Contestable AI needs Computational Argumentation

Francesco Leofante, **Hamed Ayoobi**, **Adam Dejl**, **Gabriel Freedman**, **Deniz Gorur**, **Junqi Jiang**, **Guilherme Paulino-Passos**, **Antonio Rago**, **Anna Rapberger**, **Fabrizio Russo**, **Xiang Yin**, **Dekai Zhang** and **Francesca Toni**
Computational Logic and Argumentation Group,
Department of Computing, Imperial College London, UK
{f.leofante, h.ayoobi, adam.dejl18, g.freedman22, d.gorur22, junqi.jiang20, gppassos, a.rago, a.rapberger, fabrizio, x.yin20, dekai.zhang19, f.toni}@imperial.ac.uk

Abstract

AI has become pervasive in recent years, but state-of-the-art approaches predominantly neglect the need for AI systems to be *contestable*. Instead, contestability is advocated by AI guidelines (e.g. by the OECD) and regulation of automated decision-making (e.g. GDPR). In this position paper we explore how contestability can be achieved computationally in and for AI. We argue that *contestable AI* requires dynamic (human-machine and/or machine-machine) explainability and decision-making processes, whereby machines can 1. interact with humans and/or other machines to progressively explain their outputs and/or their reasoning as well as assess grounds for contestation provided by these humans and/or other machines, and 2. revise their decision–making processes to redress any issues successfully raised during contestation. Given that much of the current AI landscape is tailored to static AIs, the need to accommodate contestability will require a radical rethinking, that, we argue, computational argumentation is ideally suited to support.

1 Introduction

AI has become pervasive in recent years, with applications ranging from autonomous driving (Muhammad et al. 2020) to finance (Cao 2022) and healthcare (Shaheen 2021). Current state-of-the-art AI systems focus on algorithmic solutions, often built from data, generating outputs in the forms of predictions, recommendations and/or decisions and, in some cases, explanations thereof. These existing solutions, however, mostly neglect the need for these AI systems to be *contestable*. Instead, the need to accommodate contestability is a crucially important problem if AI systems are to be deployed in and benefit society. Indeed, contestability is prominently advocated in some frameworks for AI ethics and in regulations. For example, the Organisation for Economic Co-operation and Development (OECD) states that information should be provided "to enable those adversely affected by an AI system to challenge its outcome based on plain and easy-to-understand information on the factors, and the logic that served as the basis for the prediction, recommendation or decision" (Principle 1.3[1]). Moreover, the ACM Global Technology Policy Council for Responsible Algorithmic Systems advocates 'Contestability and Auditability' as one of nine principles.[2] Contestability is required by law in some jurisdictions, e.g. GDPR, article 22(3),[3] states that the data subject's rights to be safeguarded shall include "at least the right [. . .] to contest the decision". Finally, contestability and redress are also identified as key principles underpinning the UK AI regulation framework[4], whereby "Guidance should clarify existing 'formal' routes of redress offered by regulators in certain scenarios". This is a major departure point from "classical" requirements for explainable systems, which typically do not require correcting the AI when faulty behaviours are exposed.

Contestability is seen by some (Hicks 2022) as a means to facilitate accountability, e.g. to prevent the use of Uber's algorithm for banning drivers because it is not taking into account bias of customers when giving bad reviews. A handful of recent works (overviewed in Section 2) bring contestability to the attention of AI researchers, practitioners and users, while providing insights into its possible interpretations in practice. Moreover, there is evidence, in the form of user studies, that contestability can affect users' perception of AI fairness (Yurrita et al. 2023). However, formal/computational forms of contestability are mostly lacking in the literature, with some exceptions, in the form of 1. indications that, in the case of machine learning, methods generating counterfactual explanations in terms of "actionable changes that an individual can make to flip the prediction of the classifier" may be interpreted as offering (limited) contestability (Venkatasubramanian and Alfano 2020) and 2. (Russo and Toni 2023), offering contestability in the very specific setting of causal discovery. Finally, existing works mostly see contestability as a post-hoc process, after predictions/recommendations/decisions have been computed.

In this position paper (specifically in Section 3) we consider what it means for (any) *algorithmic decision systems (ADS)* (Henin and Métayer 2022) to be contestable and contested (by humans or other ADSs). Our take-away message is that formalising/realising computationally contestable AI requires: *explanations* (by the ADSs which humans/other

[1]https://oecd.ai/en/dashboards/ai-principles/P7

[2]https://www.acm.org/media-center/2022/october/tpc-statement-responsible-algorithmic-systems

[3]https://gdpr-text.com/read/article-22/

[4]https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper

ADSs may want to contest); *grounds* for contestation (by the contesters); ability (by the contested) to *redress* any issues successfully raised during contestation; and ability to *interact* (by both contested and contester). We then advocate computational argumentation (CA) as being ideally suited to support contestability, providing evidence from the literature on CA (Section 4). We finally conclude (Section 5) with some pointers for future directions in what we believe is a very promising direction of research for the KR community.

2 AI Contestability in the Literature

Here, we summarise recent works focusing on advocating or proposing forms of contestability in AI.[5] We focus on the main contestability dimensions that emerge in this literature, as well as providing pointers to other (less) related research.

Contested entities. Prior work has focused on contesting various aspects of an ADS, including its general design (Almada 2019; Alfrink et al. 2022; Yurrita, Balayn, and Gadiraju 2023), training data (Kluttz, Kohli, and Mulligan 2022), training procedure (Almada 2019), its inputs (Alfrink et al. 2022) and outputs (Hirsch et al. 2017; Ploug and Holm 2020; Tubella et al. 2020; Lyons, Velloso, and Miller 2021; Henin and Métayer 2022; Hicks 2022) as well as the ADS as a whole (Lyons, Velloso, and Miller 2021; Vaccaro et al. 2021; Henin and Métayer 2022; Hicks 2022; Alfrink et al. 2023; Russo and Toni 2023). Note that the possibility of humans being contested by an ADS (e.g. if their assumptions about the subject domain are flawed) is not typically considered.

Contesting entities. All approaches we identified focus on humans as the contesting entities. However, there are considerable differences between the categories and assumed skill sets of people engaged in contestation. Several argue that decision subjects should be empowered to directly contest ADS decisions affecting them, potentially without possessing detailed knowledge of the subject domain (Almada 2019; Ploug and Holm 2020; Lyons, Velloso, and Miller 2021; Vaccaro et al. 2021; Henin and Métayer 2022; Hicks 2022). Some works suggest that individual decision subjects may not always be able to effectively contest on their own, and propose contestation via third-party representatives (Alfrink et al. 2022) or as part of a group in a "class action" (Lyons, Velloso, and Miller 2021). Others also consider contestation by professionals, subject matter experts and regulators (Tubella et al. 2020; Henin and Métayer 2022; Kluttz, Kohli, and Mulligan 2022; Alfrink et al. 2023; Russo and Toni 2023). Note that the possibility of ADSs being the contesting entities is not typically considered.

Contestation methods. Several methods have been proposed for facilitating contestation. In the realm of non-technical solutions, participatory design has been advocated as a way to involve various stakeholders in ADS development, enabling advance mitigation of possible issues and risks associated with the use of the ADS (Almada 2019; Vaccaro et al. 2021; Alfrink et al. 2023). The framework proposed in (Alfrink et al. 2022) suggests following a set of development practices over the ADS lifecycle, including incorporating ex-ante safeguards, gathering end-user feedback, implementing quality assurance, mitigating possible risks and allowing for third-party oversight. More technical approaches envision usage of operation logs (Hicks 2022) or model explanations (Almada 2019; Ploug and Holm 2020; Kluttz, Kohli, and Mulligan 2022; Russo and Toni 2023; Yurrita, Balayn, and Gadiraju 2023). We discuss these and their use of explanations in greater detail next.

The role of explanations. That explanations are needed to support contestability is acknowledged by some (e.g. by (Almada 2019; Lyons, Velloso, and Miller 2021; Alfrink et al. 2022; Wachter, Mittelstadt, and Russell 2017)), but the problem of using explanations to support contestability in practice has received little attention in explainable AI (XAI) (or AI for that matter) to date. A notable exception is (Russo and Toni 2023), where causal discovery with neural networks guides human feedback for contesting the discovered causal relations, which can be seen as a form of global explanation. Also, some current XAI methods can be seen as offering ground for some limited contestability when they suggest actionable recourse, as is the case for counterfactual explanations in terms of "actionable changes that an individual can make to flip the prediction of the classifier" (Venkatasubramanian and Alfano 2020). However, these works disregard that algorithmic decisions may actually be incorrect, e.g. because they are based on incorrect or incomplete data. Moreover, these methods are one-shot, providing no opportunities for follow-up inquiry, and shallow, revealing no information about the steps or logic that led to the explained output, and thus offer little ground for contestability. Some other XAI methods provide information on the AI models' deliberation and insecurities (Wang and Vasconcelos 2019); however, they do not support any form of contestability. Overall, explanations, as understood in state-of-the-art XAI, are seen as inadequate to support contestability (Henin and Métayer 2022). Specifically, (Yurrita, Balayn, and Gadiraju 2023) advocate a generic notion of explanations that capture the rationales behind the development and deployment of the ADS (referred to therein as "process-centric explanations").

Post-hoc and ex-ante contestability. Some take the view that contestability needs to be supported by regulatory frameworks, involving in particular policy modelling and normative reasoning (Tubella et al. 2020). These approaches see contestability as post-hoc processes, detached from explainability, "to review algorithmic decisions" (Lyons, Velloso, and Miller 2021). Ex-ante contestability is envisaged by some as a design principle (Hirsch et al. 2017; Alfrink et al. 2022), whereby systems are designed to enable users to interactively contest these systems and interaction between humans and systems is needed for "critique and correction" (Lyons, Velloso, and Miller 2021), but no technical solutions exist on how to support it.

Contestability as an interaction process. (Kluttz, Kohli, and Mulligan 2022) see contestability as an interaction process with humans, at development and deployment times, to allow humans, rather than just data, to train systems. Sev-

[5] We have identified these works by searching DBLP with 'Contestable' and 'Contestability', restricting attention to papers on AI from 2014 onwards published in peer-reviewed venues, while also considering additional references therein.

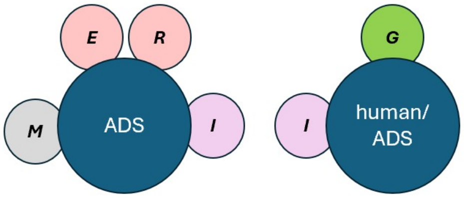

Figure 1: An abstract view of AI contestability: the contested ADS (left) is equipped with a model (M), an *explanation* method (E), and a *redress* method (R); the contester (right) is a human or an ADS equipped with a *ground generator* for contestations (G); both contested ADS and contester are able to interact (I).

eral other works (e.g. (Hirsch et al. 2017; Lyons, Velloso, and Miller 2021)) envisage post-hoc contestability as an interaction process, again between humans and ADSs alone.

Other related work. Works on *corrigibility*, such as (Soares et al. 2015; Carey and Everitt 2023; Russell, Dewey, and Tegmark 2015) focus on human overseers providing feedback/instructions to ADSs so that they align with the intentions/values of their user(s) (Gabriel 2020; Hadfield-Menell et al. 2016). While corrigibility can be viewed as enabling human users to contest the decisions of ADSs, it is narrowly focused on ADSs which optimise for reward when they have incentives to manipulate users (Ward, Toni, and Belardinelli 2022), ignore instructions (Hadfield-Menell et al. 2016), and disempower humans/other ADSs (Turner et al. 2021).

3 An Abstract View of AI Contestability

In a nutshell, towards supporting contestability, we see ADSs and humans as in Figure 1, assuming that the contested entity is an ADS and the contester is either a human or another ADS.[6]

Let $M : I \to O$ be a *model* computing outputs in O for inputs in I. M could be. e.g. a machine learning model, an expert system, or a combination of the two. M is part of the (contested) ADS, and the outputs of M are the ADS's decisions. For this ADS to be contestable (by a human or another ADS), it needs to be able to process *contestations*, which may be in reference to one of the following settings:[7]

(A) output $y = M(x)$ for some specific input x, or

(B) how M determines $y = M(x)$ from input x, or

(C) the full model M.

The first setting (A) amounts to contesting the model behaviour without referring to its internal "reasoning" process. The second setting (B) amounts to contesting (part of) the reasoning process of the model for the particular input. The third setting (C) amounts to contesting the model in general, e.g. its "philosophy", such as the underpinning input and output distributions and/or that it is biased.

Example 1. *As a simple illustration, consider M given by a (transparent) set of rules for binary classification amounting to awarding a loan (class* $\texttt{loan}$*), or not (class* $\neg\texttt{loan}$*), with binary features* employed *(*$\texttt{emp}$ *in short) and* career_breaks *(*$\texttt{breaks}$ *in short):*[8]

$$\forall X[\texttt{emp}(X) \wedge \neg\texttt{breaks}(X) \to \texttt{continuous\_emp}(X)]$$
$$\forall X[\texttt{continuous\_emp}(X) \to \texttt{loan}(X)]$$
$$\forall X[\neg\texttt{continuous\_emp}(X) \to \neg\texttt{loan}(X)]$$

For an input (loan applicant) x characterised by features $\texttt{emp}$ *and* $\texttt{breaks}$*, classified as* $\neg\texttt{loan}$ *by M, in the first setting a contestation may amount to objecting to the classification* $\neg\texttt{loan}$*, whereas in the second setting it may amount to objecting to the instances for $X = x$ of the top and bottom rules (i.e.* $\texttt{emp}(x) \wedge \neg\texttt{breaks}(x) \to \texttt{continuous\_emp}(x)$ *and* $\neg\texttt{continuous\_emp}(x) \to \neg\texttt{loan}(x)$*). In the third setting, the middle rule could be the object of contestation, without reference to any specific input.*

Explanations for contestability. In line with some of the literature, we envisage that contestability needs to be supported by explanations for M and/or its outputs for specific inputs. We thus assume that the ADS, in addition to M, is equipped with an *explanation method* E. Differently from (Yurrita, Balayn, and Gadiraju 2023), we believe that technical notions of explanations from the literature in XAI may already prove useful. These explanations may be *local* (i.e. relating to a specific input, output pair) or *global* (i.e. relating to all input, output pairs). Different forms of contestability may require and/or benefit from different forms of explanations and thus explanation methods, including as follows: 1. a *feature-attribution explanation* method, e.g. as provided by model-agnostic LIME (Ribeiro, Singh, and Guestrin 2016) or SHAP (Lundberg and Lee 2017) or by model-specific (tailored to neural classifiers) LRP (Montavon et al. 2019) or DeepLIFT (Shrikumar, Greenside, and Kundaje 2017); 2. an *abductive explanation* method, e.g. as in (Ignatiev, Narodytska, and Marques-Silva 2019); 3. a *counterfactual explanation* method e.g. as first envisaged in (Wachter, Mittelstadt, and Russell 2017; Tolomei et al. 2017); 4. a *rule-based explanation* method, e.g as in (Ribeiro, Singh, and Guestrin 2018; Guidotti et al. 2019; Ferreira et al. 2022); 5. a *mechanistic (local) explanation* method, e.g. as provided in DAX (Albini et al. 2020), SpArX (Ayoobi, Potyka, and Toni 2023), and (Wang et al. 2023); 6. a *mechanistic (global) explanation* e.g as in (Ayoobi, Potyka, and Toni 2023); 7. a *surrogate model* for interpreting M, e.g. (Kenny and Keane 2019; Tan and Kotthaus 2022; Potyka, Yin, and Toni 2023); 8. M

[6]For simplicity, we assume that the contested ADS is equipped with a single M and accompanying E and R, but in general the same ADS may be equipped with several. Similarly, we assume that the contester is equipped with a single G, but in general the same contester may be equipped with several (one per contested model). Further, the same ADS could be both contested and contester.

[7]We ignore here the trivial setting where contestations are directed at specific inputs, as these do not require altering M.

[8]These rules may correspond to the behaviour of a machine learning classifier, where $\texttt{continuous\_emp}$ is a "latent feature" of the model. We use them informally here (e.g. they may be implications in classical logic or logic programming rules, with $\neg$ as negation as failure).

itself, if already an *interpretable model* (e.g. a decision tree, or a set of rules).

We stress that satisfying some desirable properties will make the explanation methods more useful to the contestation. Notably, *faithfulness* requires that the explanations reflect the true reasoning of the explained models (Lakkaraju et al. 2019). This will guide the interactions in contestations towards the correct direction, and allow for more effective redress. Another important property is *robustness* (Jiang et al. 2024b). Non-robust explanation methods could result in drastically different explanations for two users with similar inputs (Leofante and Potyka 2024), which may jeopardise their explanatory function. In contrast, robust methods (Jiang et al. 2023; Leofante and Lomuscio 2023; Leofante, Botoeva, and Rajani 2023; Jiang et al. 2024a) improve the consistency and trustworthiness of explanations and may be better suited to support contestability.

Note that, in setting (A), an explanation is not strictly necessary, e.g. a loan applicant could contest not having received a loan no matter the reason. However, if an explanation is present, the contestation can be richer, e.g. if a loan applicant finds out that the reason for the loan refusal is their career-breaks, then they can contest the decision with evidence that they were encouraged to take career-breaks by the employer for training purposes.

Note also that, in settings (B-C), explanations are essential "windows" over the model, without which contestation cannot take place. Specifically, in the second setting, explanations are crucial and, arguably, they need to reveal the reasoning by the model for obtaining the output, e.g. as given by (faithful) mechanistic explanations. Indeed, input-output explanations alone (such as those computed by feature-attribution or counterfactual explanation methods) cannot always provide grounds for contestation. For instance, consider the model M in Example 1 and applicant x' characterised again by features `emp` and `breaks`, with the latter due to parental leaves: an explanation including the (instances for $X=x'$ of the) first and third rule in the model empowers the loan applicant to contest the decision by objecting to its bias against people who take parental leave; this is more powerful than contesting based on the feature $\texttt{breaks}(x)$ alone in a feature-attribution explanation.

While in the first two settings (A-B) local explanations suffice, in the third setting global explanations are needed. For illustration, a bank manager or regulator, inspecting (interpretable) model M in Example 1, may realise of a possible bias underpinning the strict definition of $continuous\_emp$ by the first rule, disregarding the possibility of career breaks by applicants.

For further illustration, consider the case of a black-box text classifier which, taken an input text, returns a classification, e.g. the sentiment of the text, or that the text is about a certain topic. Feature-attribution explanations for the classifications (such as those provided by LIME) can be used to pinpoint words in the input text deemed responsible for the classification, but not how the model determines the classification based on those input words – being thus unsuitable for the second and third settings.

Interaction for contestability. The process of contestation needs to be supported by suitable forms of interaction between the contested ADS and the (human or ADS) contester. Interaction may be in the form of conversations in natural language, e.g. using explanations generated by Large Language Models (Bills et al. 2023), especially if the contester is a human. Alternatively, they may be in structured format, using a formal agent communication language, e.g. in the spirit of FIPA (Poslad 2007).[9] In Figure 1, we indicate with I the method used by contested ADS and contester to engage in the interactions necessary for contestability.

The viewpoint of the contester. So far, we have taken the viewpoint of the ADS being contested. The contester may be a human, as in all the existing works on AI contestability, or, alternatively another ADS. In either case, for contestations to be acceptable, they need to be accompanied by some *grounds* for contestation, as in all earlier illustrations. Thus, we assume that the contester is equipped with a *ground generator method* G. For instance, in setting (A) for Example 1, if a loan applicant finds out that the reason for the loan refusal is their career-breaks and they decide to contest the decision with evidence that they were encouraged to take career-breaks by the employer for training purposes, this evidence forms the grounds for the contestation.

Redress. While explanations empower contestation, its complete realisation needs the ADS to have the ability to redress any issues (successfully) brought about in the contestation. Thus, we assume that the ADS is also equipped with a *redress method* R. For illustration, in the loan application Example 1, R may amount to revising the first rule by allowing for exceptions to be made when the breaks are due to training or parental leaves. As an additional illustration for text classifiers, redress may result from the a post-hoc reasoning process with the classifier's outputs and external knowledge encoded in argument schemes, as in (Carstens and Toni 2017), or with the classifier's outputs and explanations therefor, e.g. as in (Freedman et al. 2024).

Note that we see contestation, and thus redress, as a post-training process, rather than model debugging during training, such as via data augmentation (e.g. as in (Teso and Kersting 2019)) or regularisation (e.g. as in (Ross, Hughes, and Doshi-Velez 2017; Rieger et al. 2020; Shao et al. 2021; Zhang, Williams, and Toni 2024)). Thus, our view is that contestation is, in general, different from explanation-based model debugging (e.g. as in (Ghai et al. 2021) for tabular data, (Lertvittayakumjorn and Toni 2021) for text classification, or (Popordanoska, Kumar, and Teso 2020) for image classification). We, however, envisage that, in the case of models trained from data, redress may at times involve fine-tuning or retraining steps in which case the above methods, as well as methods for repairing AI models, e.g. (Henriksen, Leofante, and Lomuscio 2022; Almog and Kalech 2023), could be drawn upon. In the illustrative case of a black-box text classifier, an example of contestation of the full model (the third setting) is offered by FIND (Lertvittayakumjorn, Specia, and Toni 2020): here, LRP (Montavon et al. 2019) is used to associate output neurons in the feature extractor of a text classifier with word clouds. Users can then con-

[9] http://www.fipa.org/

test the use of individual neurons by disabling them and fine-tuning the model to no longer rely on them. This allows, specifically, to decrease model bias and reliance on artifacts (Lertvittayakumjorn, Specia, and Toni 2020). (Dreyer et al. 2024), similarly, aim to expose visual concepts learned by an image classifier using concept activation vectors (Kim et al. 2018) and fine-tune the model to mitigate the use of specific concepts using regularisation.

4 The Role of Computational Argumentation

In this position paper we argue that computational argumentation (CA) is ideally suited to support AI contestability computationally. CA, broadly understood as in (Atkinson et al. 2017; Baroni et al. 2018), is a branch of Knowledge Representation & Reasoning which represents information in terms of arguments and dialectical relations (of attack and, possibly, support) between them. CA is equipped with semantics to reach some form of consensus regarding conclusions to be drawn. As such, CA is ideally and uniquely suited to cover all aspects of the abstract view in Figure 1 organically. We support this view with reference to several lines of work in the CA literature, as follows.

CA for explanation. CA has been widely used for XAI (see (Vassiliades, Bassiliades, and Patkos 2021; Cyras et al. 2021; Guo et al. 2023) for recent overviews). It can provide abstractions of several existing, widely used models, e.g. as in (Potyka, Yin, and Toni 2023; Ayoobi, Potyka, and Toni 2023; Prakken and others 2020; Fan 2018; Cyras et al. 2019b), and can itself directly serve as the basis of models (Rago, Cocarascu, and Toni 2018; Rago et al. 2020; Cocarascu et al. 2020). Natural forms of explanations can be obtained from CA abstractions, e.g. dispute trees (Fan and Toni 2015b; Čyras et al. 2019a), defence sets (Arioua, Tamani, and Croitoru 2015) and attribution scores, e.g. gradient-based argument attributions (Yin, Potyka, and Toni 2023), Shapley-based relation attributions (Amgoud, Ben-Naim, and Vesic 2017; Yin, Potyka, and Toni 2024), amongst several others.

CA for redress. CA provides methods for revising (Snaith and Reed 2017; Baumann and Brewka 2015; Falappa, Kern-Isberner, and Simari 2009) and repairing knowledge bases (Ulbricht and Baumann 2019). Much research has been investigated in the context of forgetting (Berthold, Rapberger, and Ulbricht 2023; Baumann, Gabbay, and Rodrigues 2020) and enforcement (Rapberger and Ulbricht 2023; Baumann et al. 2021); researchers investigated the effect of expansions (Prakken 2023; Oikarinen and Woltran 2011; Cayrol, de Saint-Cyr, and Lagasquie-Schiex 2010) and changes in the knowledge base (Doutre and Mailly 2018; Booth et al. 2013; Moguillansky et al. 2013; Niskanen 2020). In addition, incomplete argumentation frameworks (Baumeister et al. 2021; Alfano et al. 2023) incorporate uncertainty which enables redress on a conceptual level.

These generic lines of work in CA are useful starting points for supporting redress of models which can be abstracted argumentatively. In addition, reasoning with argumentation frameworks drawn from models is the basis for forms of redress of (some) machine learning models, e.g. as in (Carstens and Toni 2017) and (Freedman et al. 2024) for natural language processing, and of (some) scheduling models, as in (Cyras et al. 2019b).

CA for interaction. Interactions between agents, often modelled as dialogues, have been shown to be effectively supported by various forms of CA, e.g. argument schemes (Panisson, McBurney, and Bordini 2021) or abstract argumentation frameworks (de Tarlé, Bonzon, and Maudet 2022), in a number of settings. CA's formal nature allows for principled desiderata of argumentative agent protocols in these settings, as defined in (McBurney, Parsons, and Wooldridge 2002). These settings include: computational persuasion (Fan and Toni 2012; Hunter 2018; Calegari, Riveret, and Sartor 2021; Donadello et al. 2022), framed as selecting the most effective arguments for changing the mind of the other agents; information-seeking and inquiry (Black and Hunter 2007; Fan and Toni 2015a), where agents share information comprising arguments which are private or public; and other areas such as the handling of maliciousness amongst agents (Kontarinis and Toni 2015). Approaches to multi-agent argumentation such as these have been been shown to be useful in various real-world applications, e.g. regulatory compliance (Raymond et al. 2022), recommender systems (Briguez et al. 2014; Teze, Godo, and Simari 2018; Rago et al. 2020; Rago et al. 2021) and, more recently, interactive XAI (Madumal et al. 2019; Calegari et al. 2022; Paulino-Passos and Toni 2022). One such approach which looks to have particular promise is that of (Rago, Li, and Toni 2023), where *argumentative exchanges* frame interactive explanation between agents as a conflict resolution problem, while accounting for humans' cognitive biases. These CA-based approaches to modelling interactions look to have promise for human interaction with Large Language Models, as in (Freedman et al. 2024).

CA and the viewpoint of the contester. CA-based dialogue and interaction methods, e.g. as in the aforementioned (Rago, Li, and Toni 2023), already have the potential to accommodate the contester's viewpoint, leveraging on the use of CA to provide abstractions for underpinning G. Indeed, e.g. in (Rago, Li, and Toni 2023), contested and contester are interchangeably seen as argumentation frameworks.

5 Discussion and Future Work

In this position paper, we have proposed an abstract view of contestable AI and advocated computational argumentation (CA) as ideally positioned to support this view.

Our analysis is restricted, for simplicity and brevity, but could be naturally extended to cover broader scenarios. For example, we have focused on two-party scenarios where a single ADS is contested and another ADS or a human is doing the contesting; however multi-party scenarios are also possible, e.g. when, in addition to a bank customer, a regulator contests a financial institution. Also, in line with the existing literature on contestability, we have assumed that only ADSs can be contested, but it may be useful to consider the possibility that humans may be contested by ADSs, e.g. in the spirit of (Miller 2023). In addition to broadening our view to these and other more complex scenarios, future work will be needed to provide evidence of our claim, by building forms of contestable AI that naturally use CA. This involves connecting what we see are the main requirements of contestability (explanations, grounds for contestation, redress, and interaction) in concrete applications and scenarios. CA has already been used to solve several individual challenges, e.g. see (Borg and Bex 2020; Lawrence, Visser, and Reed 2023; Cyras et al. 2019b). However, considerable engineering work is needed to combine solutions towards fully-fledged, end-to-end contestable systems. For instance, in order to support contestability with modern AI systems, we envisage that considerable effort will be required to build large-scale argumentation solvers.

Finally, we have advocated CA for contestability, but other KR techniques could provide useful support for some contestability aspects. Specifically, formal verification of AI models could be used to ensure the functional correctness of ADSs, e.g. as in (Kouvaros et al. 2023). Similarly, rigorous logic-based explainability techniques (Narodytska et al. 2019; Darwiche 2020) could be used to ensure the faithfulness and trustworthiness of explanations.

Acknowledgments

This research was partially supported by ERC under the EU's Horizon 2020 research and innovation programme (grant agreement No. 101020934), by J.P. Morgan and the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme, by Imperial College through an Imperial College Research Fellowship and by UKRI through the CDT in Safe and Trusted Artificial Intelligence (Grant No. EP/S023356/1) and the CDT in AI for Healthcare (Grant No. EP/S023283/1). Any views or opinions expressed herein are solely those of the authors. We thank Francis Rhys Ward for pointing the authors towards the literature on corrigibility, and Adam Gould, Avinash Kori and Andria Stylianou for helpful discussions.

References

Albini, E.; Lertvittayakumjorn, P.; Rago, A.; and Toni, F. 2020. DAX: deep argumentative explanation for neural networks. *CoRR* abs/2012.05766.

Alfano, G.; Calautti, M.; Greco, S.; Parisi, F.; and Trubitsyna, I. 2023. Explainable acceptance in probabilistic and incomplete abstract argumentation frameworks. *Artif. Intell.* 323:103967.

Alfrink, K.; Keller, I.; Kortuem, G.; and Doorn, N. 2022. Contestable ai by design: Towards a framework. *Minds and Machines* 1–27.

Alfrink, K.; Keller, I.; Doorn, N.; and Kortuem, G. 2023. Contestable camera cars: A speculative design exploration of public AI that is open and responsive to dispute. In *Proc. of CHI*, 8:1–8:16.

Almada, M. 2019. Human intervention in automated decision-making: Toward the construction of contestable systems. In *Proc. of ICAIL*, 2–11.

Almog, S., and Kalech, M. 2023. Diagnosis for post concept drift decision trees repair. In *Proc. of KR*, 23–33.

Amgoud, L.; Ben-Naim, J.; and Vesic, S. 2017. Measuring the intensity of attacks in argumentation graphs with shapley value. In *Proc. of IJCAI*, 63–69.

Arioua, A.; Tamani, N.; and Croitoru, M. 2015. Query answering explanation in inconsistent datalog +/- knowledge bases. In *Proc. of DEXA*, 203–219.

Atkinson, K.; Baroni, P.; Giacomin, M.; Hunter, A.; Prakken, H.; Reed, C.; Simari, G. R.; Thimm, M.; and Villata, S. 2017. Towards artificial argumentation. *AI Magazine* 38(3):25–36.

Ayoobi, H.; Potyka, N.; and Toni, F. 2023. SpArX: Sparse argumentative explanations for neural networks. In *Proc. of ECAI*, volume 372, 149–156.

Baroni, P.; Gabbay, D.; Giacomin, M.; and van der Torre, L., eds. 2018. *Handbook of Formal Argumentation*. College Publications.

Baumann, R., and Brewka, G. 2015. AGM meets abstract argumentation: Expansion and revision for dung frameworks. In *Proc. of IJCAI*, 2734–2740.

Baumann, R.; Doutre, S.; Mailly, J.; and Wallner, J. P. 2021. Enforcement in formal argumentation. *FLAP* 8(6):1623–1678.

Baumann, R.; Gabbay, D. M.; and Rodrigues, O. 2020. Forgetting an argument. In *Proc. of AAAI*, 2750–2757.

Baumeister, D.; Järvisalo, M.; Neugebauer, D.; Niskanen, A.; and Rothe, J. 2021. Acceptance in incomplete argumentation frameworks. *Artif. Intell.* 295:103470.

Berthold, M.; Rapberger, A.; and Ulbricht, M. 2023. Forgetting aspects in assumption-based argumentation. In *Proc. of KR*, 86–96.

Bills, S.; Cammarata, N.; Mossing, D.; Tillman, H.; Gao, L.; Goh, G.; Sutskever, I.; Leike, J.; Wu, J.; and Saunders, W. 2023. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index. [Accessed: 04.08.23].

Black, E., and Hunter, A. 2007. A generative inquiry dialogue system. In *Proc. of (AAMAS*, 241.

Booth, R.; Kaci, S.; Rienstra, T.; and van der Torre, L. W. N. 2013. A logical theory about dynamics in abstract argumentation. In *Proc. of SUM*, volume 8078, 148–161.

Borg, A., and Bex, F. 2020. Explaining arguments at the dutch national police. In *Proc. of AICOL*, 183–197.

Briguez, C. E.; Budán, M. C.; Deagustini, C. A. D.; Maguitman, A. G.; Capobianco, M.; and Simari, G. R. 2014. Argument-based mixed recommenders and their application to movie suggestion. *Expert Syst. Appl.* 41(14):6467–6482.

Calegari, R.; Omicini, A.; Pisano, G.; and Sartor, G. 2022. Arg2P: an argumentation framework for explainable intelligent systems. *J. Log. Comput.* 32(2):369–401.

Calegari, R.; Riveret, R.; and Sartor, G. 2021. The burden of persuasion in structured argumentation. In *Proc. of ICAIL*, 180–184.

Cao, L. 2022. Ai in finance: challenges, techniques, and opportunities. *ACM Computing Surveys (CSUR)* 55(3):1–38.

Carey, R., and Everitt, T. 2023. Human control: Definitions and algorithms. In *Proc. of UAI*, 271–281.

Carstens, L., and Toni, F. 2017. Using argumentation to improve classification in natural language problems. *ACM Trans. Internet Techn.* 17(3):30:1–30:23.

Cayrol, C.; de Saint-Cyr, F. D.; and Lagasquie-Schiex, M. 2010. Change in abstract argumentation frameworks: Adding an argument. *J. Artif. Intell. Res.* 38:49–84.

Cocarascu, O.; Stylianou, A.; Cyras, K.; and Toni, F. 2020. Data-empowered argumentation for dialectically explainable predictions. In *Proc. of ECAI*, 2449–2456.

Čyras, K.; Birch, D.; Guo, Y.; Toni, F.; Dulay, R.; Turvey, S.; Greenberg, D.; and Hapuarachchi, T. 2019a. Explanations by arbitrated argumentative dispute. *Expert Systems with Applications* 127:141–156.

Cyras, K.; Letsios, D.; Misener, R.; and Toni, F. 2019b. Argumentation for explainable scheduling. In *Proc. of AAAI*, 2752–2759.

Cyras, K.; Rago, A.; Albini, E.; Baroni, P.; and Toni, F. 2021. Argumentative XAI: A survey. In *Proc. of IJCAI*, 4392–4399.

Darwiche, A. 2020. Three modern roles for logic in AI. In *Proc. of PODS 2020*, 229–243. ACM.

de Tarlé, L. D.; Bonzon, E.; and Maudet, N. 2022. Multiagent dynamics of gradual argumentation semantics. In *Proc. of AAMAS*, 363–371.

Donadello, I.; Hunter, A.; Teso, S.; and Dragoni, M. 2022. Machine learning for utility prediction in argument-based computational persuasion. In *Proc. of AAAI*, 5592–5599.

Doutre, S., and Mailly, J. 2018. Constraints and changes: A survey of abstract argumentation dynamics. *Argument Comput.* 9(3):223–248.

Dreyer, M.; Pahde, F.; Anders, C. J.; Samek, W.; and Lapuschkin, S. 2024. From Hope to Safety: Unlearning Biases of Deep Models via Gradient Penalization in Latent Space. *Proc. of AAAI* 21046–21054.

Falappa, M. A.; Kern-Isberner, G.; and Simari, G. R. 2009. Belief revision and argumentation theory. In *Argumentation in Artificial Intelligence*. Springer. 341–360.

Fan, X., and Toni, F. 2012. Mechanism design for argumentation-based persuasion. In *Proc. of COMMA*, 322–333.

Fan, X., and Toni, F. 2015a. Mechanism design for argumentation-based information-seeking and inquiry. In *Proc. of PRIMA*, 519–527.

Fan, X., and Toni, F. 2015b. On computing explanations in argumentation. In *Proc. of AAAI*, 1496–1502.

Fan, X. 2018. On generating explainable plans with assumption-based argumentation. In *Proc. of PRIMA*, 344–361.

Ferreira, J.; de Sousa Ribeiro, M.; Gonçalves, R.; and Leite, J. 2022. Looking inside the black-box: Logic-based explanations for neural networks. In *Proc. of KR*.

Freedman, G.; Dejl, A.; Gorur, D.; Yin, X.; Rago, A.; and Toni, F. 2024. Argumentative large language models for explainable and contestable decision-making. *CoRR* abs/2405.02079.

Gabriel, I. 2020. Artificial intelligence, values, and alignment. *Minds Mach.* 30(3):411–437.

Ghai, B.; Liao, Q. V.; Zhang, Y.; Bellamy, R.; and Mueller, K. 2021. Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. *Proc. of HCI* 1–28.

Guidotti, R.; Monreale, A.; Giannotti, F.; Pedreschi, D.; Ruggieri, S.; and Turini, F. 2019. Factual and counterfactual explanations for black box decision making. *IEEE Intell. Syst.* 34(6):14–23.

Guo, Y.; Yu, T.; Bai, L.; Tang, J.; Ruan, Y.; and Zhou, Y. 2023. Argumentative explanation for deep learning: A survey. In *Proc. of ICUS*, 1738–1743.

Hadfield-Menell, D.; Russell, S.; Abbeel, P.; and Dragan, A. D. 2016. Cooperative inverse reinforcement learning. In *Proc. of NeurIPS*, 3909–3917.

Henin, C., and Métayer, D. L. 2022. Beyond explainability: justifiability and contestability of algorithmic decision systems. *AI Soc.* 37(4):1397–1410.

Henriksen, P.; Leofante, F.; and Lomuscio, A. 2022. Repairing misclassifications in neural networks using limited data. In *Proc. of SAC*, 1031–1038.

Hicks, A. 2022. Transparency, compliance, and contestability when code is(n't) law. In *Proc. of NSPW*, 130–142.

Hirsch, T.; Merced, K.; Narayanan, S. S.; Imel, Z. E.; and Atkins, D. C. 2017. Designing contestability: Interaction design, machine learning, and mental health. In *Proc. of DIS*, 95–99.

Hunter, A. 2018. Towards a framework for computational persuasion with applications in behaviour change. *Argument Comput.* 9(1):15–40.

Ignatiev, A.; Narodytska, N.; and Marques-Silva, J. 2019. Abduction-based explanations for machine learning models. In *Proc. of AAAI*, 1511–1519. AAAI Press.

Jiang, J.; Leofante, F.; Rago, A.; and Toni, F. 2023. Formalising the robustness of counterfactual explanations for neural networks. In *Proc. of AAAI*, 14901–14909.

Jiang, J.; Leofante, F.; Rago, A.; and Toni, F. 2024a. Recourse under model multiplicity via argumentative ensembling. In *Proc. of AAMAS*, 954–963.

Jiang, J.; Leofante, F.; Rago, A.; and Toni, F. 2024b. Robust counterfactual explanations in machine learning: A survey. In *Proc. of IJCAI*, 8086–8094.

Kenny, E. M., and Keane, M. T. 2019. Twin-systems to explain artificial neural networks using case-based reasoning: Comparative tests of feature-weighting methods in ANN-CBR twins for XAI. In *Proc. of IJCAI*, 2708–2715.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C. J.; Wexler, J.; Viégas, F. B.; and Sayres, R. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proc. of ICML*, 2673–2682.

Kluttz, D. N.; Kohli, N.; and Mulligan, D. K. 2022. Shaping our tools: Contestability as a means to promote responsible algorithmic decision making in the professions. In *Ethics of Data and Analytics*. 420–428.

Kontarinis, D., and Toni, F. 2015. Identifying malicious behavior in multi-party bipolar argumentation debates. In *Proc. of EUMAS*, 267–278.

Kouvaros, P.; Leofante, F.; Edwards, B.; Chung, C.; Margineantu, D. D.; and Lomuscio, A. 2023. Verification of semantic key point detection for aircraft pose estimation. In *Proc. of KR*, 757–762.

Lakkaraju, H.; Kamar, E.; Caruana, R.; and Leskovec, J. 2019. Faithful and customizable explanations of black box models. In *Proc. of AIES*, 131–138.

Lawrence, J.; Visser, J.; and Reed, C. 2023. Translational argument technology: Engineering a step change in the argument web. *J. Web Semant.* 77:100786.

Leofante, F., and Lomuscio, A. 2023. Robust explanations for human-neural multi-agent systems with formal verification. In *Proc. of EUMAS*, 244–262.

Leofante, F., and Potyka, N. 2024. Promoting counterfactual robustness through diversity. In *Proc. of AAAI*, 21322–21330.

Leofante, F.; Botoeva, E.; and Rajani, V. 2023. Counterfactual explanations and model multiplicity: a relational verification view. In *Proc. of KR*, 763–768.

Lertvittayakumjorn, P., and Toni, F. 2021. Explanation-based human debugging of NLP models: A survey. *Trans. Assoc. Comput. Linguistics* 9:1508–1528.

Lertvittayakumjorn, P.; Specia, L.; and Toni, F. 2020. FIND: human-in-the-loop debugging deep text classifiers. In *Proc. of EMNLP*, 332–348.

Lundberg, S. M., and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Proc. of NeurIPS*, 4765–4774.

Lyons, H.; Velloso, E.; and Miller, T. 2021. Conceptualising contestability: Perspectives on contesting algorithmic decisions. *Proc. of HCI* 106:1–106:25.

Madumal, P.; Miller, T.; Sonenberg, L.; and Vetere, F. 2019. A grounded interaction protocol for explainable artificial intelligence. In *Proc. of AAMAS*, 1033–1041.

McBurney, P.; Parsons, S.; and Wooldridge, M. J. 2002. Desiderata for agent argumentation protocols. In *Proc. of AAMAS*, 402–409.

Miller, T. 2023. Explainable AI is dead, long live explainable ai!: Hypothesis-driven decision support using evaluative AI. In *Proc. of FAccT*, 333–342.

Moguillansky, M. O.; Rotstein, N. D.; Falappa, M. A.; García, A. J.; and Simari, G. R. 2013. Dynamics of knowledge in *DeLP* through argument theory change. *Theory Pract. Log. Program.* 13(6):893–957.

Montavon, G.; Binder, A.; Lapuschkin, S.; Samek, W.; and Müller, K.-R. 2019. Layer-wise relevance propagation: an overview. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer. 193–209.

Muhammad, K.; Ullah, A.; Lloret, J.; Del Ser, J.; and de Albuquerque, V. H. C. 2020. Deep learning for safe autonomous driving: Current challenges and future directions. *IEEE Transactions on Intelligent Transportation Systems* 22(7):4316–4336.

Narodytska, N.; Shrotri, A. A.; Meel, K. S.; Ignatiev, A.; and Marques-Silva, J. 2019. Assessing heuristic machine learning explanations with model counting. In *Proc. of SAT*, volume 11628, 267–278.

Niskanen, A. 2020. *Computational Approaches to Dynamics and Uncertainty in Abstract Argumentation*. Ph.D. Dissertation, University of Helsinki, Finland.

Oikarinen, E., and Woltran, S. 2011. Characterizing strong equivalence for argumentation frameworks. *Artif. Intell.* 175(14-15):1985–2009.

Panisson, A. R.; McBurney, P.; and Bordini, R. H. 2021. A computational model of argumentation schemes for multi-agent systems. *Argument Comput.* 12(3):357–395.

Paulino-Passos, G., and Toni, F. 2022. On interactive explanations as non-monotonic reasoning. In *XAI Workshop@IJCAI2022*.

Ploug, T., and Holm, S. 2020. The four dimensions of contestable AI diagnostics - A patient-centric approach to explainable AI. *Artif. Intell. Medicine* 107:101901.

Popordanoska, T.; Kumar, M.; and Teso, S. 2020. Machine guides, human supervises: Interactive learning with global explanations. *arXiv preprint arXiv:2009.09723*.

Poslad, S. 2007. Specifying protocols for multi-agent systems interaction. *ACM Transactions on Autonomous and Adaptive Systems* 2(4):15–es.

Potyka, N.; Yin, X.; and Toni, F. 2023. Explaining random forests using bipolar argumentation and markov networks. In *Proc. of AAAI*, 9453–9460.

Prakken, H., et al. 2020. Pagerank as an argumentation semantics. *Proc. of COMMA* 55.

Prakken, H. 2023. Relating abstract and structured accounts of argumentation dynamics: the case of expansions. In *Proc. of KR*, 562–571.

Rago, A.; Cocarascu, O.; Bechlivanidis, C.; and Toni, F. 2020. Argumentation as a framework for interactive explanations for recommendations. In *Prof. of KR*, 805–815.

Rago, A.; Cocarascu, O.; Bechlivanidis, C.; Lagnado, D. A.; and Toni, F. 2021. Argumentative explanations for interactive recommendations. *Artif. Intell.* 296:103506.

Rago, A.; Cocarascu, O.; and Toni, F. 2018. Argumentation-based recommendations: Fantastic explanations and how to find them. In *Proc. of IJCAI*, 1949–1955.

Rago, A.; Li, H.; and Toni, F. 2023. Interactive explanations by conflict resolution via argumentative exchanges. In *Proc. of KR*, 582–592.

Rapberger, A., and Ulbricht, M. 2023. On dynamics in structured argumentation formalisms. *J. Artif. Intell. Res.* 77:563–643.

Raymond, A.; Malencia, M.; Paulino-Passos, G.; and Prorok, A. 2022. Agree to disagree: Subjective fairness in privacy-restricted decentralised conflict resolution. *Frontiers in Robotics and AI* 9.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proc. of KDD*, 1135–1144.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Anchors: High-precision model-agnostic explanations. In *Proc. of AAAI*, 1527–1535.

Rieger, L.; Singh, C.; Murdoch, W. J.; and Yu, B. 2020. Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge. *Proc. of ICML* 8086–8096.

Ross, A. S.; Hughes, M. C.; and Doshi-Velez, F. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. *Proc. of IJCAI* 2662–2670.

Russell, S.; Dewey, D.; and Tegmark, M. 2015. Research priorities for robust and beneficial artificial intelligence. *AI Mag.* 36(4):105–114.

Russo, F., and Toni, F. 2023. Causal discovery and knowledge injection for contestable neural networks. In *Proc. of ECAI*, 2025–2032. IOS Press.

Shaheen, M. Y. 2021. Applications of artificial intelligence (ai) in healthcare: A review. *ScienceOpen Preprints*.

Shao, X.; Skryagin, A.; Stammer, W.; Schramowski, P.; and Kersting, K. 2021. Right for Better Reasons: Training Differentiable Models by Constraining their Influence Functions. *Proc. of AAAI* 9533–9540.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. In *Proc. of ICML*, 3145–3153.

Snaith, M., and Reed, C. 2017. Argument revision. *Journal of Logic and Computation* 27(7):2089–2134.

Soares, N.; Fallenstein, B.; Armstrong, S.; and Yudkowsky, E. 2015. Corrigibility. In *AAAI Workshop on Artificial Intelligence and Ethics*.

Tan, H., and Kotthaus, H. 2022. Surrogate model-based explainability methods for point cloud nns. In *Proc. of WACV*, 2239–2248.

Teso, S., and Kersting, K. 2019. Explanatory interactive machine learning. In *Proc. of AIES*, 239–245.

Teze, J. C.; Godo, L.; and Simari, G. R. 2018. An argumentative recommendation approach based on contextual aspects. In *Proc. of SUM*, volume 11142, 405–412.

Tolomei, G.; Silvestri, F.; Haines, A.; and Lalmas, M. 2017. Interpretable predictions of tree-based ensembles via actionable feature tweaking. In *Proc. of KDD*, 465–474.

Tubella, A. A.; Theodorou, A.; Dignum, V.; and Michael, L. 2020. Contestable black boxes. In *Proc. of RuleML+RR*, 159–167.

Turner, A. M.; Smith, L.; Shah, R.; Critch, A.; and Tadepalli, P. 2021. Optimal policies tend to seek power. In *Proc. of NeurIPS*, 23063–23074.

Ulbricht, M., and Baumann, R. 2019. If nothing is accepted - repairing argumentation frameworks. *J. Artif. Intell. Res.* 66:1099–1145.

Vaccaro, K.; Xiao, Z.; Hamilton, K.; and Karahalios, K. 2021. Contestability for content moderation. *Proc. of HCI* 318:1–318:28.

Vassiliades, A.; Bassiliades, N.; and Patkos, T. 2021. Argumentation and explainable artificial intelligence: a survey. *The Knowledge Engineering Review* 36.

Venkatasubramanian, S., and Alfano, M. 2020. The philosophical basis of algorithmic recourse. In *Proc. of FAT**, 284–293.

Wachter, S.; Mittelstadt, B. D.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31:841.

Wang, P., and Vasconcelos, N. 2019. Deliberative explanations: visualizing network insecurities. In *Proc. of NeurIPS*, 1372–1383.

Wang, K. R.; Variengien, A.; Conmy, A.; Shlegeris, B.; and Steinhardt, J. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *Proc. of ICLR*.

Ward, F. R.; Toni, F.; and Belardinelli, F. 2022. On agent incentives to manipulate human feedback in multi-agent reward learning scenarios. In *Proc. of AAMAS*, 1759–1761.

Yin, X.; Potyka, N.; and Toni, F. 2023. Argument attribution explanations in quantitative bipolar argumentation frameworks. In *Proc. of ECAI*, 2898–2905.

Yin, X.; Potyka, N.; and Toni, F. 2024. Explaining arguments' strength: Unveiling the role of attacks and supports. In *Proc. of IJCAI*.

Yurrita, M.; Balayn, A.; and Gadiraju, U. 2023. Generating process-centric explanations to enable contestability in algorithmic decision-making: Challenges and opportunities. In *Workshop on Human-Centred Explainable AI*.

Yurrita, M.; Draws, T.; Balayn, A.; Murray-Rust, D.; Tintarev, N.; and Bozzon, A. 2023. Disentangling fairness perceptions in algorithmic decision-making: the effects of explanations, human oversight, and contestability. In *Proc. of CHI*, 134:1–134:21.

Zhang, D.; Williams, M.; and Toni, F. 2024. Targeted Activation Penalties Help CNNs Ignore Spurious Signals. *Proc. of AAAI* 16705–16713.