KLEROS

INFELICITOUS COORDINATION: THE SIGNIFICANCE OF KNOBE AND SIDE-EFFECT EFFECTS FOR KLEROS ARBITRATION

Paul A. Poenicke
Kleros Fellowship of Justice, 2023

Infelicitous Coordination: The Significance of Knobe and Side-Effect Effects for Kleros Arbitration

Consider the following story, originally presented in Knobe (2003):

> *The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.' The chairman of the board answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.' They started the new program. Sure enough, the environment was harmed. (192)*

To determine what the chairman of the board intended when coordinating policy, overruling the vice-president and focusing company action on profit, one should disregard the program's positive or negative outcome when trying to determine the chairman's intentions. What the chairman intended through the program was to simply make profit; the actual impact on the environment is a side-effect of that intention, with a mens rea ("guilty mind") ascription being separate from that action's outcome. The Knobe effect, the cognitive bias that humans consistently impute additional intentions based upon adverse side effects and fewer intentions based upon beneficial side effects, is a significant, replicated result in experimental philosophy. The Knobe effect is part of a larger set of similar biases, collectively known as the side-effect effect, which is shared across cultures and has influence in predicating epistemic concepts (knowledge), motivation states (desire), and moral notions (blameworthiness, responsibility) (Alfano, 2022b).[1]

The side-effect and Knobe effects are sources of what this paper analyzes as infelicitous coordination—a concept previously unexplored in philosophy of language, epistemology, and social-political philosophy. Infelicity, as used in philosophy of language following Austin (1962, 14-23), refers to an action that is inappropriately executed, either "incorrectly" or "incompletely" (14-15), in the relevant context. Coordination, as laid out by Schelling (1960) and Lewis (1969), is the choice, "arbitrary and self-perpetuating" (Rescoria, 2019), that allows individuals to agree upon how to interact. Coordination can converge on a variety of choices or points, none of which are required and tend to remain unchanged over iterations.

[1] Alfano (2022a) provides a fantastic overview of the expansion of research from the Knobe effect to the side-effect effect, as well as recent research on cross-cultural studies and active confirmation of results from different studies.

It may seem impossible for coordination to be infelicitous, just as long as the parties agree upon how they will interact and achieve a particular outcome. However, if the desired result is achieved in a way that is inappropriate for the coordination, then coordination is infelicitous. Marriage, one of Austin's favorite examples from How to Do Thing with Words, provides a thought experiment for an initial appreciation of infelicitous coordination. For a marriage ceremony to create a legally binding union, two individuals must coordinate so that the law recognizes the new union—in some cases, what is required is simply appearing before a judge; in other cases, participating in a wedding in a church or temple. In our example, the couple prefers a sacred ceremony but believes that the secular ceremony is necessary for them to be married legally. Religious commitments and legal requirements motivate the couple to plan two ceremonies, with the religious ceremony before the secular ceremony. However, the couple is mistaken about the law: their favored religious ceremony ends up being the necessary legal act, and it is their confusion that ultimately results in their legal marriage via the first, preferred wedding.

One of the goals of this paper is to explicate the idea of infelicitous coordination and apply it to the context of Kleros arbitration. Harnessing coordination for arbitration has incredible value, but infelicitous coordination is a very real threat, especially when the infelicitous aspect of interaction—and not the intentions, incentives, or other features of coordination—achieves the desired result (as in the case of the mishandled marriage). Noted philosophers and legal scholars believe Knobe and side-effect effects already encourage infelicitous coordination in law: If defendants are perceived as having greater intentionality, desire, knowledge, blameworthiness, or responsibility related to negative side-effects, this may increase the punishment of individuals for negative effects unrelated to a case (a worry expressed by Kneer, 2017; Nadelhoffer, 2006; Prochownik, 2021).

Of significance for Kleros's court system is that the side-effect effect has been revealed in study participants, who are being asked about the effect, predicating more blameworthiness and moral responsibility for violating non-pertinent and silly legal and non-legal norms (Güver and Kneer, 2022; Alfano, Machery, Plakias, and Loeb, 2022). Norms have a coordinating function, so it is expected that individuals see them as signals for coordination. However, when those norms are unimportant or silly, the norm should not be salient or binding; it should also not be a reason to increase predications of blame and moral responsibility. Unfortunately, this has not been what experimental philosophers and lawyers have discovered—salient and non-salient norms both increased predications of blame and moral responsibility.

As Kleros advances in providing justice based upon coordination, infelicitous coordination becomes more salient. In this paper, I will explain why cognitive biases, specifically the Knobe and side-effect effects, despite incentives to properly coordinate, encourage infelicitous coordination. I will also offer solutions for how Kleros can help stop such

mishandling of justice. Section 1 begins by outlining how Kleros utilizes coordination in arbitration proceedings, explicates the notion of infelicitous coordination, and explains why this form of coordination is dangerous for Kleros's innovative way of pursuing justice. Section 2 describes the Knobe and side-effect effects to demonstrate how cognitive biases can encourage coordination yet still be infelicitous—an inadequate, inappropriate application of coordination-based justice. Section 3 extends the critique to irrelevant norms, which are especially troubling for a legal system that utilizes coordination points to determine justice. Section 4 outlines a study that will test how significant Knobe and side-effect effects are in Kleros's court system.

Significance of Infelicitous Coordination for Kleros

A short, focused discussion of incentives for coordination within Kleros's arbitration system is sufficient to reveal the significant threat of infelicitous coordination. Kleros's system works because of incentivized coordination, and one of the most important incentivizations occurs in disputes. Disputes are voted on by jurors, with coordination being incentivized for jurors to converge on the appropriate result for the dispute in question. If a juror votes with the ultimate ruling, they receive PNK, to cover their stakes from jurors not voting with the ruling, as well as arbitration fees. A second incentive to coordinate is that if a juror votes in favor of the ruling and receives additional PNK, then the PNK can be staked towards being drawn into another jury, with a greater PNK stake increasing the probability of being part of the jury.[2]

The epistemology of coordination provides a second incentive structure that, when explicated, demonstrates why incentivized, felicitous coordination is essential for Kleros's arbitration system. Coordination, particularly outlined by Lewis, forces individuals to know why others ought to coordinate on a particular point; further, coordination requires individuals to know why their partner expects to coordinate at that point, as well as knowing why everyone has roughly the same coordination preferences (Rescorla, 2019). Lewis and other thinkers focus on coordination as a basis for future interaction, while Kleros utilizes coordination to determine the arbitration decision, ideally encouraging jurors to coordinate on truth. Incentivized coordination makes Kleros's juries epistemic engines that encourage honesty, fairness, and, as noted before, truth, akin to how Longino (1990) approaches scientific coordination and objectivity: for a particular individual to flourish in such a system, they must know why a particular coordination

[2] This abridged account is taken from Lesaege, Ast, and George, 2019, and Lesaege, George, and Ast, 2021.

point should be chosen, as well as why others would find such a point to be justifiable and, ultimately, why all coordinators have reason to find that point to be justified.

This movement from first-person knowledge to second-person justification, and, finally, to knowledge of what is justifiable from a variety of positions (a kind of objectivity), makes Kleros's jury system a particularly effective epistemic engine. Jurors who are adept at these three kinds of knowledge—justification to one's self, to one's possible jurors, and to the legal community—will predominate over self-interested individuals, who perhaps became a juror in hopes of rigging a ruling. Similarly, jurors adept at these three kinds of knowledge will tend to surpass jurors who are loath to fully appreciate the values and perspectives of others, as well as anyone uninterested in understanding how various perspectives in communities justify claims and to know how relevant peers think to coordinate. This is how coordination can be truthful and reflective of the facts, even if jurors cannot speak to one another.

If Kleros's approach to coordination functions as described above, then there does not seem to be a problem as long as jurors follow incentives and rulings are perceived as legitimate by all participants. However, infelicitous coordination is consistent with coherent incentives and legitimate rulings. The trouble is that the resulting coordination could be, in various ways, incomplete or incorrect for arbitration, even when coordination is appropriately incentivized to achieve a normative ideal.

Consider the following analogy, inspired by Cohen (2009), to understand infelicitous coordination when it comes to coordination on a normative ideal, justice.[3] A group of individuals agree to go on a camping trip. In addition to the coordination necessary for agreeing upon the camping decision, coordination is necessary to make sure the group is fed, housed, and entertained during their time in the outdoors. Our campers are an ideological bunch: they want to coordinate to realize justice in their group, so that (1) each person has an equal opportunity, removing undeserved benefits or harms, to achieve their goals, and (2) each person cares about one another, as well as caring that others care about one another and the larger camping project. To encourage these goals, incentives are created such that (1) campers receive an equal opportunity to achieve their life plans, (2) economic inequality is limited, and (3) measures to support care for others and the ideals behind the trip remain throughout the process.

Assume that our group realizes their goals not through incentivized coordination but through a means that, consistent with incentives and interaction, provides the same outcome. Perhaps the campers are an extended family: a desire to create memories, a desire for harmony, or love between family members would be the reason behind realizing the campers' goals. Just as in the marriage case, more exotic possibilities are

[3] This case presents infelicitous coordination in a more thorough fashion than the wedding case, which may strike the reader as too potted and philosophical. This case also shows how one can infelicitously coordinate based upon a normative ideal, which is akin to Kleros's goal of truth.

possible: perhaps each family member is being framed by a powerful mob boss to go on a camping trip and realize these altruistic goals, or there is a requirement from extended family for the immediate family to do something together and work out the difference. There are many, more plausible reasons, why individuals might be encouraged to coordinate within an incentive structure to achieve a particular goal. The problem is that in these examples, beyond the fact that such infelicitous coordination could obviously lead to poor outcomes, the reason for coordination is bias, coercion, or psychological direction, making the actual reason for coordination incongruous with the goals of the trip. These are examples of infelicitous coordination: there is incentivized coordination, but the actual motivation behind the coordination is, beyond being consistent with the incentive structure and possibly leading to an unwanted outcome, illegitimate and incongruous with the normative ideals for the context in which the coordination occurs.

The previous analogy demonstrates that it is possible to have coordination that, despite attempts to realize some normative goal through incentivization, may still be illegitimate and incongruous within the context of that interaction. The Knobe and side-effect effect are ways to coordinate—but coordinating around cognitive biases is illegitimate for the goals of justice. The same is true for coordinating around unimportant or trivial norms: these are points for coordination, though their connection to the truth—the goal of justice—is questionable. In the following two sections, I will further explain why these forms of coordination can occur in Kleros's system and are harmful to the context of determining justice.

Infelicities Regarding the Knobe Effect and Side-Effect Effect

This paper opened with an explanation of the Knobe effect as a species of side-effect effects. Knobe began by examining individuals' predication of intentionality by going to a park, presenting individuals with stories that had either positive or negative cues in the vignette that resulted in negative or positive outcomes, and having participants record on a Likert scale about whether they felt that the protagonist—the individual whose intentionality was being measured—acted more or less intentionally. Except for the natural setting, the experimental philosophy continued Knobe's approach, actively confirming Knobe's work and completing large projects that attempted to ascertain whether the results could be confirmed in non-Western countries (Alfano, Machery, Plakias, and Loeb, 2022; Maćkiewicz, Kuś, Paprzycka-Hausman, and Zaręba, 2022; Cova, F., Strickland, B., Abatista, A. G. F., Allard, A., Andow, J., Attie, M., … Zhou, X., 2019).

Troublingly, Knobe's results appeared for other predications. Beebe was one of the first to recognize that a similar effect applied to knowledge (Beebe and Gerken, 2014), and is credited with making popular "side-effect effect," the larger set of cognitive biases that increases a predication based upon negative results. In addition to epistemic concepts, philosophers found that these predications appeared in, and replicated in cross-cultural experiments on, epistemic concepts (knowledge), motivation states (desire), and moral notions (blameworthiness, responsibility). The confirmation and transnational research into Knobe effects is mirrored in other side-effect effect predications, with robust confirmations coming for all of the mentioned predications.

Even early on in the research (Nadelhoffer, 2006), the legal significance of Knobe effects, and eventually side-effect effects, was clear. If individuals tended to credit knowledge, intentionality, desire, blameworthiness, and responsibility based upon results, then defendants who inflicted more negative impacts would be judged differently than individuals whose same actions had a positive result. Within the last decade, the experimental legal field has begun to investigate the issue with more sensitivity towards the legal context (Macleod, 2006). Unfortunately, for the legal system, further investigation revealed that jurors as well as judges fell prey to such a cognitive bias (Kneer, 2017).

Cognitive biases driving Knobe and side-effect effects provide a clear example of infelicitous coordination. With incentives in place, a jury could agree on a judgment, which could be accepted by both parties and eliminate the felt need for an appeal. However, what is driving the judgment is not incentives that encourage honesty—what causes jurors to agree, and for parties to be satisfied with the result, is a cognitive bias that impacts the entire set of individuals involved with the case. Coordination is taking place but there is more than just an intuition that something has gone wrong. The context of determining justice has been impacted by incorrect coordination, the result of a cognitive bias that impacts all individuals in the case (jurors and individuals in the dispute), and is consistent with the carefully managed system of Kleros arbitration.

Significance of Uninformative Norms

A more relevant issue for Kleros is the impact of the side-effect effect on norms. Norms being natural coordination points, it would be troublesome for Kleros's system if the cognitive bias initiated by the side-effect effect was triggered by norms that were uninformative or unimportant to the situation at hand. Cases that had such norms would be infelicitously decided if the natural Schelling Point in a case—what was in fact true—was not perceived by participants as the focal coordinating point. Unfortunately,

results from the literature suggest that the side-effect effect does indeed have such additional significance for Kleros's legal system, as a species of the side-effect effect has been demonstrated (Güver, L., and Kneer, M., 2022; Knobe, J., and Shapiro, S. J., 2021).

An example from Güver and Kneer (2022) demonstrates how norms impact predications of various kinds that are of significance for justice. In testing whether blameworthiness is impacted by side-effect effects, Güver and Kneer (2022) created the following vignette:

> *One recent summer afternoon, Mark was rollerblading outside. The path Mark is on is commonly used by cyclists, rollerbladers, and pedestrians.*
>
> *One of these pedestrians is Lauren, who is walking ahead of Mark.*
>
> *Suddenly a cat jumps onto the path right in front of Lauren. Lauren is startled and steps to the left to evade it.*
>
> *Mark, who is approaching speedily on rollerblades from behind, collides with Lauren. The collision sweeps her off her feet and knocks her to the ground. Lauren sustains bruises all over. (162)*

In other versions of the vignette, the norm becomes either unimportant or nonsensical, with the following sentence occurring after the second sentence in the vignette: "However, it is forbidden to be on the path as a cyclist or rollerblader unless one wears a helmet. Mark is not wearing a helmet. He is thus not allowed to be on the path." or "However, it is forbidden to be on the path as a cyclist or rollerblader unless one wears a gray t-shirt. Mark is not wearing a gray T-shirt. He is wearing a blue t-shirt. He is thus not allowed to be on the path." The problem is that the side-effect effect remained statistically significant when Mark violated the unimportant or nonsensical norm and the result was negative.

The problem with coordination around unimportant or uninformative norms adds further weight to the problem of Knobe and side-effect effects for Kleros arbitration. Arbitration exists within a set of norms, including promising, supplying, and punishment, all of which, ideally, are set up in the contract between individuals. If other norms can impact juries, then it is possible for the norms established in a contract to be only part of the set of norms considered by jurors. Incentives would naturally focus on norms that are part of the contract—but these norms are not the only ones that can be focal points for jurors. Even if jurors follow the correct norms as laid out in contracts, given that one party broke a norm and had a negative consequence (needing to go to arbitration), punishments for norm-breaking parties would be more severe than norm-upholding parties that did the same action.

Responding to the Challenges of Infelicitous Coordination—A Study Proposal

How should Kleros respond to the challenges set out in this paper? A key assumption is that the literature provides identical challenges for Kleros's arbitration system as it does for various philosophical theories. There are reasons to doubt this claim. First, individuals tend to be more careful when there is a financial benefit or harm: putting money on the line is an effective way to encourage people to care and act more responsibly. The incentive structure for jurors discourages dishonest, ignorant jurors, and it may also discourage cognitive biases, including the Knobe and side-effect effects.

More importantly, the arbitration system that exists now can be reinforced to avoid infelicitous coordination. Information and training on side effects and legal concepts would be easy to integrate into the juror incentive program; supplemented by pre-case checks--simple five-minute confirmations that jurors are not prone to Knobe or side-effect--jurors could receive significant training and consideration unavailable in traditional jury situations. Jurors who passed these tests, as well as pre-case checks, could be given badges for competence or be incentivized through increased PNK.

The fact that Kleros is online provides additional safety precautions. Given that the aforementioned side effects may be the result of non-affective, cold System 2 thought (Diaz, 2017), explicit jury instructions (Macleod, 2016) may be useful by closing the interpretative diversity of concepts associated with these side effects. Such additional safety precautions could limit, or even eliminate, the various biases that cause of the Knobe effect: a focal bias (anchoring effect) has been suggested by Beebe and Gerken (2014) as the cause of the Effect, while Ostillio and Bukat (2018) offer the availability heuristic bias as the cause. Finally, the cases themselves could be recomposed and presented in alternative syntax (which helps avoid a Knobe effect (Strickland, Fisher, Knobe (2012)), or with or without explicit moral (or epistemic, motivational, etc.) censure against a possible side-effect, which Lindauer and Southwood (2021) find as a way to avoid a side-effect.

Unfortunately, the Knobe and side-effect effects have yet to be evaluated in a court setting, to the frustration of researchers (including Nadelhoffer, 2006 and Prochownik, 2021). Without an actual court proceeding to evaluate, all we have are studies and theorizing that suggest these effects may have a significant impact on how justice is carried out. A jury study, deliberating on cases reflecting the concerns of Nadelhoffer and

Prochownik, seems impossible to do in a traditional court setting, especially if we add in other desiderata (i.e., having a record of jurors' reactions as they deliberate, or responses of the potential jury pool to the vignettes from the literature).

With a combination of qualitative and quantitative methods, and examining the right court system, there is a more than feasible option of testing the impact of the Knobe and side-effect effect on jurors. Kleros's courts provide a chance to not only test vignettes from Knobe, Nadelhoffer, and others, but also to see how jurors react to those cases on relevant communication channels. The proposed study would utilize a mixed methods approach: it would include jury cases, as well as information from a survey with the vignettes mentioned in this paper,[4] reactions to jury cases on Discord and Kleros-related Telegram channels, and a review of previous cases to see if earlier juries faced the conceptual issues in this paper. The proposed mixed methods study has significant advantages, beginning with evaluating the statistical significance of results from the questionnaire, which asks for 1-7 Likert scores on 20 various side-effect effect vignettes. A significant result, showing Kleros's potential jurors were even relatively immune from the Knobe and side-effect effects, would be of interest to the experimental philosophy and jurisprudence communities. Further, the proposed mixed methods approach allows a way to gauge juror reactions to cases; add information from previous cases; and gain a nuanced understanding of juror views of each case, from juror comments on Discord or Telegram to the frequency of appeals in jury cases.

The foundation of the study would be to craft a group of jury cases for arbitration that utilized a vignette from the relevant literature and adjusted it to fit the functioning of Kleros's court system. Ideally, the set of vignettes studied would include the five major cognitive biases replicated in the literature—misprediction of intention, knowledge, desire, blameworthiness, and responsibility. However, picking two, preferably intention and knowledge, given their significance in the literature and for law, would be a focused use of time and resources. Additionally, blameworthiness will be a part of the norm bias study, discussed in the following pages.

If we consider the vignette of Knobe's greedy chairman that began this paper, the jury would be given information mirroring all the relevant details that, at its core, reflected the original vignette (the original is on the left; revised version on the right):

[4] The survey has already been completed and can be found at: https://docs.google.com/forms/d/1t-Ypv7M5bgUJqJ7FfVGd0sM25PPctN56JrP842WDNTg/edit

The vice-president of a company went to the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.'

The chairman of the board answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's start the new program.'

*They started the new program. Sure enough, the environment was **harmed**.*

Question for the jury: Did the chairman intend the environment to be harmed?

A company contact went to a freelance artist and said, 'I am thinking of revising the contracted art project. Its vivid styling will help us increase profits, but materials for the project will also harm the environment.'

The freelance artist answered, 'I don't care at all about harming the environment. I just want to make as much profit as I can. Let's continue the project.'

*They agreed to continue the project. Sure enough, the environment was **helped**.*

Question for the jury: Did the free-lance artist intend the environment to be helped?

The help case (the case on the right) would be tweaked to present a similar situation for the harm alternative, which would be assigned to another jury. While simplicity is one reason to simply present these vignettes and have the jury decide on the question, there is still an artificial feel to the revised vignettes. Supplemental evidence and further background information, which is expected in other jury cases, may be necessary to have jurors take these cases as seriously as others—though this is a point for further consideration if the study goes forward.

What about the knowledge vignette? The first paper on knowledge and the side-effect effect (Beebe and Buckwalter, 2003) used Knobe's chairman case and asked an epistemic question about the vignette: whether the chairman knew/did not know the environment would be helped or harmed, based on the harm/help versions of Knobe's vignette. In other words, testing the knowledge condition is straightforward: simply using a revised intention vignette, one can then ask the jury to deliberate whether the chairman knew more if the environment was harmed. Obviously, a second jury would consider whether the chairman knew more if the environment was helped, and the example would be tweaked with a different vignette (e.g., a businessman choosing whether to encourage/discourage the use of renewables to support green energy by a construction team).

The second set of jury cases concerns the cognitive biases of unimportant or silly norms. These jury cases will focus on the impact of norm bias on blameworthiness, one of several predications tested by Güver and Kneer (2022). Güver and Kneer set up four studies by slightly changing their base vignette: the no norm condition did not feature

any norm; the important norm condition provided a norm that was salient for action; the unimportant norm condition offered a norm that was not salient for action; and the silly norm condition laid out an implausible norm seemed unjustified in any context. Before reviewing these conditions in detail, it is important to understand how Güver and Kneer's vignettes functioned.

Repeating the Güver and Kneer vignette (from pages 11-12) at this point is helpful to see how to integrate it into the study:

> *One recent summer afternoon, Mark is rollerblading outside. The path Mark is on is commonly used by cyclists, rollerbladers, and pedestrians. (**NORM MENTIONED: Important, Unimportant, Silly**)*
>
> *One of these pedestrians is Lauren, who is walking ahead of Mark.*
>
> *Suddenly a cat jumps onto the path right in front of Lauren. Lauren is startled and steps to the left to evade it.*
>
> *Mark, who is approaching speedily on rollerblades from behind, collides with Lauren. The collision sweeps her off her feet and knocks her to the ground. Lauren sustains bruises all over.*
>
> *Question asked to participants: To what extent do you think that Mark is blameworthy, if at all, for the accident?*

This vignette contains several aspects that must be carefully reflected in the proposed study to apply Güver and Kneer's findings. First, there must be an activity that can be (I) norm-bound in important, unimportant, and silly ways, and (II) something causally connecting the two individuals in the story. Second, the person who breaks the norm (Mark) negatively impacts the other person (Lauren), such that, third and finally, Mark the norm breaker can be considered as blameworthy for the accident. Güver and Kneer's findings reveal that breaking various norms—including unimportant and silly norms—encourages respondents to make Mark more blameworthy for the accident. *This is the case even though all of the relevant features of his action are the same!*

Now we are in a position to present how the study would simulate Güver and Kneer's four vignettes. Let me lay out the basic story before adding in the norms:

> *A noted purveyor of recreations of Chinese art during China's dynastic period (1600BCE-1911CE), Lauren wants to feature more paintings of ordinary Chinese individuals, themes, and images. To that end, she commissions Zhuang Zhou, a noted artist specializing in reproductions of the dynastic period, to produce a painting. (**NORM MENTIONED: Important, Unimportant, Silly**)*

Zhuang's work is good, but he falls far short of his regular excellence. The painting takes weeks to sell, forcing Lauren to retain the painting longer than she wanted.

Because of Zhuang's credentials, Lauren never expected to reduce the price by more than half—a significant loss for a very expensive piece by a noted artist.

Eventually, Lauren sells the painting. Due to the delay in selling and the significantly lower selling price, Lauren is forced to lay off three employees.

Question for the jury: To what extent do you think that Zhuang is blameworthy, if at all, for Lauren laying off her three employees?

The story replaces physical pain with monetary loss, and the causal connection between the two is relevant to a business agreement, rather than interacting on a path. The *no norm condition* is simply the story as presented above. The *important norm* condition would add in the following in the NORM MENTIONED space: "Lauren wants an ordinary Chinese dynastic theme, but she gives Zhuang the freedom to paint what he desires. Zhuang paints a dragon scene, knowing that dragons were only illustrated for emperors. This fact is known by, and in turn, discourages, potential buyers." The *unimportant norm* is as follows: "Lauren commissions a traditional portrait of Confucius. Zhuang executes the portrait in the traditional fashion, which illustrates Confucius as an unattractive, wise sage." The silly norm is: "Dynastic Chinese paintings never used the color red exclusively for landscape painting. Lauren gives Zhuang the freedom to paint a dynastic landscape. Zhuang uses shades of red to reveal a valley at sunset."

This is a representation of what each of the four vignettes for each jury would require. The study's analysis of norms and cognitive bias could be limited to three juries: juries are needed for unimportant and silly norms. Because the no norm and important norm cases acted as controls, choosing either case would work. Just like the chairman case, we would need slightly different versions of the basic vignette to test with each jury. While this would require creativity, the cognitive function of norms, possibly as a bias, is of such significance for Kleros's court system that, in my view, it should be studied.

In conclusion, the study I am proposing would have the following steps:

1. Request individuals associated with Kleros to fill out the questionnaire. This can be ahead of any item on the list and functions as a way to understand the possible members of juries. Using a more powerful survey tool—e.g., Survey Monkey—would allow for easy collection and analysis of the data, which, if significant, could be shared with members of the experimental philosophy community.

2. Review previous jury cases for issues regarding infelicitous coordination. Another

step that can be done before others, and one that I pursued to a significant extent. I reviewed a few hundred cases but did not find anything.

3. Decide which types of cases to submit for juries, determine whether the vignette and jury question are enough, and submit these cases to juries. It would be good to space the cases out to prevent jurors from considering the case's background and, more importantly, discussing with other jurors that they are not ruling on real cases. Once the construction of the vignettes and related issues are resolved, watching and recording how jurors react and respond to the proposed jury cases becomes the primary aspect of the study. During cases, monitoring Discord and Telegram may provide as much information as the jury's deliberation.

Conclusion: Opportunity and Challenge

The challenges presented in this paper face all legal systems (Nadelhoffer, 2006), and being able to see how, through a mixed methods approach, the Knobe and side-effect effect impact a jury would be a significant advancement for the study of law. If Kleros can demonstrate it is not impacted by these biases, unlike judges and juries in state systems, then it would receive greater respect and, hopefully, increased engagement from the public either familiar or unfamiliar with crypto. Furthermore, individuals unassociated with crypto are frequently unfamiliar with the benefits of Kleros's approach to juries as "epistemic engines" that, with the right tweaks to avoid infelicitous coordination, reliably incentivize jurors to coordinate on truth.

The benefits of this study do not stop at revealing Kleros's juries as effective epistemic engines or the improved public status of Kleros as an organization committed to thoroughly reviewing its functioning in the name of truth. Demonstrating how Kleros can be adjusted (e.g., the suggestions on pages 13-14) to avoid cognitive biases and allow jurors to converge on truth, despite potential biases towards non-optimal solutions (i.e., infelicitously coordinating on what is true), would be an effective, non-technical validation of Kleros's judicial system. Finally, a successful demonstration would encourage a virtuous circle: if coordination towards truth continues to reoccur, despite cognitive hurdles to that process, then this would be an ethical result, increasing the system's perceived fairness, with additional coordination resulting in increases in perceived fairness (George, 2018). This would be especially true when the proposed study is the first of its kind: despite philosophers' significant concerns about the side-effect effect's significance for juries (a worry expressed since 2006), there has yet to be a study on the side-effect effect done in an actual legal setting, other than with college students or participants on MTurk. It is time Kleros helped itself by providing the opportunity to test the efficacy of the Knobe and side-effect effect in court.[5] [6]

---

[5] I'd like to thank everyone at Kleros for their patience as I tried to complete this research under personal misfortune. While I think I could've completed the study in July, for all concerned, a shorter paper focused on the conceptual issues and an outline of the study was best. The study can be done if there is still interest. My personal thanks go to Federico Ast for his kind heart and willingness to talk football/soccer (and being an interesting philosopher, natch). William George answered questions that probably were useless yet his patience did not wane. Yann Aouidef generously provided research contacts and helpful comments. Jamilya Kamalova ran excellent review sessions, provided encouragement, and suggested that I focus my paper on the theoretical framework of my research, which was a wise action. Finally, Abeer Sharma provided significant feedback and disabused me of silly ideas early in my research.

[6] If there is still interest in running the study, I would be willing to do so for free. I am genuinely interested in the results, and as I completed this report, I realized that several individuals with funding may be able to help pay for juries. Knobe has his experimental philosophy lab (https://campuspress.yale.edu/joshuaknobe/) and is working to confirm studies across experimental philosophy. My undergraduate institution even has its own funding for experimental philosophy, and I may be able to convince a former professor to help, even though his current research focus is on the philosophy of religion (https://xphi.hillsdale.edu/). I have met Knobe, Beebe, and most of the other major thinkers in experimental philosophy, so I can keep asking different groups if they would like to sponsor this study.

Works Cited

Alfano, M., Machery, E., Plakias, A. and Loeb, D. (2022a). Experimental Moral Philosophy. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/fall2022/entries/experimental-moral

Alfano, M., Machery, E., Plakias, A. and Loeb, D. (2022b). Notes to Experimental Moral Philosophy. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/Archives/fall2022/entries/experimental-moral/notes.html

Austin, J.L. (1962). How to Do Things with Words. Cambridge, MA: Harvard University Press.

Beebe, J. R., & Buckwalter, W. (2010). The epistemic side-effect effect. Mind & Language, 25(4), 474-498.

Beebe, J. and Gerken, M. (2014). "Knowledge in and out of Contrast." Nous, 50 (1): Volume 50 (1): 133-164.

Brogaard, B. (2010). "'Stupid people deserve what they get': The effects of personality Infelicitous Coordination 24
assessment on judgments of intentional action." Behavioral and Brain Sciences, 33(4), 332-334. Doi:10.1017/S0140525X1000169X

Bukat, M. and Ostillio, T. (2019). "The Knobe Effect with Probable Outcomes and Availability Heuristic Triggers." Logos and Episteme, 10 (4): 363-377.

Cohen, G.A. (2009). Why Not Socialism? Princeton University Press.

Cova, F., Strickland, B., Abatista, A. G. F., Allard, A., Andow, J., Attie, M., … Zhou, X. (2018). "Estimating the Reproducibility of Experimental Philosophy." oi.org/10.31234/osf.io/sxdah.

Díaz, R. (2017). "Cold Side-Effect Effect: Affect Does Not Mediate the Influence of Moral Considerations in Intentionality Judgments." Frontiers in Psychology, 8:295.

Fisher, M., Knobe, J., and Strickland, B. (2012). "Moral Structure Falls Out of General Event Structure." Psychological Inquiry, 23 (2):198-205.

George, W. (2018). "Kleros and Mob Justice: Can the Wisdom of the Crowd Go Wrong?"

https://medium.com/kleros/kleros-and-mob-justice-can-the-wisdom-of-the-crowd-go-wrong-ef311209ea36

Guilherme, A., Knobe, J., Struchiner, N. and Hannikainen, I. (forthcoming). "Purposes Infelicitous Coordination 25
in law and in life: An experimental investigation of purpose attribution." Canadian Journal of Law and Jurisprudence.

Güver, L., & Kneer, M. (2022). "Causation and the Silly Norm Effect." In S. Magen & K. Prochownik (Eds.), Advances in Experimental Philosophy of Law (to appear). Bloomsbury Publishing.

Kneer, M. (2017). "Mens rea ascription, expertise and outcome effects: Professional judges surveyed." Cognition, 169: 139-146.

Kneer, M. and Bourgeois-Gironde, S. (2017). "Mens rea ascription, expertise and outcome effects: Professional judges surveyed." Cognition, 169: 139-146.

Knobe, J. (2003). "Intentional action and side effects in ordinary language." Analysis, 63(3): 190–194.

Knobe, J. (2019). "Philosophical Intuitions Are Surprisingly Robust Across Demographic Differences." Epistemology and Philosophy of Science, 56 (2):29-36.

Knobe, J., & Shapiro, S. J. (2021). "Proximate cause explained: An essay in experimental jurisprudence." University of Chicago Law Review, 88, 165–236.

Lesaege, C., Ast, F. and George, W. (2019). Kleros Short Paper v 1.0.7. Infelicitous Coordination 26
https://kleros.io/whitepaper.pdf

Lesaege, C., George, W., and Ast, F. (2021). Kleros Long Paper v2.0.2.
https://kleros.io/yellowpaper.pdf

Lewis, D. (1969). Convention. Cambridge: Harvard University Press.

Lindauer, M. and Southwood, N. (2021). "How to Cancel the Knobe Effect: The Role of Sufficiently Strong Moral Censure." American Philosophical Quarterly, 58 (2): 181–186.

Longino, H. (1990). Science as Social Knowledge. Princeton University Press.

Macleod, J. (2016). "Belief States in Criminal Law." Oklahoma Law Review, 68, 3: 497-554.

Maćkiewicz, B., Kuś, K., Paprzycka-Hausman, K., and Zaręba, M. (2022). "Epistemic Side-Effect Effect: A Meta-Analysis." Episteme (first view): 1-35.

Michael, J. and Szigeti,A. (2019) "'The Group Knobe Effect': evidence that people intuitively attribute agency and responsibility to groups." Philosophical Explorations, 22:1, 44-61. Infelicitous Coordination 27

Nadelhoffer, T. (2006). "Bad acts, blameworthy agents, and intentional actions: Some problems for jury impartiality", Philosophical Explorations, 9: 203–220.

Prochownik, K. M. (2021). "The experimental philosophy of law: New ways, old questions, and how not to get lost." Philosophy Compass, e12791. https://doi.org/10.1111/phc3.12791

Rescorla, M. (2019). Convention. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/sum2019/entries/convention/

Schelling, T. (1960). The Strategy of Conflict. Cambridge: Harvard University Press.

Struchiner, N., Hannikainen, I., and de Almeida, G. (2020). "An experimental guide to vehicles in the park." Judgment and Decision Making, 15 (3):312-329.

Tobia, K. (2021). "Legislative Intent and Acting Intentionally." Advances in Experimental Philosophy of Law, Stefan Magen & Karolina Prochownik, eds. 2022. https://ssrn.com/abstract=3812474