AI Modelling of Counterfactual Thinking for Judicial Reasoning and Governance of Law

Luís Moniz Pereira[1], Francisco C. Santos[2], António Barata Lopes[3]

[1]NOVA-LINCS, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa,
2829-516 Monte da Caparica, Portugal
lmp@fct.unl.pt (+351) 212 948 536
[2]INESC-ID and Instituto Superior Técnico, Universidade de Lisboa,
Tagus Park, 2744-016 Porto Salvo, Portugal
franciscocsantos@tecnico.ulisboa.pt (+351) 210 407 091
[3]ANQEP – "Agência Nacional para a Qualificação e Ensino Profissional" and Agrupamento de Escolas de Alvalade,
Av. 24 de Julho 138, 1200-771 Lisboa, Portugal
lopesab@msn.com (+351) 213 943 700

Abstract

When speaking of moral judgment, we refer to a function of recognizing appropriate or condemnable actions and the possibility of choice between them by agents. Their ability to construct possible causal sequences enables them to devise alternatives in which choosing one implies setting aside others. This internal deliberation requires a cognitive ability, namely that of constructing counterfactual arguments. These serve not just to analyse possible futures, being prospective, but also to analyse past situations, by imagining the gains or losses resulting from alternatives to the actions actually carried out, given evaluative information subsequently known.

Counterfactual thinking is in thus a prerequisite for AI agents concerned with Law cases, in order to pass judgement and, additionally, for evaluation of the ongoing governance of such AI agents. Moreover, given the wide cognitive empowerment of counterfactual reasoning in the human individual, namely in making judgments, the question arises of how the presence of individuals with this ability can improve cooperation and consensus in populations of otherwise self-regarding individuals.

Our results, using Evolutionary Game Theory (EGT), suggest that counterfactual thinking fosters coordination in collective action problems occurring in large populations and has limited impact on cooperation dilemmas in which such coordination is not required.

Keywords Counterfactual Thinking, Evolutionary Game Theory, AI Governance, Judicial Reasoning.

1. Introduction and Motivation

The Law clearly says that its theory of causation is counterfactual dependency (Moore 2009, p. 371). The focus on counterfactual theory lies in morality. The social minimum is that we do no harm. Our moral responsibility is naturally captured by a certain kind of counterfactual test, one that compares how the world is after our actions with how the world would have been if, contrary to fact, we had not done the actions in question. Similarly, one can reason about alternative actions which would have improved the world or produced a greater good (Roese and Olson 1995; Pereira and Saptawijaya 2016a, 2016b, 2017).

The class of statements we deem counterfactual are conditional statements conjoined with the falsity of both the antecedent and their consequent clauses (Pearl 2010). Counterfactuals, possibility, and the hypothetical are part of the genesis of what there is, and what there is is what it is because it was *otherwise*. In (Dietz Saldanha et al. 2015, 2021) we also consider conditionals whose antecedents are *unknown* and evaluate the conditional by applying revision and abduction in order to satisfy it. The laws of physics, for example, can be interpreted as counterfactual assertions, such as 'Had the weight on this spring doubled, its length would have doubled as well' (Hooke's Law) (Pearl and Mackenzie 2018).

Causation as a prerequisite to legal liability is intimately related to causation as a natural relation lying at the heart of scientific explanation. Moral responsibility supervenes on natural properties like causation, intention, and the like. The counterfactual theory of causal relations is dominant in both Law and recent Philosophy. We are more blameworthy when we cause some evil, than merely trying to cause it. We experience regret when we have caused some harm even though we were not at all culpable. It is not regret but guilt that disturbs us, in those cases we judge ourselves to be blameworthy. It is guilt, not regret, that is consistent with such self-judgements (Moore 2009, p. vi, vii, 30-32).

When people are moved to think counterfactually, they generally think about how things might have turned out better ('upward counterfactuals' in the parlance of experimental psychology). When thinking how events might have been worse, one speaks of 'downward counterfactuals' (Byrne 2005). Already the Greek Lysias, in the aftermath of the Peloponnesian war (382 B.C) says "if we had remained united and every man had done as I did, the oligarchy and civil war would not have happened. Another Greek, the historian and general Thucydides, not only emphasizes how terrible the war really was but underlines moments when it might have been worse for Athens and its citizens (Tordoff 2014, p. 116).

According to judgement dissociation theory, upwards counterfactuals tend to focus on the functional goal of identifying ways in which a negative outcome would have been prevented. These thoughts can undo outcomes not only by negating direct causes, but also by negating enabling conditions or adding in disabling conditions. This suggests there are more ways an actor could prevent an outcome than ways it could cause it. Hence, self-implicating upward counterfactuals are likely to draw attention to blame-implicating actions. Research suggests that prison programs

designed to stimulate and explore prisoners' upward counterfactual thoughts about their crime, arrest, conviction, and sentence may increase prisoners' attributions of self-blame, and enhance their feelings of guilt (Mandel et al. 2005). Our own theoretical study of guilt (Pereira et al. 2017), grounded on Evolutionary Game Theory (EGT), provides evidence that, in a population wherein there exists from the start a modicum of guilt-feeling agents, a better cooperation tends to arise as guilt tends to spread.

For decades or even centuries, lawyers have used a relatively straightforward test of a defendant's culpability called 'but-for-causation': "The injury would not have occurred *but-for* the defendant's action." Given just the conditional "If a defendant does action A, then injury I follows," its related counterfactual can promote the antecedent to a cause of the consequent: "If the defendant would not have done action A, then injury I would not have occurred." *But-for* clauses can also be indirect. If Joe blocks a building's fire exit with furniture, and Judy dies after she could not reach the exit, then Joe is legally responsible for her death even though he did not light the fire (Pearl and Mackenzie 2018). Similarly, the central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, natural origin, etc.) (Greiner 2008).

Recent social and cognitive psychology theories propose a 'dual-processing' mental architecture. Most of what the mind does is achieved by quick, automatic, heuristic-laden processing, our visual system being an example. This first cognitive system is often called the automatic system, or intuitive system, or simply 'system 1'. But occasionally, we need to think about a problem, consider counterfactual situations, entertain suppositions, weigh possibilities, and consciously decide upon a solution. This sort of thinking, is slow, laboured, and easily disrupted by other tasks; it is sometimes called the reasoning system, or controlled processing, or simply 'system 2'. Yet, there is nothing about system 2 that precludes the conscious deliberate use of heuristics, colloquially referred to as rules of thumb, a staple domain of study in AI. Laws and legislative procedures may induce people to use both systems (Gigerenzer and Engel 2006).

In (Pereira and Saptawijaya 2016b, 2017; Pereira and Santos 2019; Pereira and Lopes 2020a, 2020b), we have examined how counterfactual reasoning can be employed to discuss moral responsibility and, moreover, shown how it can be utilised to henceforth produce greater good and avoid harm, after knowing the joint outcomes of one's and another's actions in abstract social games.

In this chapter, we concentrate on using EGT to evince why and how AI regulated counterfactual reasoning can be a promoter of cooperation within a population, and on its incidence in the domain of Law governance and Law application. We will not address in detail the issue of governance of AI innovation by the Law, for we have done so elsewhere (Han et al. 2020, 2021, 2022; Cimpeanu et al. 2022) but we provide, in section 5, an outline of the issues of such AI regulation.

The remainder is organized as follows. Firstly, we recall some societal and historical background with regard to alternative pasts and prospective futures. Next, we provide basic notions about counterfactual reasoning. That is followed by its use

in evolutionary game theory models, intuitively illustrated with the well-known Stag-Hunt example (Skyrms 2004). Henceforth, we make the case for the use of counterfactual reasoning in law, namely in what regards improved joint Plea Bargaining, by analogy with the Stag-Hunt game, and elaborate on its positive juridical consequences. Thereafter, we delve in more detail into the usage of counterfactual thinking in evolutionary games modelling, and finally conclude with some remarks.

2. Some Societal and Historical Background

Living in a better society first requires conjecturing what that better society might be. Now, this task is not at all easy. Throughout History, human beings have always been imagining utopias. When we think of Plato's ideal *Republic*, or St. Augustine's *City of God*, or Thomas Moro's *Utopia*, or Karl Marx's *Classless Society*, we are always a long way from concrete societies. Throughout our History we have inhabited the *world-as-it-is*, but imagining alternatives that would make it better. This dialectic game between the descriptive domain and the prescriptive realm has been extremely rich and fruitful. Of course, we have never achieved any utopia so far; moreover, we are not sure whether, had we done so, it would have been good for humanity. Still, for better or worse, utopias have played a key role in our individual and collective decisions.

From a collective standpoint, they have provided an elicitation model for what we imagine the ideal destination to be. We are used to thinking that having a destination, or a comprehensive purpose, is highly positive. However, this goal has also given rise to much violence between groups with opposing interests. Suffice to think of the various Proletarian Dictatorships that have proliferated across this planet, and how, under the possible pretext of creating an egalitarian and just society, they have sanctioned acts of extreme violence, with massive killings of human beings. On the other hand, without a range of possible utopias, we would be relatively lost, because we would not have enough diversity in the answer to the collective question of where we wish to go. We need this diversity not to become dependent on just one possibility. Imagine a single answer –religious in nature, say– to this question. It will not be accepted by all believers, let alone by non-believers.

Even without reaching a consensus on what an ideal society is and accepting the idea that multiple conjectures about it can coexist, we will unreservedly agree that human societies should not be used as a pretext for the enrichment of a meagre ten percent of the world's population. Nor is it likely that consuming all, each one would give credible meaning to our individual and collective lives. However, this is what we are witnessing more and more. That means we are treading dangerous paths, both in the field of our capacities for idealization (or lack thereof), and in the realm of what –concretely– we are doing to try and improve the present.

Reflecting on these issues requires the exercise of critical thinking, a capacity we acknowledge to be rare. Indeed, the data from Social Psychology is quite emblematic in this field; we know –from Salomon Asch's experiments– that the percentage of conformists in a given population is much higher than the percentage of nonconformists. We also know –at least since Stanley Milgram's (Milgram 1974)

experiments– that the tendency toward obedience to an authoritative-looking figure is very strong amongst humans. If the order giver is credible, if he maintains a close relationship with the order follower, the latter will do practically anything he is ordered to do, without resisting. In this context, we must raise the issue of critical thinking and the conception of alternative worlds. Expecting everyone to be non-conformist, critical and informed will imply confidence in a highly unlikely social change, with consequences very difficult to predict.

On the other hand, in the domain of individual morality, one of the structuring requirements to be able to affirm that a certain act is moral consists in the possibility of the same not being enacted. Duty is not about a constraining obligation. Even knowing what good is, as Saint Paul acknowledged, we can do evil: it is in this tension that the dignity of all acts is founded. To the extent that, even in Christian theology, the problem of free-will finds an answer compatible with the question of evil. That is, God allows it in the name of a greater good, which is freedom. If we were left with only one possible option, there would be no dignity in choosing it. In the realm of emotions as well, the imagination of alternative scenarios occupies a prominent place. Consider the situation of Camus's character in *The Stranger*: If it had not been so hot, if there had not been the resulting despair, would he have killed the Arab? Would he still have subjected himself to an unnecessary death sentence? Most likely not.

This game between what is and what could have been, evidence of a higher cognitive function, underpins every speculation about possible worlds, and allows us to anticipate response scenarios. Now, this possibility of pre-adaptation, outcome evaluation, and speculation about strategic revisions, is at the heart of counterfactual hypothetical reasoning. How can a scientific approach to this issue help us better understand such a role, and how does it speak to the issue of morality?

3. On Counterfactual Reasoning

Counterfactual Thinking (CT) is a human cognitive ability studied in a wide variety of domains, namely Psychology, Causality, Justice, Morality, Political History, Literature, Philosophy, Logic, and AI. In particular, within AI, there is an ongoing effort in the development of algorithmic solutions capable of identifying counterfactual explanations to the decisions produced by automated systems (Chou et al. 2022). CT captures the process of reasoning about a past event that did not occur, namely, what would have happened had the event occurred, which may take into account what we know today. CT is also used to reason about an event that did occur, concerning what would have followed if it had not; or if another event might have happened in its place.

An example situation: Lightning hits a forest, and a devastating forest fire breaks out. The forest was dry after a long hot summer and many acres were destroyed. A counterfactual thought is: If only there had not been lightning, then the forest fire would not have occurred.

Today there is a rediscovery and appreciation of the role of counterfactuals in the fields of Literature, History research, Cognitive Psychology, Moral Psychology and AI, just to name a few of the more relevant areas.

Specifically, in this example, counterfactual reasoning consists in the imagining of an alternative scenario in relation to the one that indeed happened, and the exploration of its consequences: "If the forest floor had not been covered with dry leaves after the long hot summer, then the lightning would not have caused such a tremendous fire."

Applied to the morality of groups, its relevance is as much related to the construction of alternative hypothetical and credible scenarios about the past as to the choices made or about the events that occurred and, concomitantly, the assessment of the various consequences that would have followed. Properly conducted, counterfactual reasonings can provide very relevant insights into the ways ahead in the domains where they are applied. Thus, they are an excellent tool for understanding and explaining the mutability of certain behaviours, supported by the review of strategies, re-examining the past in the light of what we *a posteriori* know today. We can identify some of the reasons that make individuals build counterfactuals: The need to improve future performance, or to work over a factual event to make it more acceptable to themselves, or justifiable to others, either why we did not pursue the alternatives, or by teaching us from experience about what we could rather have done differently to what we did. This way of reasoning may apply as well to events that did not happen but could have happened.

For example, to conjecture what the urban areas of the United States would look like if, instead of building the great railroads, investment had bet even more on rivers as a means of communication. Or about events that occurred, thereby reasoning about what would follow had they not occurred; for example, imagining that the Portuguese Revolution of April 25th, 1974, had not happened, and what the evolution of its prior so-called “Marcellist Spring” would have been. Or if a particular event had not occurred, but another would have in its place, for example, if massive exploitation of fossil fuels had not taken place, and if we had already then moved on to solar and wind energy exploitation. And even to verify if the alternatives would be indifferent with respect to relevant consequences.

In a sense, we can consider that all scientific laboratories are places of counterfactuality, because they create alternative scenarios, which are simplifiers of reality, where a given variable can be tested. To wit, reality is too rich and complex to serve as an appropriate place for certain scientific tests. If we want to know if "x" is the cause of "y" we will have to create a counterfactual scenario where this can be made evident. The fact is, we may be foreseeing the occurrence of "y" in a temporal sequence where "x" has already happened, and this happens successively because "x" is associated with "z" and it is "z" that actually causes "y" and also "x". Finding this out by observing reality may be utterly impossible –the number of items in co-presence is too high and may lead to unnecessary misconceptions and unfounded convictions. Thus, in the laboratory, having a good conjecture and testing one variable at a time enables us to observe unsuspected and unambiguous causal networks. When Galileo conjectured that –in a void– all objects fall at the same speed, gaining

equal speeds at equal times, regardless of their mass, he had no technical means to test the theory. It was from his mental experience that he devised a system of highly polished conduits through which spheres with different masses rolled. Conduit polishing and ball perfection could minimize the inexistence of a vacuum chamber at the time, inasmuch friction was made minimal. Galileo thus constructed the possible scenario in his days to test a theory that very few would be willing to accept. Albeit, the perfect vacuum, as today we know, is impossible, for it is necessarily composed of vacuum fluctuations, without which Heisenberg's Uncertainty Principle would be violated.

4. Counterfactual Reasoning and Conflicts of Interest in Large Populations

Specifically, about applications of counterfactual reasoning in the domain of AI, a scientific approach to the question of morality and judgment can be treated by its consideration as one case of computer implemented game theoretical models. Game theory is nowadays the common language to encode any conflict of interest, with applications spanning from theology to economics, encompassing computer science, mathematics, physics, anthropology, psychology, and many other disciplines. Games are also recognized as one of the key testbeds underlying progress in artificial intelligence (AI), aptly referred to as the "Drosophila of AI" (McCarthy 1997).

Generally, game theory studies how, in a strategic relationship, rationally acting players promote the best outcome for themselves. To do this, each player must analyse the game, and identify the strategies available to achieve its goal. Typically, classical game theory approaches disregard the large-scale dynamical processes that accrue to many social scenarios and modern economic and political systems. Instead, here we will focus on analysing counterfactual reasoning occurring in large populations, adopting a dynamic variant of game theory called Evolutionary Game Theory (EGT).

EGT considers a population of players interacting via a game, a metaphor of a conflict. The payoffs obtained from a given set of interactions are added up and associated with social success or individual fitness. In a natural setting, we may say that strategies that do well reproduce faster. In a social system, successful strategies tend to be imitated more often and thus will spread in the populations. This translates into a convenient (formal and dynamical) similarity between social learning and Darwinian evolution. In the context of human systems, EGT allows the discovery of the most likely behavioural patterns to be found in human populations, together with the mechanisms that will enable one to reach those states. It also allows for novel quantitative descriptions of the dynamics of peer influence, including bounded rationality and cognitive biases pertaining to most social processes.

Here we shall illustrate these ideas in the context of simple conflicts of interest, described by non-cooperative games. The questions related to whether to collaborate or not are pertinent in areas as diverse as Evolutionary Psychology, Evolutionary Biology, Economics, or the Law, among others. Thus, it is important to know whether or not counterfactual reasoning is an essential tool for understanding

behavioural dynamics, and for improving individual as well as collective gains in contexts where the greatest advantage is afforded by evolved collaboration (Santos et al. 2012, 2018).

Given its broad spectrum and cognitive value, a relevant scientific question, and auspicious in terms of research, is what is the effective, if sufficient, role of a small minority of individuals endowed with this counterfactual rationality within some given population. More specifically, to understand if this minority –say ten per cent of the individuals– can influence the whole group, encouraging cooperative behaviours by virtue of their ability to think counterfactually regarding a common good. It is of paramount relevance to determine if counterfactual reasoning, even when adopted by a minority, can influence the collective behavioural patterns.

Importantly, this minority can represent a different set of individuals eager to adopt more detailed reasoning when compared with individuals that simply learn from others. This minority may also be seen as artificial agents or algorithms, mimicking the present challenge of understanding the hybrid world we will soon face, comprising humans and machines (Paiva et al. 2018; Santos et al. 2019). Indeed, besides aiming to understand human decisions better, AI research will continue to investigate how we may foster prosocial behaviours in situations in which cooperation either remains absent or has the potential not to emerge. This may be achieved in different yet subtle ways by transforming the properties of the dilemma humans face, as illustrated below.

We also allude to the extremely complex problem that has arisen from morals suspended on a religious or philosophical system. To avoid the resulting problems, a scientific approach will select aspects that are fundamental to group morality, assignable to all contexts, regardless of the original culture of each group, or the fact that the autonomous agent be biological, or silicon based. It will address in the abstract the elements –say, atomic ones– of all moral systems, such as: collaborating or not collaborating, acknowledging guilt and apologizing, acknowledging or expressing intentions, etc.; and the way in which these aspects may or may not, individually or intertwined with one another, foster group cohesion.

5. Stag Hunting and Law: From Plea Bargaining to International Agreements and AI Regulation

Equipped with the two abovementioned forewarnings, let us delve into our approach to the role of counterfactuals (Pereira and Santos 2019). In the well-known case of the game *Stag Hunt*, a cooperation dilemma is contemplated, which helps us establish the importance of building counterfactuals. It is a game played by any two agents in a population, and the mission of those involved is to hunt stag, a task that must be performed together to maximise the possibility of success and with large payoff. As such, we may also see it as a metaphor of a coordination problem. Each player may decide not to collaborate and choose instead to try and hunt hare on their own. Although it is a less rewarding alternative, the decision can be interpreted as safer, since the hunter depends only on himself, and hare is easier to hunt than stag.

The dilemma results from each hunter not knowing what the other will do; that is, whether he will collaborate and hunt stag, or will act on his own, deciding to defect and hunt hare. So, each one can be tempted to protect himself by hunting hare. In other words, the most cooperative scenario (both players opting for stag) is not achieved due to fear that the other will not follow the same path. The returns differ according to each option taken. One may, for instance, consider a reward R of 4 units for the decision to hunt stag, if taken simultaneously by both players; a return of 3 units for the decision to hunt hare alone; and 0 units for the player who decides to hunt stag without the other doing so. We are thus facing a cooperation dilemma in which maximization of the outcome depends on the effective decision on cooperating by both players. In the context of EGT, players review their strategies, watching each other's actions and copying the most successful ones.

In the domain of the Law, examples of such coordination dilemmas abound. The strategy known as Plea Bargain (PB) could substantially improve its results if informed by the abstract conclusions of the *Stag Hunt* game. Imagine a situation of double whistleblowing, in which each of two culprits –in a payoff context like that of the *Stag Hunt* players– confesses to the wider guilt of both, thereby obtaining an advantageously increased PB, advantageous for the Law's side as well, then our resulting conclusions validate a substantial improvement in the current view and use of the PB, including an improved governance of the Law. Double whistleblowing is not now put forth as more individually rewardable, since whatever it validates is validated by one of the whistleblowers alone, not adding value to the proof. This may make sense in the context of criminal proceedings blame assignment; however, double whistleblowing may afford the Law a wider and confirmatory testimonial evidence. Additionally, analysed from the point of view of the morality of groups, this stance about PB can be seen as promoting multiple PBs. It not only fosters the acknowledgment of guilt in the population from which those indicted for crime come from, something we know is desirable (Pereira et al. 2017), but can also be relevant for the putting together of stronger forensic evidence. A research field is thus opened for legal philosophers interested in evolutionary morality and judgmental topics using the tools of EGT.

From a more general perspective, Stag-Hunt games constitute also the prototypical example of a social contract, a collective agreement between the ruled and their rulers, defining the duties and rights of each. In this realm, one can find instances of Stag-Hunt games in the writings of Rousseau, Hobbes, and Hume (Skyrms 1996, 2004). Smith and Szathmary (1997) have also discussed analogues of social contracts implicit in various natural settings, which can be understood through the lens of adaptive dynamics, cultural evolution, and social learning (Skyrms 2014).

To include the group dynamics associated with this type of problems, the Stag-Hunt can be readily generalisable to an N-player situation where a minimum number of cooperators is required to hunt stag (Pacheco et al. 2009). Imposing such a threshold mimics situations common to most of the public endeavours, where a minimum combined effort is needed to achieve a collective goal. This is also the case in international agreements, which often demand a minimum number of ratifications to come into practice. Adoption of new laws, both at national or international

levels, such as the ones related to climate action and regulation, offer key examples of collective endeavours which can be framed as a N-player Stag-Hunt of coordination games. Antibiotic abuse, vaccination hesitancy, and even coordinating the population to comply with SARS-CoV-2 regulations, provide further examples of this class of dilemmas. In all cases, the non-linear nature of the returns associated with these complex adaptive systems (e.g., as in the case of public health measures), naturally leads to such thresholds and critical levels of adoption to produce a measurable impact (Santos and Pacheco 2011). Climate and public health "games" do have additional complexities due to the time-delayed and uncertain nature of the returns (Santos and Pacheco 2011; Domingos et al. 2020), a complexity which we shall not elaborate on here.

Another dilemma of this class naturally emerges from the ongoing discussions on AI regulation. Rapid technological advancements in AI, as well as the growing deployment of intelligent technologies in new application domains, have generated anxiety and a fear of missing out among different stakeholders, fostering a racing narrative (Han et al. 2020). Whether real or not, the belief in such a race for domain supremacy through AI can make it real, simply from its consequences. These consequences may be negative, as racing for technological supremacy creates a complex ecology of choices that could push stakeholders to underestimate or even ignore ethical and safety procedures. Consequently, different actors are urged to consider both the normative and social impact of these technological advancements, contemplating the use of the precautionary principle in AI innovation and research. This, however, creates novel regulation dilemmas, where non-linearities and thresholds as the ones described above would undoubtedly play an important role. Agreeing or not with implementing these measures involves yet another N-player coordination game, coupled with the innovation dynamics associated with AI systems. Game theoretical models can also be used in this context. In (Han et al. 2020, 2021), we show how these regulatory measures may provide solutions for particular scenarios, depending on the development timeframe of an AI product and the risk of negative externalities. Yet, they may also overshoot their targets, thereby stifling innovation, and hindering investments in developing novel innovations as they become too risky an endeavour.

Now, irrespectively of the conflict or example we are interested in, if we wish to have machines endowed with moral capacity, capable of selecting moral decisions that optimise the expected results and, at least, maximise the expected utility (using here the utilitarian paradigm, with due reservations), it is crucial that we learn to program them with the capacity to develop counterfactual scenarios. These prove to be excellent tools for selecting alternatives not available in the behavioural portfolio for just mimicking and may result in improved cohesion and cooperativeness within groups. This statement has significant experimental relevance; according to a study conducted by the UK Department of Justice (2013), in the context of rehabilitation of delinquents condemned in court cases, recidivism cases are strongly mitigated by strategies that involve the use of counterfactual reasoning. In fact, in mentoring activities that aim delinquents to make other life still alternatives, it is

proven that those who process stimuli to the point of desiring other existential alternatives are the ones who least relapse into criminality.

In all these examples, counterfactual reasoning is also usable for judging, morally, the intentions of an agent's act. One counterfactually assumes that a certain noxious side effect that occurred might not have occurred. Even so, would the purpose of the acting agent have been accomplished? If not, then this side effect was indispensable and, therefore might have been intentional. If so, then it was not necessary to achieve the agent's goal, and therefore, the noxious effect did not need to be intended (Pereira and Saptawijaya 2017).

6. Evolutionary Games with Counterfactual Thinking (CT)

In this section, we illustrate how application of counterfactual thinking to the *Stag Hunt* – contrary to what happens with the mimetic process proposed by social learning theory –the individual can conjecture what would happen if he had used another strategy as his own (such as collaborating) rather than the one he in fact used (such as defecting). We depart from the usual computer-modelling of artificial agents to illustrate that counterfactual reasoning is much more efficient and fruitful in revising strategies than simply mimicking of the most successful strategies used by the adversary. Note that the game also shows that the creation of counterfactuals is a merely instrumental mental activity solely dependent on oneself. That is, it is also a resource available to those who systematically opt for selfish strategies. There exists counter-factuality for the good, and for the evil… (say, the Mafia).

Given the wide cognitive empowerment of CT in the human individual, the question arises of how the presence of individuals with CT-enabled strategies affects the evolution of cooperation in a population comprising individuals of diverse interaction strategies. Importantly, depending on the game and associated strategies, individuals may revise their strategies in different ways. The common assumption of classic game theory is that players are rational, and that the Nash Equilibrium constitutes a reasonable prediction of what self-regarding rational agents adopt (Fudenberg and Tirole 1991). Often, however, players have limited cognitive skills or resort to simpler heuristics to revise their choices. Evolutionary game theory (EGT) (Hofbauer and Sigmund 1998) offers an answer to this situation, adopting a population description of game interactions in which individuals resort to social learning and imitation. As a result, strategies that do well spread in the population.

Yet, contrary to social learning, more sophisticated agents (such as humans) might instead imagine how a better outcome could have turned out, if they would have decided differently, and thence self-learn by revising their strategy. This is where Counterfactual Thinking (CT) comes in. Here, we have previously proposed a mathematical model to study the impact on cooperation of having a population of agents resorting to such counterfactual kind of reasoning, when compared with a population of just social learners (Pereira and Santos 2019). Specifically, we answered for the positive to three main questions:

1. Can we formalize counterfactual behavioural revision in large populations (taking cooperation dynamics as an application case study)?
2. Will cooperation emerge in collective dilemmas if, instead of evolutionary dynamics and social learning, individuals revise their choices through counterfactual thinking?
3. What is the impact on the overall levels of cooperation of having a fraction of counterfactual thinkers in a population of social learners? Does cooperation benefit from such diversity in learning methods?

CT can be exercised after knowing one's resulting payoff following a single playing step with a co-player. It employs the counterfactual thought: *Had I played differently, would I have obtained a better payoff than I did?* This information can be easily obtained by consulting the game's payoff matrix, assuming the co-player would have made the same play, that is, other things being equal. In the positive case, the CT player will learn to next adopt the alternative play strategy. In EGT, a frequent standard form of learning is so-called Social Learning (SL). It basically consists in switching one's strategy by imitating the strategy of a more successful individual in the population, compared to one's success. CT, instead, can be envisaged as a form of strategy update learning akin to debugging, in the sense that: *if my actual play move was not conducive to a good, accumulated payoff, then, after having known the co-player's move, I can imagine how I would have done better had I made a different strategy choice*.

When compared with SL, this type of reasoning is likely to have a minor impact in games of cooperation with a single Nash equilibrium (or a single evolutionary stable strategy, in the context of EGT) such as the Prisoner's Dilemma or the Public Goods game, where defection-dominance prevails. However, as illustrated below, counterfactual thinking has the potential to have a strong impact in games of coordination, characterized by multiple Nash Equilibria: CT will allow for a meta-reasoning on which equilibria provide higher returns.

Let us consider a population of size Z in which individuals engage in a N-Stag-Hunt dilemma (see above) characterized by a limited set of behaviours: *to cooperate* or *to defect*. The cooperators (Cs) contribute a cost c to the public good, whereas defectors (Ds) refuse to do so. The accumulated contribution is multiplied by an enhancement factor F, and the ensuing result equally distributed among all individuals of the group, irrespective of whether they contributed or not. The requirement of coordination is introduced by noticing that often we find situations where a minimum number M of Cs is required within a group to create any sort of collective benefit.

What is the impact on the overall levels of cooperation of having a fraction of counterfactual thinkers in a population of social learners? Does cooperation benefit from such diversity in learning methods? To answer these questions, we developed a new population dynamics model based on evolutionary games, which allows for a direct comparison between the behavioural dynamics created by individuals who

revise their behaviours through social learning and through counterfactual thinking (Pereira and Santos 2019).

In Figure 1A, we illustrate the behavioural dynamics both under CT and SL for the same parameters of the N-person Stag-Hunt game. For each fraction of co-operators (Cs), if the gradient G (for both SL or CT) is positive (negative), then it is likely the fraction of Cs will increase (decrease). As shown, in both cases, the dynamics is characterized by two basins of attraction and two interior fixed points: one unstable (also known as a coordination point), and a stable co-existence state between Cs and Ds. To achieve stable levels of cooperation (in a co-existence state), individuals must coordinate to be able to reach the cooperative basin of attraction on the right-hand side of the plot, a common feature in many non-linear public goods dilemmas (Pacheco et al. 2009). Figure 1 also shows that CT allows for the creation of new playing strategies, absent before in the population, since new strategies can appear spontaneously based on individual reasoning. By doing so, CT interestingly leads to different results if compared to SL. In this particular scenario, it is evident how CT may facilitate coordination of action, as individuals can reason on the sub-optimal outcome associated with non-reaching the coordination threshold, and individually react to that.

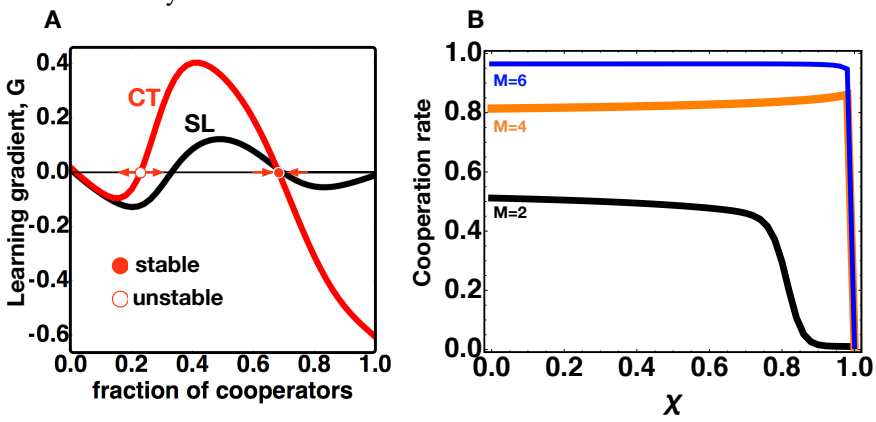

Figure 1. A) Left panel: Learning gradients for social learners (SL, black line) and counterfactual learners (CT, red line) for the N-person SH game. If the learning gradient is positive (negative), the fraction of cooperators will tend to increase (decrease). Empty and full circles represent the finite population analogue of unstable and stable fixed points, respectively. Right panel: Stationary distribution of the Markov processes created by the transition probabilities pictured in the left panel; it characterizes the prevalence in time of each fraction of cooperators in finite populations. B) Right panel: Overall cooperation as a function of the prevalence of individuals resorting to social learning (SL, χ) and counterfactual reasoning (CT, 1-χ). It shows that only a relatively small prevalence of counterfactual thinking is required to nudge cooperation in an entire population of self-regarding agents. Other parameters: Z=50, N=6, F=5.5. M=N/2 (panel A), c=1.0, μ=0.01, β_{SL}=β_{CT}=5.0.

In Figure 1A, individuals can either revise their strategies through social learning or counterfactual reasoning. However, one could also envisage situations where

each agent may resort to CT and to SL in different circumstances, a situation prone to occur in Human populations. To encompass such heterogeneity at the level of agents, let us consider a simple model in which agents resort to SL with a probability χ, and to CT with a probability $(1-\chi)$.

In Figure 1B, we show the impact χ on the average cooperation levels in a N-person Stag-Hunt dilemma in which, in the absence of CT, cooperation is unlikely to persist. Remarkably, our results suggest that a tiny prevalence of individuals resorting to CT is enough to nudge an entire population of social learners towards highly cooperative standards, providing further indications on the robustness of cooperation prompted by counterfactual reasoning. This result becomes more evident whenever coordination is harder to achieve (i.e., larger coordination thresholds, M).

This result may have various interesting implications, if heterogeneous populations are considered. For instance, we can envision a near future made of hybrid societies comprising humans and machines. In such scenarios, it is not only important to understand how human behaviour changes in the presence of artificial entities, but also to understand which properties should be included in artificial agents capable of leveraging cooperation among humans. Our results suggest that a small fraction of artificial CT agents in a population of Humans social learners can decisively influence the dynamics of cooperation towards a cooperative state.

7. Concluding Remarks

We have argued that counterfactual reasoning or thinking is a cognitive device with a long human history, which supplies a basis for causal explanations, and hence for the attribution of blame in moral and judicial judgments. We illustrate the potential impact of counterfactual reasoning in the context of non-linear public goods dilemmas, also known as N-player Stag-Hunt game, a class of dilemmas of relevance in a broad range of domains, from law and public health to international agreements and AI regulation. Our results suggest that counterfactual learners foster coordination in collective dilemmas of this kind, transforming the behavioural dynamics typically associated with these games (Pereira and Santos 2019).

We also showed how these counterfactual learners may influence others. Particularly, in an era increasingly shaped by intelligent systems and artifacts that amplify the human ability to manipulate information, it urges to understand how such instruments can change human behaviour and augment our capacity to cooperate (Paiva et al. 2018; Santos et al. 2019). In this realm, our results suggest that a small fraction of artificial agents resorting to CT is able steer human cooperation whenever placed in hybrid populations comprising humans and machines. A similar effect has been shown to be present in the context of other dilemmas (Santos et al. 2019).

Obviously, real decision-making processes among humans involve a complexity beyond the limits we use to illustrate these ideas. On the other hand, the conceptual simplicity of these models makes them generally applicable to a broad range of problems involving collective cooperative action, which emerges in numerous

conflicting situations in nature and societies, thereby providing insights into the richness, beauty, variety, and complexity of collective social interactions.

Finally, our previous work on machine ethics (Pereira and Lopes 2020a, 2020b) enticed us to consider and argue for the positive effect on Law governance and its application regarding the advantage of promoting joint Plea Bargaining, based on its situational analogy with the Stag-Hunt evolutionary game and the latter's results. Moreover, our previous work on guilt (Pereira et al. 2017), points to the advantage of training detainees in counterfactual thinking about their acts and alternative options, with a view to honing their moral sense, speeding their conditional parole, and improving their future behaviour.

Indeed, there is a compelling intuition that the anticipation of regret (over undesired outcomes) is a significant factor in decision making. Most generally, regret theories imply that the attractiveness of an option cannot be evaluated without reference to the context of other available options. Because regret is a response to the counterfactual outcome of a different choice, the knowledge that the decision maker expects to have about that outcome should affect the anticipation of regret. The knowledge of the payoff matrix of a game permits the evaluation of possible alternative payoffs (Kahneman 1995).

Acknowledgements

L.M.P. acknowledges support by Future of Life Institute grant 372 RFP2-154 and is supported by NOVA LINCS (UIDB/04516/2020) with the financial support of FCT-Fundação para a Ciência e a Tecnologia, Portugal, through national funds. F.C.S. acknowledges support from FCT Portugal's grants PTDC/CCI-INF/7366/2020, PTDC/MAT-APL/6804/2020, and UIDB/50021/2020. A.B.L. acknowledges the support of ANQEP – "Agência Nacional para a Qualificação e Ensino Profissional."

References

Byrne RMJ (2005) The rational imagination: how people create alternatives to reality. MIT Press, Cambridge, MA

Chou YL, Moreira C, Bruza P, Ouyang C, Jorge J (2022) Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications. Inf Fusion 81:59–83

Cimpeanu T, Santos FC, Pereira LM, Lenaerts T, Han TA (2022) Artificial intelligence development races in heterogeneous settings. Sci Rep 12:1723

Dietz Saldanha EA, Hölldobler S, Pereira LM (2015) On conditionals. In: Gottlob G, Sutcliffe G, Voronkov A (eds) Global conference on artificial intelligence. EPiC Computer Science, Tbilisi, Georgia, p 79–92

Dietz Saldanha EA, Hölldobler S, Pereira LM (2021) Our themes on abduction in human reasoning: a synopsis. In: Shook JR, Paavola S (eds) Abduction in cognition and action: logical reasoning, scientific inquiry, and social practice. Springer, Cham, Switzerland, p 279–293

Domingos EF, Grujić J, Burguillo JC, Kirchsteiger G, Santos FC, Lenaerts T (2020) Timing uncertainty in collective risk dilemmas encourages group reciprocation and polarization. iScience 23:101752

Fudenberg D, Tirole J (1991) Game theory. MIT Press, Cambridge, MA

Gigerenzer G, Engel C (2006) Heuristics and the law. MIT Press, Cambridge, MA

Greiner D (2008) Causal inferences in civil rights litigations. Harv Law Rev 81:533–598

Han TA, Lenaerts T, Santos FC, Pereira LM (2022) Voluntary safety commitments provide an escape from over-regulation in AI development. Technol Soc 68:101843

Han TA, Pereira LM, Lenaerts T, Santos FC (2021) Mediating artificial intelligence developments through negative and positive incentives. PLoS One 16:e0244592

Han TA, Pereira LM, Santos FC, Lenaerts T (2020) To regulate or not: a social dynamics analysis of an idealised ai race. J Artif Intell Res 69:881–921

Hofbauer J, Sigmund K (1998) Evolutionary games and population dynamics. Cambridge University Press, Cambridge, UK

Kahneman D (1995) Varieties of counterfactual thinking. In: Roese N, Olson J (eds) What might have been: the social psychology of counterfactual thinking. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, p 375–396

Mandel R, Hilton D, Catellani P (2005) The psychology of counterfactual thinking. Routledge, Milton Park, UK

McCarthy J (1997) AI as sport. Science 276:1518–1519

Milgram S (1974) Obedience to authority — An experimental view. Harpercollins, New York, NY

Moore M (2009) Causation and responsibility — An essay in law, morals, and metaphysics. Oxford University Press, Oxford, UK

Pacheco JM, Santos FC, Souza MO, Skyrms B (2009) Evolutionary dynamics of collective action in N-person stag hunt dilemmas. Proc R Soc B Biol Sci 276:315–321

Paiva A, Santos FP, Santos FC (2018) Engineering pro-sociality with autonomous agents. In: Proceedings of the thirty-second AAAI conference on artificial intelligence and thirtieth innovative applications of artificial intelligence conference and eighth AAAI symposium on educational advances in artificial intelligence, AAAI Press, New Orleans, LA, p Article 994

Pearl J (2010) Causality – Models, reasoning, and inference. Cambridge University Press, Cambridge, UK

Pearl J, Mackenzie D (2018) The book of why – The new science of cause and effect. Basic Books, New York, NY

Pereira L, Lopes A (2020b) Máquinas éticas - da moral da máquina à maquinaria moral. NOVA.FCT Editorial, Costa da Caparica, Portugal
Pereira L, Saptawijaya A (2016a) Programming machine ethics. Springer, Berlin, Germany
Pereira LM, Lenaerts T, Martinez-Vaquero LA, Han TA (2017) Social manifestation of guilt leads to stable cooperation in multi-agent systems. In: Proceedings of the 16th conference on autonomous agents and multiagent systems, International Foundation for Autonomous Agents and Multiagent Systems, São Paulo, Brazil, p 1422–1430
Pereira LM, Lopes AB (2020a) Machine ethics: from machine morals to the machinery of morality. Springer, Cham, Switzerland
Pereira LM, Santos FC (2019) Counterfactual thinking in cooperation dynamics. In: Fontaine M, Nepomuceno-Fernández Á, Magnani L, Salguero-Lamillar FJ, Barés-Gómez C (eds) Model-based reasoning in science and technology. Springer, Cham, Switzerland, p 69–82
Pereira LM, Saptawijaya A (2016b) Counterfactuals in critical thinking with application to morality. In: Magnani L, Casadio C (eds) Model-based reasoning in science and technology: logical, epistemological, and cognitive issues. Springer, Cham, Switzerland, p 279–289
Pereira LM, Saptawijaya A (2017) Counterfactuals, logic programming and agent morality. In: Urbaniak R, Payette G (eds) Applications of formal philosophy: the road less travelled. Springer, Cham, Switzerland, p 25–53
Roese N, Olson J (1995) What might have been: the social psychology of counterfactual thinking. Lawrence Erlbaum Associates Inc, New Jersey, NJ
Santos FC, Pacheco JM (2011) Risk of collective failure provides an escape from the tragedy of the commons. Proc Natl Acad Sci U S A 108:10421–10425
Santos FC, Pinheiro FL, Lenaerts T, Pacheco JM (2012) The role of diversity in the evolution of cooperation. J Theor Biol 299:88–96
Santos FP, Pacheco JM, Paiva A, Santos FC (2019) Evolution of collective fairness in hybrid populations of humans and agents. In: Proceedings of the thirty-third AAAI conference on artificial intelligence and thirty-first innovative applications of artificial intelligence conference and ninth AAAI symposium on educational advances in artificial intelligence, AAAI Press, Honolulu, HI, p Article 754
Santos FP, Santos FC, Pacheco JM (2018) Social norm complexity and past reputations in the evolution of cooperation. Nature 555:242–245
Skyrms B (1996) Evolution of the social contract. Cambridge University Press, Cambridge, MA
Skyrms B (2004) The stag hunt and the evolution of social structure. Cambridge University Press, Cambridge, MA
Skyrms B (2014) Social dynamics. Oxford University Press, Cambridge, UK
Smith JM, Szathmary E (1997) The major transitions in evolution. Oxford University Press, Oxford, UK
Tordoff R (2014) Counterfactual history and thucydides. In: Wohl V (ed) Probabilities, hypotheticals and counterfactuals in ancient Greek thought. Cambridge University Press, Cambridge, UK, p 101–121
UK Department of Justice (2013) Transforming rehabilitation: a summary of evidence on reducing reoffending. ISBN 978-1-84099-608-1, Ministry of Justice Analytical Series, UK.https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/243718/evidence-reduce-reoffending.pdf/ Accessed 19 April 2022
Wohl V (2014) Probabilities, hypotheticals, and counterfactuals in ancient Greek thought. Cambridge University Press, Cambridge, UK