CrossRef]
35. Awad, E.; Dsouza, S.; Shariff, A.; Rahwan, I.; Bonnefon, J.-F. Universals and Variations in Moral Decisions Made in 42 Countries by 70,000 Participants. *Proc. Natl. Acad. Sci. USA* **2020**, *17*, 2332–2337. [CrossRef] [PubMed]
36. Hanun Rodríguez, O.; Ximénez Camilli, C. The Neurobiological and Environmental Origin of Ethics: Analysis of Biological, Social and Religious Determinism. *Bioeth. Update* **2018**, *4*, 92–102. [CrossRef]
37. Darwin, C. *The Descent of Man, and Selection in Relation to Sex*; Princeton University Press: Princeton, NJ, USA, 2008; ISBN 978-1-4008-2006-1.
38. Du, W.; Ding, S. A Survey on Multi-Agent Deep Reinforcement Learning: From the Perspective of Challenges and Applications. *Artif. Intell. Rev.* **2021**, *54*, 3215–3238. [CrossRef]
39. Yang, Y.; Wang, J. An Overview of Multi-Agent Reinforcement Learning from Game Theoretical Perspective. *arXiv* **2021**, arXiv:2011.00583. *preprint*
40. Henrich, J. Cultural Group Selection, Coevolutionary Processes and Large-Scale Cooperation. *J. Econ. Behav. Organ.* **2004**, *53*, 3–35. [CrossRef]
41. Nowak, M.A. Five Rules for the Evolution of Cooperation. *Science* **2006**, *314*, 1560–1563. [CrossRef]
42. Stewart, A.J.; Plotkin, J.B. Collapse of Cooperation in Evolving Games. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 17558–17563. [CrossRef]
43. Press, W.H.; Dyson, F.J. Iterated Prisoner's Dilemma Contains Strategies That Dominate Any Evolutionary Opponent. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 10409–10413. [CrossRef]
44. Stewart, A.J.; Plotkin, J.B. From Extortion to Generosity, Evolution in the Iterated Prisoner's Dilemma. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 15348–15353. [CrossRef]
45. Ellickson, R.C. *Order without Law: How Neighbors Settle Disputes*; Harvard University Press: Cambridge, MA, USA, 1991; pp. 145–148, ISBN 978-0-674-64169-3.
46. Skyrms, B. *The Stag Hunt and the Evolution of Social Structure*; Cambridge University Press: Cambridge, UK, 2004; ISBN 978-0-521-53392-8.
47. Skyrms, B. *The Stag Hunt*; American Philosophical Association: Philadelphia, PA, USA, 2001; Volume 75, pp. 31–41. [CrossRef]
48. Crawford, V.P. Fair Division. In *Game Theory*; Eatwell, J., Milgate, M., Newman, P., Eds.; Palgrave Macmillan UK: London, UK, 1989; ISBN 978-1-349-20181-5.
49. Steinhaus, H. The Problem of Fair Division. *Econometrica* **1948**, *16*, 101–104.
50. Myerson, R.B. Mechanism Design. In *Allocation, Information and Markets*; Eatwell, J., Milgate, M., Newman, P., Eds.; The New Palgrave; Palgrave Macmillan UK: London, UK, 1989; pp. 191–206, ISBN 978-1-349-20215-7.
51. Foley, D.K. *Resource Allocation and the Public Sector*; Yale University: New Haven, CT, USA, 1967.
52. Alexander, J.M. *The Structural Evolution of Morality*; Cambridge University Press: Cambridge, UK, 2010; ISBN 978-0-521-15269-3.
53. Majhi, S.; Perc, M.; Ghosh, D. Dynamics on Higher-Order Networks: A Review. *J. R. Soc. Interface* **2022**, *19*, 20220043. [CrossRef] [PubMed]

54. Szabó, G.; Fáth, G. Evolutionary Games on Graphs. *Phys. Rep.* **2007**, *446*, 97–216. [CrossRef]
55. Nowak, M.; May, R. Evolutionary Games and Spatial Chaos. *Nature* **1992**, *359*, 826–829. [CrossRef]
56. Durrett, R.; Levin, S. The Importance of Being Discrete (and Spatial). *Theor. Popul. Biol.* **1994**, *46*, 363–394. [CrossRef]
57. Veit, W.; Browning, H. Why Socio-Political Beliefs Trump Individual Morality: An Evolutionary Perspective. *AJOB Neurosci.* **2020**, *11*, 290–292. [CrossRef]
58. Gintis, H. Social Norms as Choreography. *Polit. Philos. Econ.* **2010**, *9*, 251–264. [CrossRef]
59. Helbing, D.; Huberman, B.A. Coherent Moving States in Highway Traffic. *Nature* **1998**, *396*, 738–740. [CrossRef]
60. Mueller, D.C. Rational Egoism versus Adaptive Egoism as Fundamental Postulate for a Descriptive Theory of Human Behavior. *Public Choice* **1986**, *51*, 3–23. [CrossRef]
61. Ahlgren, J.; Berezin, M.E.; Bojarczuk, K.; Dulskyte, E.; Dvortsova, I.; George, J.; Gucevska, N.; Harman, M.; Lämmel, R.; Meijer, E.; et al. WES: Agent-Based User Interaction Simulation on Real Infrastructure. In Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops, Seoul, Republic of Korea, 27 June–19 July 2020; pp. 276–284, ISBN 978-1-4503-7963-2.
62. Aaronson, S. Why Philosophers Should Care About Computational Complexity. *Comput. Turing Gödel Church Beyond* **2013**, *261*, 327.
63. Daskalakis, C.; Goldberg, P.W.; Papadimitriou, C.H. The Complexity of Computing a Nash Equilibrium. *SIAM J. Comput.* **2009**, *39*, 195–259. [CrossRef]
64. Borgs, C.; Chayes, J.; Immorlica, N.; Kalai, A.T.; Mirrokni, V.; Papadimitriou, C. The Myth of the Folk Theorem. In Proceedings of the fortieth annual ACM symposium on Theory of computing, Victoria, Canada, 17–20 May 2008; pp. 365–372.
65. Gintis, H. *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*, 1st ed.; Princeton University Press: Princeton, NJ, USA, 2009; ISBN 978-0-691-14052-0.
66. Morsky, B.; Akçay, E. Evolution of Social Norms and Correlated Equilibria. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 8834–8839. [CrossRef] [PubMed]
67. Papadimitriou, C.H.; Roughgarden, T. Computing Correlated Equilibria in Multi-Player Games. *J. ACM* **2008**, *55*, 1–29. [CrossRef]
68. Mead, B. The Meaning of Value: A Review and Analysis of A Sociological Concept. Master's Thesis, College of William & Mary: Williamsburg, VA, USA, 1976. [CrossRef]
69. Capraro, V.; Rand, D.G. Do the Right Thing: Experimental Evidence That Preferences for Moral Behavior, Rather than Equity or Efficiency per Se, Drive Human Prosociality. *Judgm. Decis. Mak.* **2018**, *13*, 99–111. [CrossRef]
70. Dosilovic, F.K.; Brcic, M.; Hlupic, N. Explainable Artificial Intelligence: A Survey. In Proceedings of the 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 21–25 May 2018; pp. 0210–0215.
71. Juric, M.; Sandic, A.; Brcic, M. AI Safety: State of the Field through Quantitative Lens. In Proceedings of the 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia, 28 September–2 October 2020; pp. 1254–1259.
72. Krajna, A.; Brcic, M.; Kovac, M.; Sarcevic, A. Explainable Artificial Intelligence: An Updated Perspective. In Proceedings of the Proceedings of 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO) 2022, Opatija, Croatia, 23–27 May 2022.
73. Krajna, A.; Brcic, M.; Lipic, T.; Doncevic, J. Explainability in Reinforcement Learning: Perspective and Position. *arXiv* **2022**, arXiv:2203.11547. *preprint*
74. Gallardo, J.E.; Cotta, C.; Fernández, A.J. Solving Weighted Constraint Satisfaction Problems with Memetic/Exact Hybrid Algorithms. *J. Artif. Intell. Res.* **2009**, *35*, 533–555. [CrossRef]
75. Talbert, B. Overthinking and Other Minds: The Analysis Paralysis. *Soc. Epistemol.* **2017**, *31*, 545–556. [CrossRef]
76. Schwartz, S.H. Basic Human Values: An Overview. *Rev. Française Sociol.* **2006**, *47*, 249–288.
77. Schwartz, S.H. Values and Culture. In *Motivation and Culture*; Routledge: New York, NY, USA, 1997; pp. 69–84. ISBN 978-0-415-91509-0.
78. O'Connor, C. *Games in the Philosophy of Biology*; Cambridge University Press: Cambridge, UK, 2020; ISBN 978-1-108-61673-7.
79. Cavagnetto, S.; Gahir, B. Game Theory—Its Applications to Ethical Decision Making. *CRIS—Bull. Cent. Res. Interdiscip. Study* **2014**, *2014*, 1–19. [CrossRef]
80. Friedman, D. On Economic Applications of Evolutionary Game Theory. *J. Evol. Econ.* **1998**, *8*, 15–43. [CrossRef]
81. Smith, J.M. *Evolution and the Theory of Games*; Cambridge University Press: Cambridge, UK, 1982; ISBN 978-0-521-28884-2.
82. Dugatkin, L.A.; Reeve, H.K. *Game Theory and Animal Behavior*; Oxford University Press: New York, NY, USA, 2000; ISBN 978-0-19-513790-3.
83. Matsubara, S.; Deeming, D.C.; Wilkinson, A. Cold-Blooded Cognition: New Directions in Reptile Cognition. *Curr. Opin. Behav. Sci.* **2017**, *16*, 126–130. [CrossRef]
84. Veit, W. Modeling Morality. In *Model-Based Reasoning in Science and Technology*; Fontaine, M., Barés-Gómez, C., Salguero-Lamillar, F., Magnani, L., Nepomuceno-Fernández, Á., Eds.; Springer: Berlin/Heidelberg, Germany, 2019; pp. 83–102.
85. Mathew, S.; Boyd, R.; Van Veelen, M. Human Cooperation among Kin and Close Associates May Require Enforcement of Norms by Third Parties. In *Cultural Evolution*; The MIT Press: Cambridge, MA, USA, 2013; ISBN 978-0-262-01975-0.

86. Boyd, R.; Richerson, P.J.; Henrich, J. Rapid Cultural Adaptation Can Facilitate the Evolution of Large-Scale Cooperation. *Behav. Ecol. Sociobiol.* **2011**, *65*, 431–444. [CrossRef]
87. Muthukrishna, M.; Doebeli, M.; Chudek, M.; Henrich, J. The Cultural Brain Hypothesis: How Culture Drives Brain Expansion, Sociality, and Life History. *PLoS Comput. Biol.* **2018**, *14*, e1006504. [CrossRef]
88. Fletcher, J.A.; Zwick, M. The Evolution of Altruism: Game Theory in Multilevel Selection and Inclusive Fitness. *J. Theor. Biol.* **2007**, *245*, 26–36. [CrossRef]
89. Kameda, T. Does Social/Cultural Learning Increase Human Adaptability? Rogers's Question Revisited. *Evol. Hum. Behav.* **2003**, *24*, 242–260. [CrossRef]
90. Clark, A.; Karmiloff-Smith, A. The Cognizer's Innards: A Philosophical and Psychological Perspective on the Development of Thought. *Mind Lang.* **1993**, *8*, 487–519. [CrossRef]
91. Baimel, A.; Juda, M.; Birch, S.; Henrich, J. Machiavellian Strategist or Cultural Learner? Mentalizing and Learning over Development in a Resource-Sharing Game. *Evol. Hum. Sci.* **2021**, *3*, E14. [CrossRef]
92. Ostrom, E. Collective Action and the Evolution of Social Norms. *J. Econ. Perspect.* **2000**, *14*, 137–158. [CrossRef]
93. Cripps, M. Correlated Equilibria and Evolutionary Stability. *J. Econ. Theory* **1991**, *55*, 428–434. [CrossRef]
94. Arifovic, J.; Boitnott, J.F.; Duffy, J. Learning Correlated Equilibria: An Evolutionary Approach. *J. Econ. Behav. Organ.* **2019**, *157*, 171–190. [CrossRef]
95. Reed, S.; Zolna, K.; Parisotto, E.; Colmenarejo, S.G.; Novikov, A.; Barth-Maron, G.; Gimenez, M.; Sulsky, Y.; Kay, J.; Springenberg, J.T.; et al. A Generalist Agent. *arXiv* **2022**, arXiv:2205.06175. *preprint* [CrossRef]
96. Li, X.; Hao, G.; Wang, H.; Xia, C.; Perc, M. Reputation Preferences Resolve Social Dilemmas in Spatial Multigames. *J. Stat. Mech. Theory Exp.* **2021**, *2021*, 013403. [CrossRef]
97. Bhui, R.; Chudek, M.; Henrich, J. How Exploitation Launched Human Cooperation. *Behav. Ecol. Sociobiol.* **2019**, *73*, 78. [CrossRef]
98. Campbell, R. Learning from Moral Inconsistency. *Cognition* **2017**, *167*, 46–57. [CrossRef]
99. Rawls, J. *A Theory of Justice*; Harvard University Press: Cambridge, MA, USA, 1971; ISBN 978-0-674-88010-8.
100. Cliquet, R.; Avramov, D. *Evolution Science and Ethics in the Third Millennium: Challenges and Choices for Humankind*, 1st ed.; Springer: Berlin/Heidelberg, Germany, 2018; ISBN 978-3-319-73090-5.
101. Curry, O.S. Morality as Cooperation: A Problem-Centred Approach. In *The Evolution of Morality*; Shackelford, T.K., Hansen, R.D., Eds.; Evolutionary Psychology; Springer International Publishing: Cham, Switzerland, 2016; pp. 27–51, ISBN 978-3-319-19671-8.
102. Hardin, G. The Tragedy of the Commons: The Population Problem Has No Technical Solution; It Requires a Fundamental Extension in Morality. *Science* **1968**, *162*, 1243–1248. [CrossRef]
103. Ostrom, E.; Gardner, R.; Walker, J. *Rules, Games, and Common-Pool Resources*; University of Michigan Press: Ann Arbor, MI, USA, 1994; ISBN 978-0-472-06546-2.
104. Floridi, L. *Philosophy and Computing: An Introduction*, 1st ed.; Routledge: London, UK, 1999; ISBN 978-0-415-18024-5.
105. Paolucci, M.; Kossman, D.; Conte, R.; Lukowicz, P.; Argyrakis, P.; Blandford, A.; Bonelli, G.; Anderson, S.; de Freitas, S.; Edmonds, B.; et al. Towards a Living Earth Simulator. *Eur. Phys. J. Spec. Top.* **2012**, *214*, 77–108. [CrossRef]